



THE PROBLEM

The 50 Million Token Problem

Why Visual Data Breaks LLMs, and How to Fix It

Gabi Kazav, VP R&D, Visual layer.

THE PROBLEM

50,000,000

tokens for a single query

That's what a single metadata query generates on a 100M-image dataset. Context windows never stood a chance.

Claude Opus 4.8

1,000,000 tokens

Gemini 3.5 Pro

2,000,000 tokens

The Gap

25–50× beyond the largest mainstream window

THE PROBLEM

The World Is Going Visual

80–90% of enterprise data is unstructured, and teams want to talk to it in natural language.



Medical Imaging

Diagnostic scans, labels, and confidence scores



Satellite Imagery

Geo-tagged frames with change-detection metadata



Manufacturing

Frame-level defect detections and similarity clusters



Autonomous Driving

Hundreds of millions of tagged frames per fleet, per day



THE PROBLEM

What the LLM Actually Sees

LLMs can't process 100M images directly — they reason over text metadata. And that metadata explodes.

- 2M unique labels × 50 tokens = 100M tokens
- user tags × metadata = hundred thousands of tokens
- videos × frames = Millions of tokens
- custom fields × values = Millions of tokens

Total: >100M Tokens

No model handles this with brute force. The architecture must change.

THE PROBLEM

The Context Window Race

Windows grew 1,000×+ since 2020. Visual metadata is still off the chart.



FAILURE MODE 1 OF 3

"Just Dump Everything In"

Concatenate all metadata into the prompt. 2M labels $\times$ 50 tokens = 100M tokens. Impossible — even the 10M frontier ceiling is **10 $\times$ short**.

Prompt Size

~100M tokens

Model Limit

1M mainstream $\cdot$ 10M max

Result

10 $\times$ –100 $\times$ gap



FAILURE MODE 2 OF 3

"Pass the Search Results"

Filter first, then pass results. Smarter — but still broken.

The High-Cardinality Problem

A search for "cat" matches 3.7M images, 127K unique labels. That's 5MB of JSON. One search exhausts context.

Why It's Not Like Document

RAG

Visual metadata is relational. Labels → images → clusters → scores. Grab top 10 and you lose the graph.

FAILURE MODE 3 OF 3

"Death by Conversation"

Each turn fits. The accumulation doesn't.

50K

Turn 1

Manageable start

250K

Turn 5

Quarter of window gone

1M

Turn 20

Context destroyed

Lost in the Middle (Liu et al., 2023): LLMs ignore the middle. A 50M window still degrades.

THE SOLUTION

"Treat the LLM as an orchestrator, not a database."

The Pass-by-Reference Pattern

Stop treating the LLM as a database that needs to hold all the data. Instead, treat it as a decision-maker that works with summaries and delegates to specialized tools.

✗ Old Thinking

All Data → LLM → Response. Fails at scale.

✓ New Thinking

Summary → LLM → Tools → Data. Full state lives outside context.

THE SOLUTION

Query-Aware Sampling

Not all metadata fields are created equal. The system exploits this ruthlessly to achieve **99.95% token reduction** on high-cardinality fields.

Enum Fields → Include All

camera_type: 3 values = 45 tokens. Send everything.

String Fields → Top 50

label (query: "cat"): 850 tokens vs. 1.9M for all labels.

Numeric Fields → Min/Max

temperature: 38K values → "15.0°C – 42.0°C" = 12 tokens.

1.9M

Naive Tokens

Full serialization

907

Smart Tokens

Query-aware sampling

99.95%

Reduction

High-cardinality fields

The Math

Metric	Naive Approach	Artifact Pattern
Per Query	~50,000 tokens	~2,000 tokens
At Turn 20	1,000,000 tokens	~17,000 tokens*
Feasibility	Impossible	Production-ready

**Prior turns are summarized and compressed — growth is sub-linear, not additive.*

98.3%

Token Reduction

1M → 17K tokens at turn 20

17K

Tokens at Turn 20

Sub-linear via summarization.

"Not faster. Not cheaper. Possible."

THE SOLUTION

The Pattern Generalizes

High-cardinality metadata is everywhere. The architecture is universal.



Genomics

3.2B base pairs, relational annotations.



IoT / Sensor Data

Millions of readings, unbounded cardinality.



E-Commerce Catalogs

100M+ SKUs with attributes.



Security / Log Analysis

Billions of events, accumulates over turns.



Any domain where metadata exceeds context needs this: summarize up, reference down.



Thank you!

gabi@visual-layer.com

[linkedin.com/in/gabikazav](https://www.linkedin.com/in/gabikazav)

Gabi Kazav, VP R&D, Visual layer.