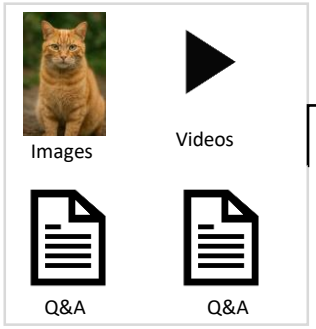
Customizing Vision-Language Models for Real-World Applications

Monika Jhuria

Technical Marketing Engineer

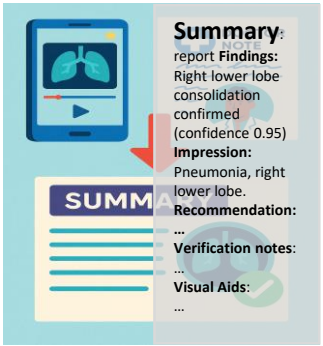
Nvidia

Vision-Language Models (VLMs): Where Images Meet Words



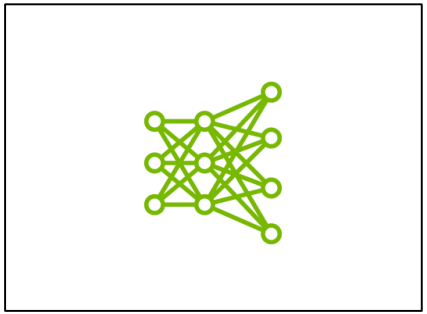
Vision Metadata

Vision-Language-Models

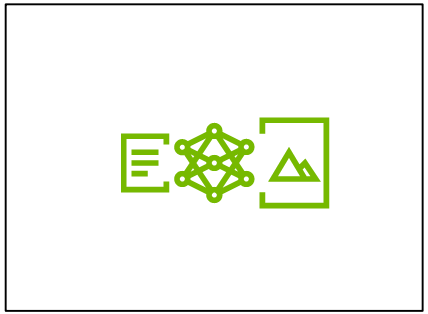


Why Vision Language Models are Important

Multimodal models capable of understanding and processing text, image and video



ZERO-SHOT LEARNING



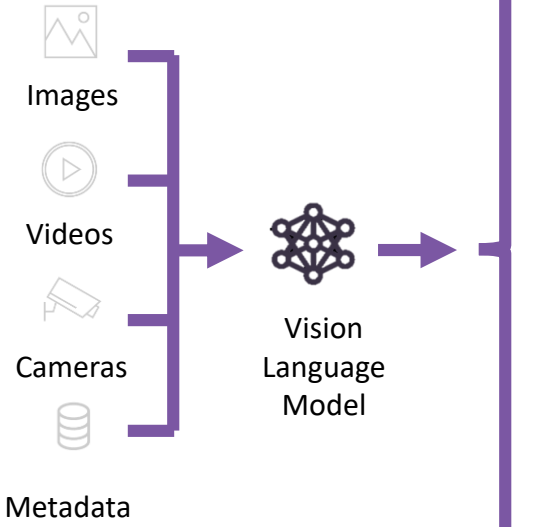
MULTIMODAL UNDERSTANDING

REASONING & COMPREHENSION

ENHANCED RETRIEVAL CAPABILITES

Vision Language Models For Insight Generation

Easily interact with your visual media to obtain valuable insights



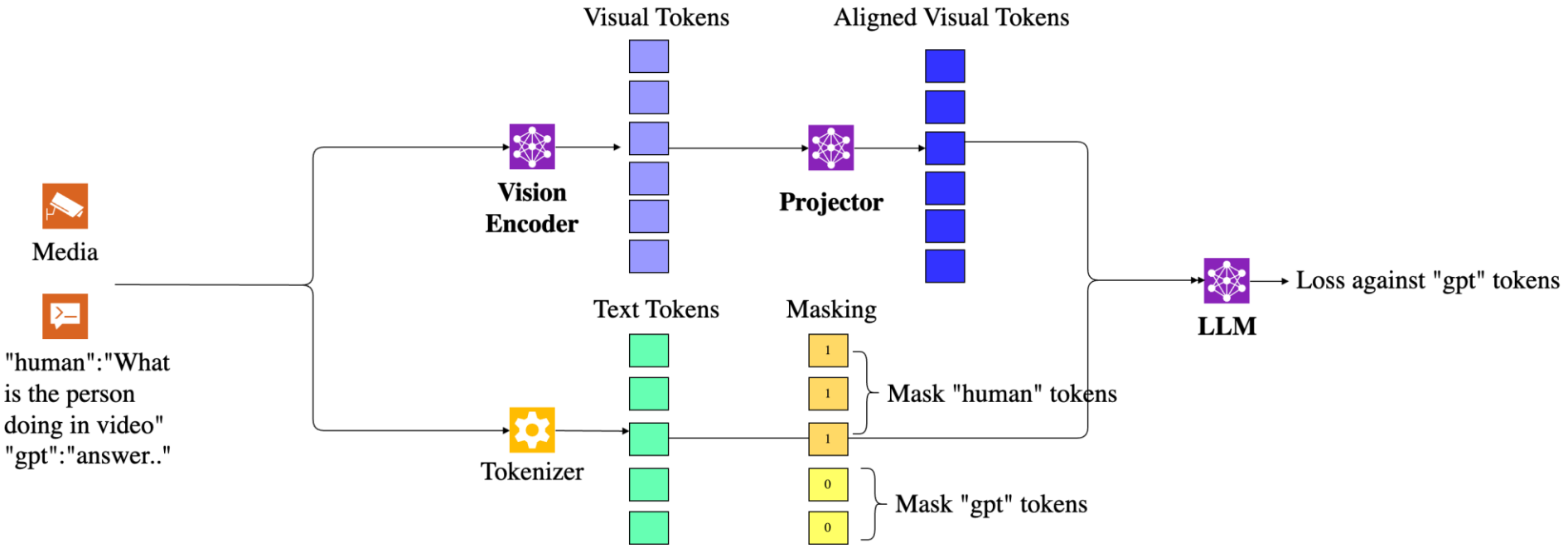
Describe any safety hazards and provide details what is being done to fix them.

The video shows a large fire burning near a body of water, with thick smoke rising into the air. Firefighters are working to extinguish the fire, with several fire trucks and firefighters visible in the scene. The firefighters are using hoses to spray water on the fire, and they appear to be working together to control the blaze.

Elaborate what the worker is doing and provide information on what he is wearing.

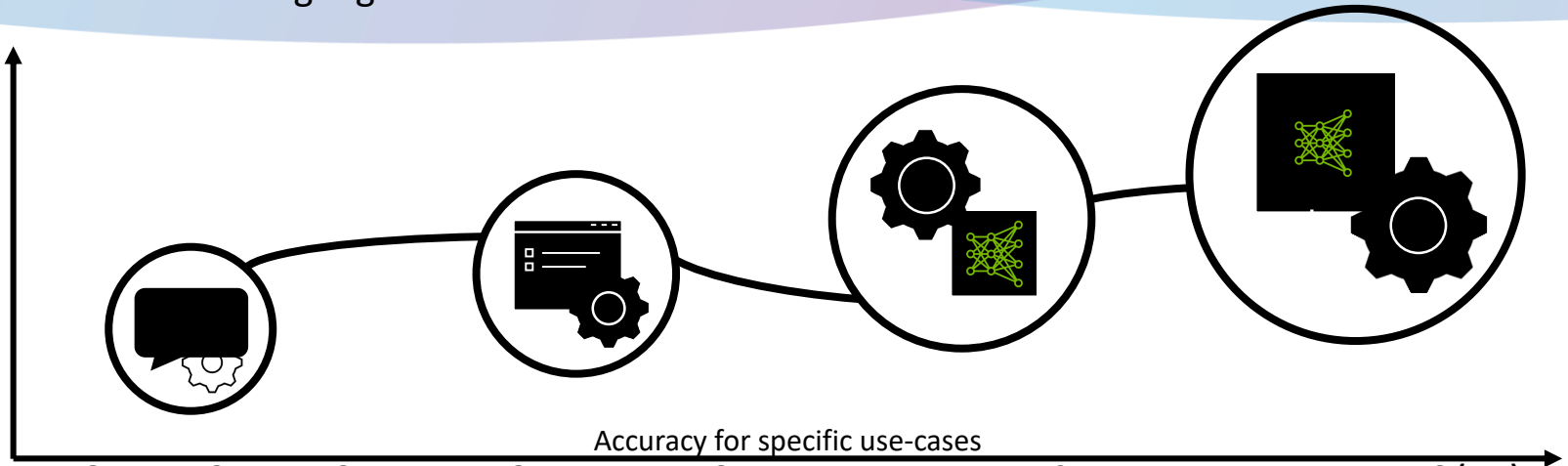
The worker is wearing a neon vest, a yellow hard hat, and black pants. He is pulling a yellow caution tape from the left side of the aisle to the right side, stretching it across the aisle to block it off. The worker's actions suggest that he is setting up a restricted area, possibly for safety or maintenance purposes.

Understanding VLM Architecture



Model Customization Options

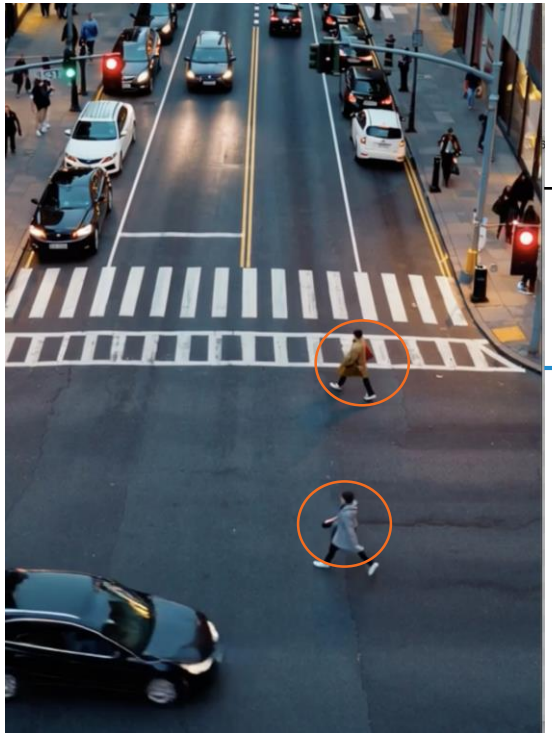
Ways To Customize Vision Language Models For Your Use-Cases



	PROMPT ENGINEERING	PROMPT LEARNING	PARAMETER EFFICIENT	FULL FINE TUNING (FFT)
Techniques	• Few-shot learning • Chain-of-thought reasoning • System prompting	• Prompt tuning • P-tuning	• LoRA	• SFT
Benefits	• Good results leveraging pre-trained VLMs • Lowest investment • Least expertise	• Better results leveraging pre-trained VLMs • Lower investment • Will not forget old skills	• Best results leveraging pre-trained VLMs • Will not forget old skills	• Best results leveraging pre-trained VLMs • Change all model parameters
Challenges	• Limited ability to utilize proprietary data and skills	• Limited ability to use proprietary data and change model parameters	• Medium investment • Takes longer to train • More expertise needed	• May forget old skills • Large investment • Most expertise needed

VLM Finetuning

When not to Finetune VLMs



GOAL: Find and describe unsafe pedestrian behavior.

Describe the scene

Zero-shot

The image depicts a busy city street at night, with multiple cars and pedestrians crossing the street. There are at least 12 people visible in the scene, with some walking and others crossing the street.

Some of the pedestrians are carrying handbags, indicating that they might be going about their daily activities. The street is well-lit, with multiple streetlights illuminating the area. The traffic lights are also present, ensuring smooth flow of traffic.

This is a traffic intersection scene. Look for zebra crossings and sidewalk. Describe any risky pedestrian behavior and describe what they are wearing.

Prompt learning

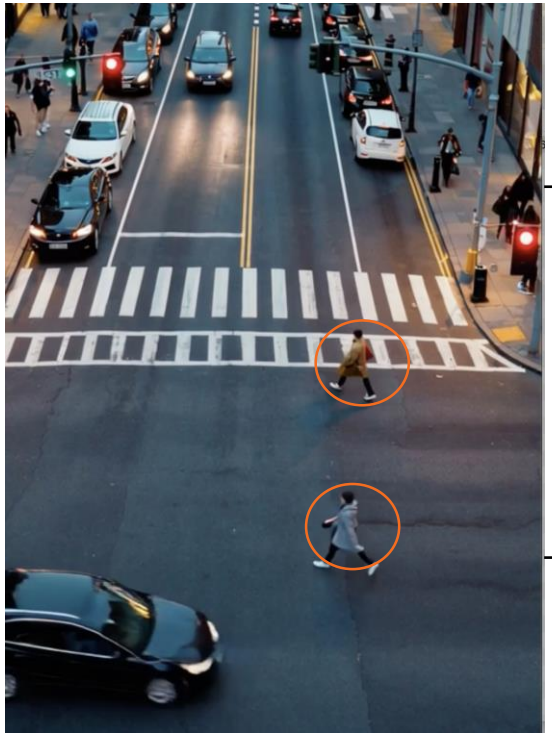
Grey hooded coat, dark pants, white sneakers — jay-walking mid-intersection, far outside the marked crosswalk while vehicles have a green light.

Tan trench coat, dark pants, black sneakers, red shoulder bag — inside the stripes but stepping off during the vehicle-green phase, crossing against the signal.

Used XXX model
Few shot learning paper from OpenAI

VLM Finetuning

When to Finetune VLMs



GOAL: Find and describe unsafe pedestrian behavior.

Zero-shot →

yes, there is an anomaly in the way persons are crossing the road. The person is not close to any car, and they are not really close to any vehicle.

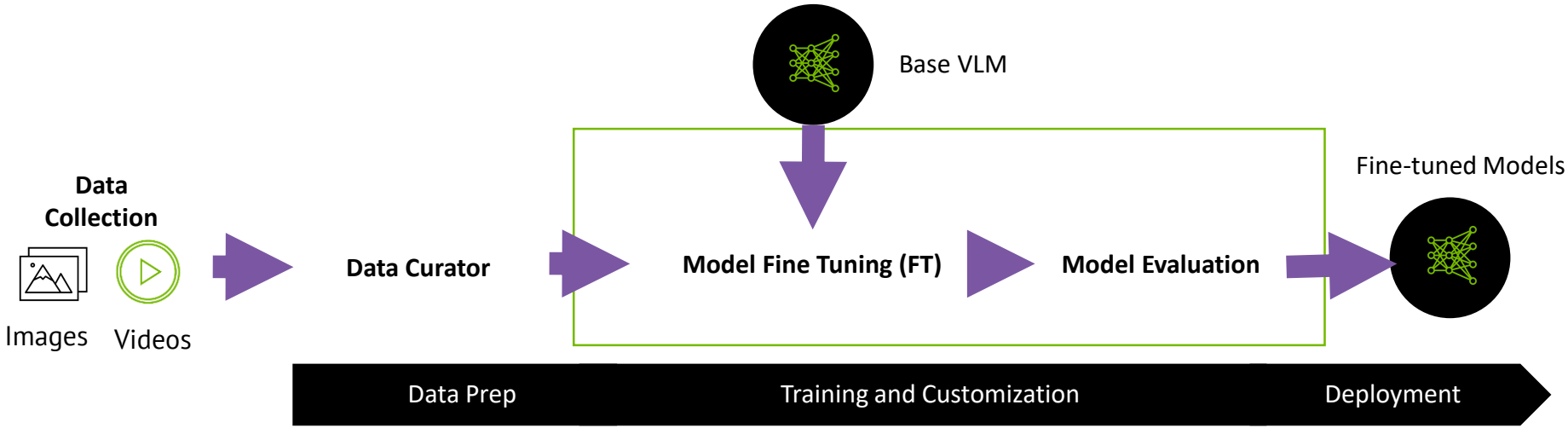
Prompt

Is there any anomaly the way people are crossing the road crossing? If so, how far the persons are from the vehicles?

After FT →

Tan-coat pedestrian: Walking ~1 m below the crosswalk, illegally in the lane and exposed to traffic ~7 m away.
Grey-coat pedestrian: Jay-walking ~10 m below the zebra stripes, only ~3 m ahead of an approaching black sedan—high collision risk.

Fine-Tuning Framework For Vision Language Models

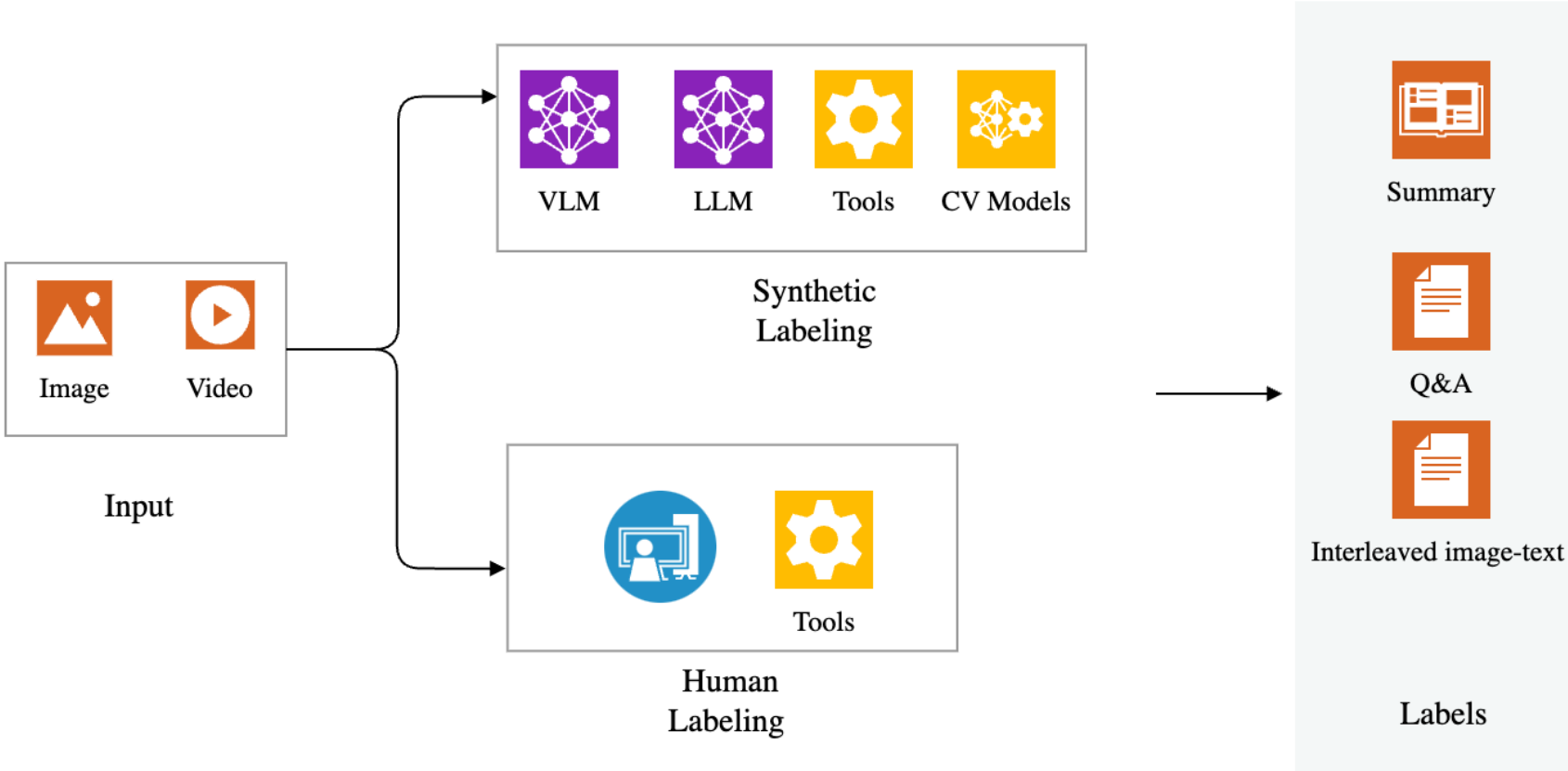


DETECT SPECIFIC OBJECTS OR EVENTS

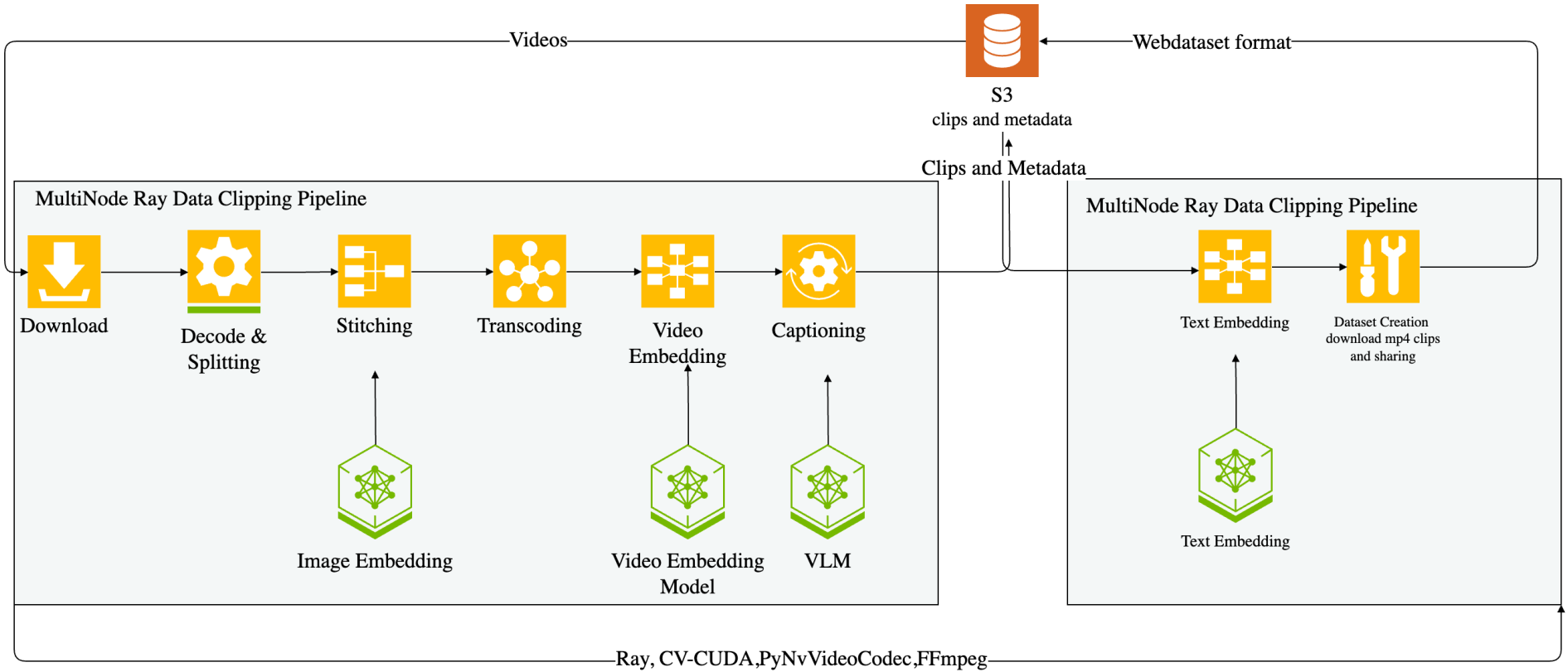
DOMAIN ADAPTION

MULTIMODAL UNDERSTANDING

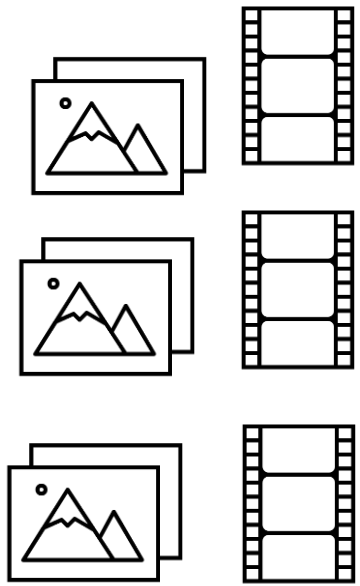
Data Curation



Data Curation
(Synthetic labeling)



Data Curation
(Type of Labels)



Media

Q&A

MCQ Style:

Questions: What Kind of anomaly is present in this scene?
Options:
A: Pedestrian jaywalking
B: Vehicle running a red light
C: Pedestrian crossing at marked crosswalk

GQA Style:

Questions: What unsafe action is happening and why is it dangerous?
Answer: The person is crossing outside the designated crosswalk….

Spatial Style:

Questions: “Describe the anomaly, including where and how it’s occurring?”

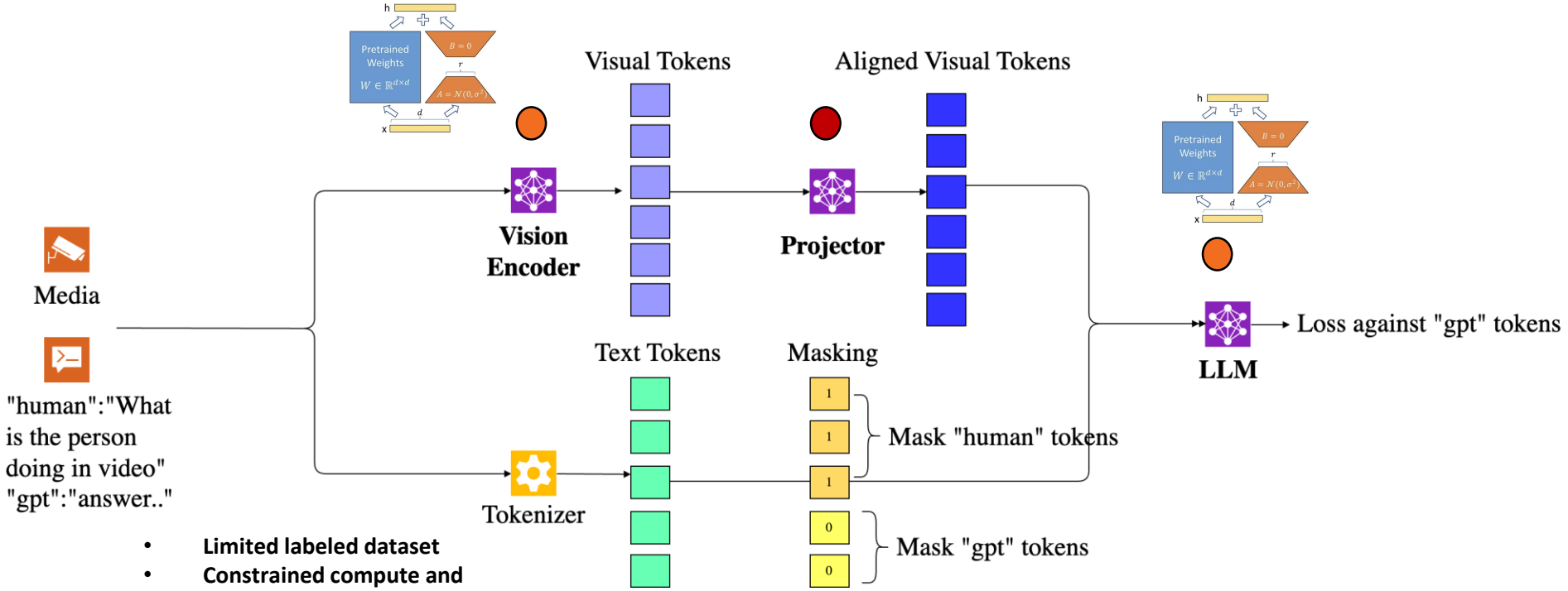
Answer: “ A pedestrian is jaywalking at the bottom-right of the frame (around 70% from the left, 85% from the top), stepping onto the road outside the marked crosswalk..”

Structured Style:
Question: Describe the occurred anomaly in the traffic scene. Response in a json format: {“anomaly_type”: , “bbox”:.., “alert”:..}

Answer:
{
“**anomaly_type**”: “pedestrian jaywalking”,
“bbox”: [450, 300, 500, 600],
“**alert**”: “Pedestrian jaywalking detected”
}

Model Finetuning

(Memory Efficient: LoRA/QLoRA…)

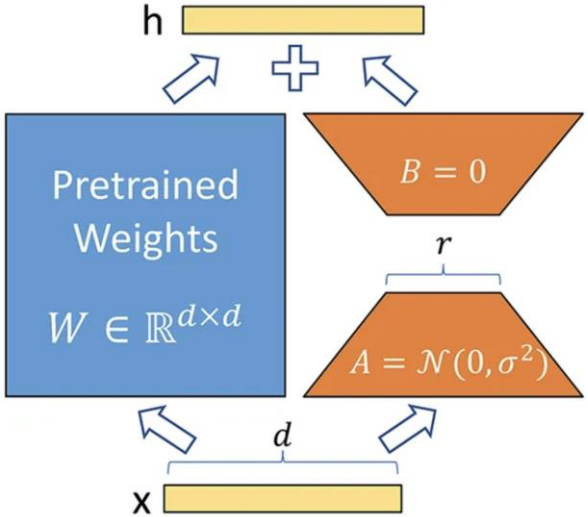


- **Limited labeled dataset**
- **Constrained compute and memory**
- **Modular deployments:** multi-task specific heads

Ref: LoRA image from: https://arxiv.org/abs/2106.09685

Model Finetuning

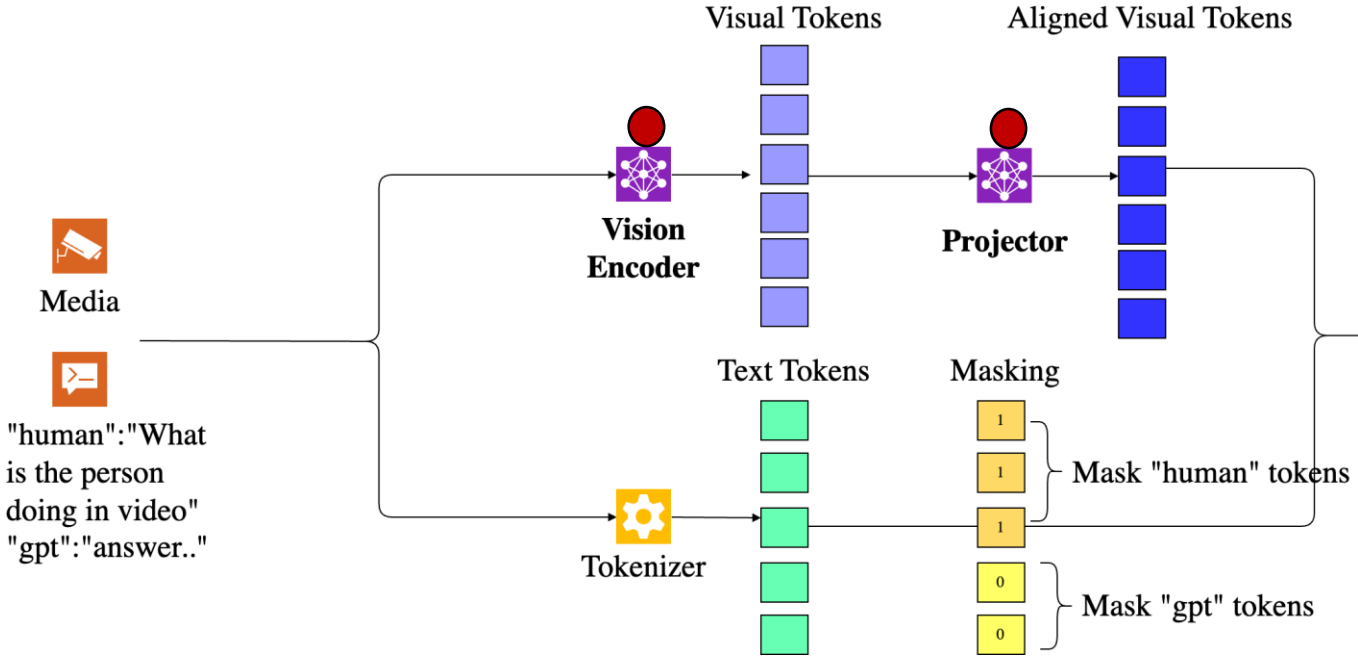
(Memory Efficient: LoRA/QLoRA…)



Config:
rank (r): smaller the r less params will be trained: decide based on available HW and complexity in usecase
Alpha: step size for adapter. Usually alpha=rank/2

- **Higher diversity** → higher rank
- **Small dataset** → lower rank (to avoid overfitting)
- **Higher rank** → more GPU RAM and compute
- **Complex reasoning** → may benefit from higher rank
- **Higher rank** -> Higher training time

Model Finetuning
(Full finetuning)



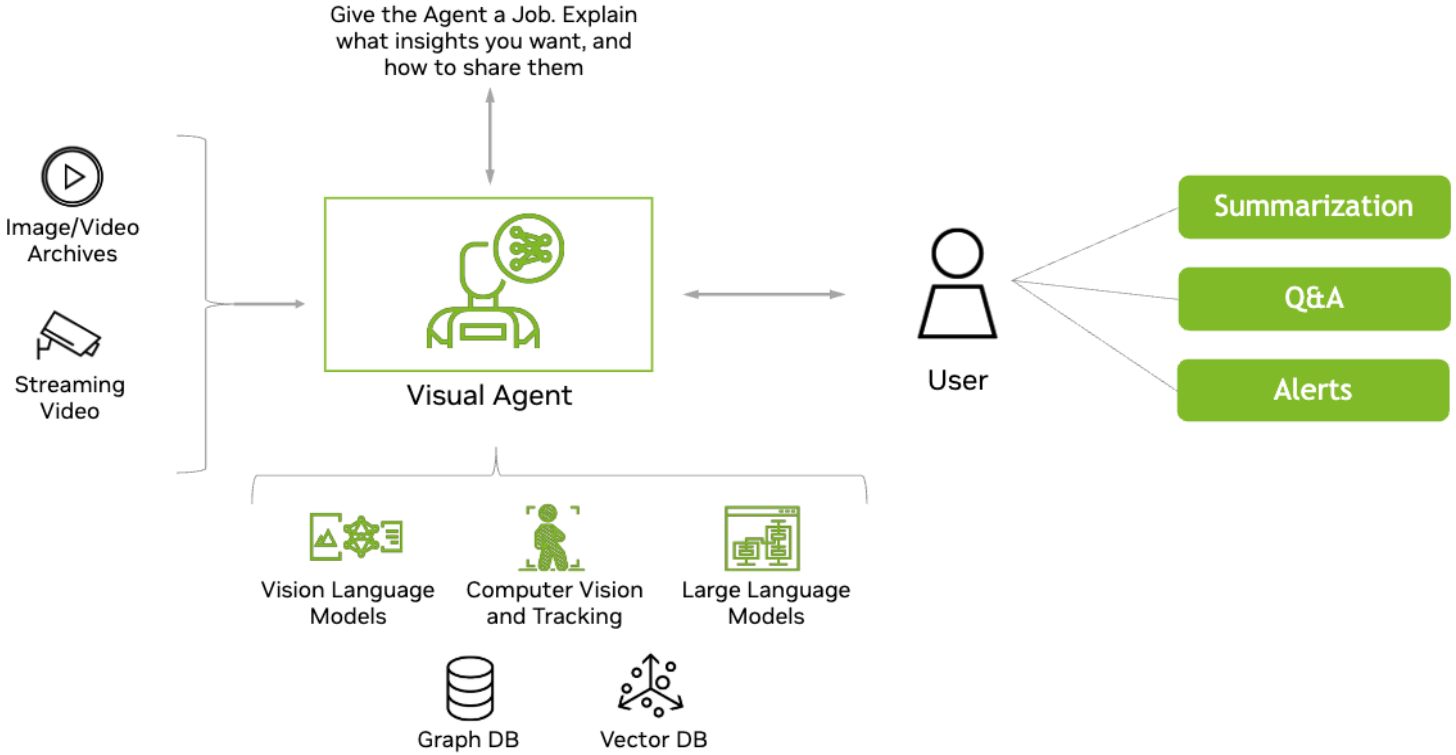
- **Adapt for new task**: visual + text labels are significantly different
- **High compute and memory** resources available
- **Adjust for new tokens** , positional encoding: example finetune for structured data, temporal
- **High accuracy**

VLM Evaluation Methods

Method	Evaluates	When to Use	Use Cases
BLEU	N-gram overlap between generated and reference text	Quick automated quality checks for text generation	Machine translation, image captioning, dialogue systems
MMMU	Cross-disciplinary reasoning and expert-level understanding	Testing AGI potential in complex, real-world scenarios	Academic subjects (STEM, humanities), multimodal problem-solving
CIDER	Adversarial pattern detection in cross-modal inputs	Security hardening of MLLMs against jailbreak attacks	Malicious query detection, image perturbation analysis
VHELM	9 key aspects including fairness, toxicity, and multilingual capabilities	Comprehensive model comparison across diverse criteria	Holistic safety audits, bias detection, multilingual performance analysis

Build Applications with VLMs and RAG

Unlock knowledge and insights from camera streams and archived videos



NVIDIA Explore NIM Docs Forums

Search API Catalog

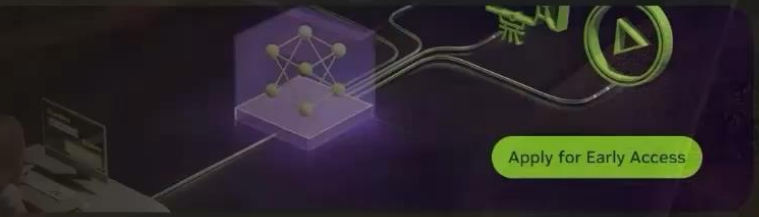
Login

Build a Video Search and Summarization Agent

Ingest massive volumes of live or archived videos and extract insights for summarization and interactive Q&A

LLAMA-3_1-70B-INSTRUCT • **NV-EMBEDQA-E5-V5** • **NV-RERANKQA-MISTRAL-4B-V3**

agent blueprint | chat | generative ai | video-to-text | vision



Apply for Early Access

Experience NIM Blueprint Card

AI models generate responses and outputs based on complex algorithms and machine learning techniques, and those responses or outputs may be inaccurate, harmful, biased or indecent. By testing this model, you assume the risk of any harm caused by any response or output of the model. Please do not upload any confidential information or personal data unless expressly permitted. Your use is logged for security purposes.

View Examples

Video Input

GOVERNING TERMS: This trial is governed by the **NVIDIA API Trial Terms of Service**. NVIDIA Retrieval QA Mistral 4B Reranking: Apache license, NVIDIA Retrieval QA E5 Embedding v5: NV-EmbedQA-E5-v5: MIT license; NV-EmbedQA-Mistral7B-v2: Apache 2.0 license, and Snowflake arctic-embed-l: Apache 2.0 license - The use of these models is governed by the **AI Foundation Models Community License Agreement**. ADDITIONAL INFORMATION: **Llama 3.1 Community License Agreement, Built with Llama**

Reset Launch Agent

Summarize the media

To begin chatting with the AI, start by clicking Launch Agent

Conclusion: From Concepts to Real-World VLMs

What is a VLM?
- Combines **vision + language** for multimodal understanding.
- Powers apps like image captioning, visual Q&A, search, and copilots.

VLM Data Curation
- **Synthetic generation**, human annotations.
- **Heuristic filtering** and synthetic pairing.
- Goal: build clean, aligned vision-text datasets.

Fine-Tuning Methods
- **Prompting** → Fast, flexible.
- **LoRA** → Efficient, low-resource finetuning.
- **Full Finetuning** → Deep control, high cost.
- Choose based on task size, data, and hardware.

Building Real Applications
- Combine VLMs with **RAG** for grounded answers.
- Use cases:
    - Visual search & analytics
    - Enterprise automation

Resources

- NeMo Curator for developers: https://developer.nvidia.com/nemo-curator
- VLM Prompting: https://developer.nvidia.com/blog/vision-language-model-prompt-engineering-guide-for-image-and-video-understanding/
- Nvidia Vision NIM: https://build.nvidia.com/explore/vision
- VLM finetuning MS: https://developer.nvidia.com/vlm-fine-tuning-microservice-early-access
- Nvidia VLMs: https://docs.nvidia.com/nim/vision-language-models/latest/introduction.html
- Nvidia's Video search and summarization Microservice: https://build.nvidia.com/nvidia/video-search-and-summarization

Disclaimer

The views and opinions expressed in this presentation are my own and do not necessarily reflect those of my employer.