Mikshi VLM: Real-Time Video Intelligence

Turns unstructured video feeds into decision ready insights

Mikshi Research Team | mikshivlm.ai

Abstract

Video streams are inherently unstructured, making it difficult to derive reliable, real time insights from fast changing scenes. Mikshi VLM bridges this gap by transforming raw video into semantic intelligence. It goes beyond frame level perception to deliver complex scene understanding, precise temporal grounding, natural language video QA, and real time anomaly detection.The system can support a wide range of applications across traffic analytics, crisis response and security monitoring.

1 INTRODUCTION

The evolution of Vision Language Models (VLMs) has largely focused on static image comprehension. However, industrial applications ranging from autonomous traffic management to security surveillance require a nuanced understanding of temporal dynamics. Mikshi VLM is engineered to bridge the gap between heavy, high latency foundation models and the need for real time, grounded video analysis.

2 CORE CAPABILITIES

Mikshi VLM extends beyond frame by frame processing to provide a holistic narrative of video sequences.

- **Complex Scene Understanding:** Distinguishes critical actions from background noise in high density feeds.
- **Temporal Grounding:** Utilizes frame aware embeddings to pinpoint exact event timestamps.
- **Visual QA:** Natural language interface for querying past and live video data.
- **Anomaly Detection:** Real time risk assessment for collision, intrusion, or operational deviations.

3 TECHNICAL ARCHITECTURE

The Mikshi framework relies on four primary technological pillars to optimize the trade off between reasoning depth and inference speed.

3.1 Golden Dataset Creation

We created a proprietary Golden Dataset comprising 170 challenging video sequences and 1,000 curated QA pairs to rigorously evaluate model performance across event detection, scene description, temporal grounding, and other industrially relevant tasks.

3.2 Dynamic Context Engineering

The model uses a context aware prompting mechanism that dynamically switches the system prompt based on the application domain

3.3 Hybrid RAG Architecture

To manage long form context without token overflow, Mikshi VLM integrates:

1. **Vector DB:** For high speed semantic retrieval of similar visual patterns.
2. **Graph DB:** To maintain spatial and causal relationships between objects over time.

3.4 Re-Inference Engine

A secondary verification loop evaluates cases where the RAG module retrieves insufficient or low quality context. If the retrieved evidence cannot support a confident answer, the engine triggers a re-inference which improves the overall accuracy

4 PERFORMANCE EVALUATION

Mikshi VLM was benchmarked against the Qwen3-VL series, focusing on both academic metrics and real world latency.

4.1 Accuracy Benchmarks

As shown in Table 1, Mikshi VLM maintains a competitive advantage in temporal grounding and event detection, outperforming larger models in specialized tasks.

Table 1: Accuracy Comparison (%)

Benchmark / Metric	Mikshi VLM	Qwen3-VL 30B	Qwen3-VL 8B	Qwen3-VL 4B
GOLDEN DATASET				
Reasoning	75.0	**75.5**	69.2	61.3
Temporal Grounding	**70.4**	66.7	67.2	61.5
Event Detection	**72.4**	71.9	70.2	64.9
Collision	70.7	**71.5**	65.3	61.3
STANDARD DATASETS				
VideoMME	75.38	**78.29**	75.80	73.56
MVBench	46.93	**50.77**	46.88	44.62

4.2 Latency Analysis

Efficiency is the primary differentiator for Mikshi VLM. In comparative tests (Table 2), Mikshi VLM achieved the lowest latency, making it the only viable candidate for real time risk detection.

Table 2: Inference Latency (Seconds)

Model	Latency	Result
Mikshi VLM	**4.55**	**Optimal**
Qwen3-VL 4B	10.93	High Latency
Qwen3-VL 8B	11.79	High Latency
Qwen3-VL 30B	11.82	High Latency

5 USE CASE DEPLOYMENTS

The system is currently optimized for these domains:

- **Traffic analytics:** Managing vehicle queues, accidents and signal violations.
- **Crisis Response:** Fire, Disaster detection.
- **Security:** Automated anomaly detection in secure zones.
- **Analytics:** Tracking 80+ object categories in urban grids.
- **Sports Analytics:** Analyze any sport events and get your queries answered.