Moonshot: Spatial Intelligence in the Physical World

(A Technical Architecture & Development Report)

Abstract

Moonshot enables supervisors to upload POV or observational videos from construction sites. The platform aggregates state-of-the-art models such as Cosmos-Reason2-8B for visual understanding and dense captioning, Llama 3.1 8B for summarization and reasoning, and NVIDIA's EmbedQA and RerankQA for retrieval. This is to power AI-driven analysis through the NVIDIA Video Search and Summarization (VSS) architecture. Google Gemini then structures the output into actionable reports with safety, productivity, and quality metrics.

Technology Stack

Our system is powered by Vision-Language Models (VLMs) and Large Language Models (LLMs), which serve as the reasoning engine.

Model Type	Model Name	Architecture Usage
Large Language Model (LLM)	meta/llama-3.1-8b-instruct	Used for structured reasoning, summarization, metric extraction, and enforcing operational output constraints.
Embedding Model	nvidia/llama-3.2-nv-embedqa-1b-v2	Used to embed structured outputs for indexing, similarity search, and historical report retrieval within MongoDB.
Reranker	nvidia/llama-3.2-nv-rerankqa-1b-v2	Used to prioritize the most relevant activity segments and refine structured outputs before final report generation.
Vision-Language Model (VLM)	nvidia/Cosmos-Reason2-8B	Used within the NVIDIA VSS architecture for: action reasoning, worker-object interaction understanding, and temporal event extraction Context-aware activity classification

We fine-tuned system behavior through structured prompting rather than retraining base model weights (no point reinventing the wheel). Specifically, we:

- Directed the models to focus on worker actions, task transitions, and duration
- Instructed them to ignore irrelevant background details
- Enforced structured JSON output formats
- Constrained outputs to operational metrics rather than descriptive narration
- Implemented schema validation to guarantee structured responses

Application Layer

Frontend:

Next.js	React	TailwindCSS	shadcn/ui	Zod

Data Access Layer:

MongoDB	UploadThing

The supervisor dashboard uses a split-screen interface:

- Left panel: Video playback
- Right panel: Structured analysis report
- Supervisors can review, edit, and manage reports with full CRUD functionality.

Video Intelligence Engine

We utilized the NVIDIA VSS architecture to power our backend VM as the perception layer of the system by aggregating data and inferences from various models.

NVIDIA VSS performs:

Object detection	Worker tracking	Action classification	Timestamp generation

This provides the foundational visual intelligence before structured reasoning is applied.

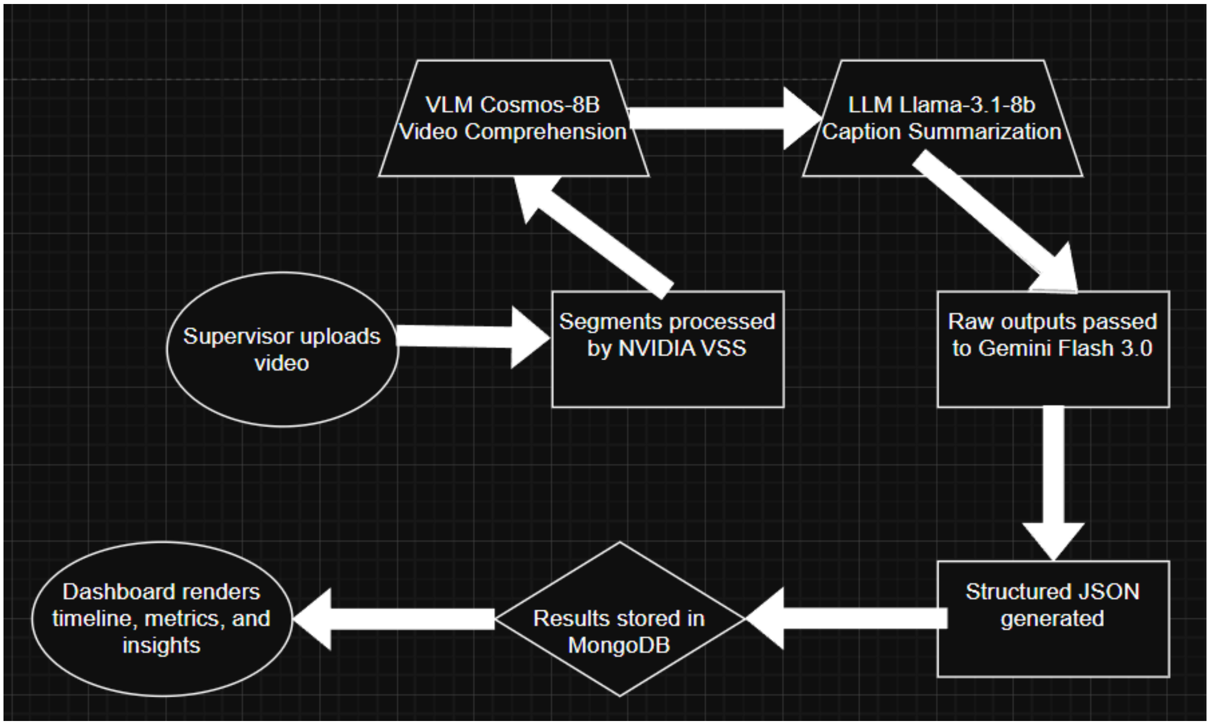
Infrastructure

All models run on a Vast.ai VM with a:

- 1x RTX PRO 6000 Blackwell Workstation GPU (96GB VRAM)
- AMD EPYC 7K62 48-Core Processor

This hardware enables high-throughput video inference, parallel clip processing, and structured reasoning without performance bottlenecks.

System Flow


VLM Cosmos-8B Video Comprehension
LLM Llama-3.1-8b Caption Summarization
Supervisor uploads video
Segments processed by NVIDIA VSS
Raw outputs passed to Gemini Flash 3.0
Dashboard renders timeline, metrics, and insights
Results stored in MongoDB
Structured JSON generated

Conclusion

Ultimately, by combining structured LLM reasoning, targeted model conditioning, automated segmentation, and NVIDIA-powered visual inference, we were able to develop an intelligent system that can translate raw construction footage into operational intelligence, which is an important representation of how a layered AI system can go beyond the visual description and provide decision-ready insights.