Review

Not peer-reviewed version

Managing the Unmanageable: Multimodal Artificial Intelligence for Unstructured Data Management and Analysis

Chong Ho Yu *, Nino Miljkovic, Zhaoyang Wang

Posted Date: 15 April 2026

doi: 10.20944/preprints202604.1121.v1

Keywords: multimodal AI; RAG; LLM; personal intelligence; neuro-symbolic AI; world model; unstructured data; world model

 doi:10.20944/preprints202604.1121.v1

Review

Managing the Unmanageable: Multimodal Artificial Intelligence for Unstructured Data Management and Analysis

Chong Ho Yu *, Nino Miljkovic and Zhaoyang Wang

Hawaii Pacific University

* Correspondence: cayu@hpu.edu

Abstract

Today, data are no longer confined to numerical values arranged in row-by-column matrices or stored neatly within relational databases. One of the defining characteristics of big data is its high variety, encompassing unstructured and multimodal forms such as text, audio, images, and video. These data types dominate contemporary domains including social media, digital humanities, biomedical research, education, and surveillance systems, yet they remain difficult to manage and analyze using traditional data management architectures. To cope with this shift, modern data management systems must move beyond schema-driven designs and incorporate multimodal artificial intelligence capable of understanding, integrating, and reasoning across heterogeneous data modalities. This article examines how multimodal AI—particularly large multimodal foundation models—can be leveraged to support the ingestion, representation, organization, and analysis of unstructured data. It discusses emerging multimodal data management frameworks, outlines a conceptual pipeline for multimodal data analysis, and highlights key challenges related to scalability, interpretability, and governance. By situating multimodal AI at the core of data management, this work argues that effective data analysis in the era of big data requires systems that treat meaning, context, and cross-modal relationships as first-class computational objects rather than afterthoughts.

Keywords: multimodal AI; RAG; LLM; personal intelligence; neuro-symbolic AI; world model; unstructured data; world model

1. Introduction

Traditionally, data have been equated with numerical representations, reflecting a realist assumption that reality can be objectively measured, quantified, and modeled [66,69]. Qualitative researchers have long challenged this view, arguing that many complex human phenomena—such as lived experience, meaning-making, consciousness, and social interaction—cannot be adequately reduced to numbers without significant loss of depth and context. This critique is closely aligned with phenomenology, particularly the work of Edmund Husserl, who argued that modern science had become detached from the lifeworld, the pre-theoretical world of everyday experience that gives phenomena their meaning [28]. From this perspective, numerical data may describe how much or how often, but they often fail to capture how it is experienced. Later qualitative methodologists extended this insight, emphasizing that human realities are socially constructed, context-dependent, and interpretive, requiring methods such as in-depth interviews, participant observation, and textual analysis to access meanings that resist quantification [12,55]. Rather than rejecting numbers outright, qualitative scholars argue for epistemological pluralism: numeric data are powerful for certain questions, but understanding complex phenomena—identity, suffering, belief, learning, or ethical judgment—demands approaches that honor subjectivity, intentionality, and lived experience alongside measurement [67].

 doi:10.20944/preprints202604.1121.v1

For much of the twentieth century, critiques of quantitative dominance were often ignored or minimized because numerical methods were widely perceived as more scientific, objective, and rigorous, aligning neatly with positivist ideals of measurement, replicability, and statistical inference [66]. However, this hierarchy has been increasingly destabilized by the explosive growth and practical importance of unstructured data—including text, audio, images, and video—which do not naturally conform to tabular or numeric formats.

For example, in industry cybersecurity now relies heavily on biometric modalities such as facial recognition, voice authentication, and behavioral patterns, all of which originate as rich, non-numeric signals that must be interpreted rather than simply counted [31,32]. Customer support systems routinely archive massive volumes of chat transcripts, call recordings, screenshots, and photos uploaded by users, where meaning, intent, and context are embedded in language and visuals rather than in predefined variables. Beyond these examples, healthcare increasingly analyzes clinical notes, medical images, and radiology scans; autonomous vehicles depend on video streams and sensor fusion; social media analytics centers on posts, images, and videos; and legal and financial sectors mine contracts, emails, and compliance documents for patterns and risk signals [15]. These developments have forced a conceptual shift: Rigor is no longer equated solely with numerical abstraction, but with the ability to systematically analyze meaning-rich, high-dimensional, and context-dependent data. As a result, qualitative insights and interpretive methods—now often augmented by machine learning and AI—are being revalued as essential components of scientifically robust inquiry rather than as inferior alternatives to quantitative analysis.

As such, the digital ecosystem is currently navigating a profound structural inflection point, characterized by the transition from deterministic data management—predicated on structured, tabular information—to probabilistic systems of reasoning capable of ingesting and interpreting the chaotic reality of unstructured media [22]. For decades, the enterprise data stack was architected around the Relational Database Management System (RDBMS) and structural query language (SQL), a paradigm that, while efficient for transactional records, inherently excluded the vast majority of human-generated information. Estimates consistently indicate that approximately 80% of enterprise data exists in unstructured forms: video footage, audio recordings, complex PDF contracts, satellite imagery, and social media interactions [61].

Historically, these virtually "unmanageable" data were relegated to "data swamps"—passive object storage accessible only via limited metadata or brittle, specialized computer vision pipelines that required extensive manual feature engineering. With the advance of text mining, which is based on computational linguistics and natural language processing, analysts could analyze open-ended data using tools like SAS Text Miner, JMP Pro, and IBM SPSS Modeler [67–72]. However, the modality is still limited to textual data, rather than audiovisual data.

To a substantial degree, NoSQL (Not Only SQL) did alleviate the structural rigidity of the RDBMS, but it solved the problem of storage and schema rather than the problem of meaning. While the RDBMS required data to be normalized into strict rows and columns (a "schema-on-write" approach), NoSQL introduced schema-on-read. This allowed enterprises to ingest unstructured and semi-structured data—like JSON documents, sensor logs, and social media feeds—without first redesigning a rigid database schema [25]. While NoSQL solved the structural limitation (how to store a PDF or a log file), it did not inherently solve the semantic limitation. Specifically, a NoSQL database can store a million video files or medical notes, but it cannot "read" them to understand their content. Let alone analysis them.

The maturation of Large Multimodal Models (LMMs) and Multimodal Large Language Models (MLLMs) has fundamentally inverted this dynamic [49]. These foundation models act as universal interpreters, projecting disparate modalities—text, image, audio, video—into a shared, high-dimensional semantic space [52]. This capability transforms unstructured data from a static liability into a dynamic asset, enabling "any-to-any" retrieval and analysis. We are moving from "Systems of Record," which capture what happened, to "Systems of Reasoning," which can analyze why it happened by synthesizing insights across modalities [59].

 doi:10.20944/preprints202604.1121.v1

This comprehensive review provides an exhaustive technical analysis of this paradigm shift. It examines the emerging architectural frameworks, specifically the evolution from static Retrieval-Augmented Generation (RAG) to dynamic, autonomous Agentic RAG workflows [41,58]. It evaluates the comparative performance of frontier models—including Google's Gemini, OpenAI's GPT, and Anthropic's Claude—across critical enterprise modalities, ranking them by reliability, reasoning depth, and user-friendliness. Furthermore, this review dissects the formidable infrastructure challenges accompanying this shift, including the scalability of vector search, the interpretability of opaque embeddings, and the rigorous governance required to manage bias and Personally Identifiable Information (PII) in probabilistic outputs [11]. Last but not least, use cases of Gemini's person intelligence is discussed to illustrate how multimodal AI could be utilized as an ecosystem that synthesizes diverse data, including structured and unstructured data.

2. From Unimodal to Multimodal

2.1. Limitations of the Unimodal Era

The history of data management is largely a history of enforcing structure upon chaos. The "Big Data" era of the early 2010s expanded the volume and velocity of data ingestion but struggled to address variety effectively. Traditional analytical workflows created deep silos: text was processed by Natural Language Processing (NLP) pipelines, images by Computer Vision (CV) Convolutional Neural Networks (CNNs), and tabular data by SQL engines. These silos were hermetically sealed; a sentiment analysis model processing customer reviews had no mathematical way to "see" the image of the damaged product attached to the same support ticket [52].

This unimodal fragmentation resulted in "brittle" workflows. Engineering a pipeline to analyze traffic footage, for example, required training specific object detectors for cars, pedestrians, and traffic lights. If the requirement changed to detecting "erratic driving behavior," the entire pipeline often had to be re-engineered. The cost of this feature engineering meant that only the most high-value unstructured data were ever analyzed, leaving the "long tail" of enterprise information dormant in cold storage [61].

2.2. The Rise of the Multimodal Foundation Model

The advent of the Transformer architecture and subsequent Large Multimodal Models (LMMs) introduced a unifying substrate: the embedding. Unlike traditional feature vectors, which represent specific attributes (e.g., "contains_red_pixel"), multimodal embeddings represent semantics in a continuous vector space. Models like CLIP (Contrastive Language-Image Pre-training) were trained on billions of image-text pairs to minimize the geometric distance between an image and its accurate textual description [2]. Further, advanced algorithms are utilized to compress vision token, making multimedia storage and retrieval viable [34].

This mathematical alignment implies that in the vector space, the numerical representation of a photograph of a golden retriever is proximal to the numerical representation of the text string "dog playing in the grass." This "Any-to-Any" capability allows for cross-modal interoperability without explicit metadata tagging. A user can search a video archive using a text query ("Find the moment the CEO gives a demo of the robot"), or search a document database using an image query ("Find contracts with this specific logo") [52].

The implications for data management are radical. The database is no longer just a storage engine; it becomes a dynamic perception engine or an engine of pattern recognition. By integrating LMMs, databases can "watch" video and "listen" to audio, indexing the semantic content dynamically. This creates a "Multimodal Lakehouse" where raw data files and their semantic indices coexist, accessible via natural language queries rather than complex code [38].

 doi:10.20944/preprints202604.1121.v1

2.3. From Systems of Record to Systems of Reasoning

The ultimate trajectory of this technology is the creation of Systems of Reasoning. A System of Record answers questions like "What was the total revenue in Q3?"—a deterministic lookup and a typical quantitative question. A System of Reasoning answers questions like "Why did revenue drop in Q3, considering the customer sentiment in support calls and the visual evidence of supply chain disruption in our logistics video feeds?" This query goes beyond superficial numbers to qualitative insight [40].

Achieving this requires more than just storage; it requires an architecture that combines memory (vector databases), perception (multimodal embeddings), and logic (agentic workflows). The following sections detail the conceptual pipeline required to build such a system. The implications for data management are radical. The database is no longer just a storage engine; it becomes a perception engine. By integrating LMMs, databases can "watch" video and "listen" to audio, indexing the semantic content dynamically. This creates a "Multimodal Lakehouse" where raw data files and their semantic indices coexist, accessible via natural language queries rather than complex code [38]. Table 1 summarizes the conceptual shift from "Systems of Record" to "Systems of Reasoning".

Table 1. Evolution of Data Management Paradigms.

Feature	RDBMS	Multimodal AI
Data Focus	Structured (Table/Row)	Unstructured (Video/Audio/Image/Text)
Approach	Schema-on-write	Schema-on-read / Embedding
Query Type	Keyword / ID Lookup	Semantic / Concept Search
System Goal	Systems of Record	Systems of Reasoning
Function	Static Storage	Dynamic Perception and Pattern Recognition
Inference	Deterministic	Probabilistic

3. The Architecture and Process of Multimodal Intelligence

To operationalize multimodal AI, organizations must transition from ad-hoc data science projects to robust, scalable engineering pipelines. The emerging reference architecture for this is the AI-Native Data Platform, which fundamentally rethinks the traditional Extract-Transform-Load (ETL) process.

3.1. Stage 1: Ingestion and the "ObjectRef" Paradigm

The first challenge in multimodal data management is the physical disconnect between analytical engines (Data Warehouses) and unstructured storage (Data Lakes/Object Storage). Historically, a data warehouse like Snowflake or BigQuery stored metadata (filenames, timestamps), while the actual.mp3,.mp4 or.pdf files resided in AWS S3, Microsoft Azure, or Google Cloud Storage. Analyzing the file content required moving data out of the warehouse, processing it, and writing results back—a slow, expensive, and insecure round-trip.

Several methods have been developed to ease this bottleneck, such as SAS's in-memory analytics [68] and edge computing [65]. A critical innovation addressing this is the ObjectRef data type, exemplified by recent updates to Google BigQuery [40]. ObjectRef allows the SQL engine to treat an unstructured file in cloud storage as a first-class citizen within a table. A table row can contain a customer ID (integer), a transaction date (timestamp), and an ObjectRef pointing to the audio file of the customer's support call.

This architectural unification enables Zero-ETL analysis. Users can run SQL queries that invoke remote functions (powered by Vertex AI or similar services) to process the referenced object directly. For instance, a query could look like:

 doi:10.20944/preprints202604.1121.v1

SELECT customer_id, ML.TRANSCRIBE(audio_file) FROM support_calls WHERE duration > 300

This integrates unstructured inference directly into the structured query planner, reducing data movement and simplifying governance [40].

For organizations without the resources to build custom ingestion pipelines, managed services like Amazon Bedrock Data Automation have emerged. These services act as an intelligent ingestion layer that automatically identifies file types, parses content (e.g., extracting text from PDFs, transcribing audio), and prepares it for downstream indexing [61]. This abstracts away the complexity of managing libraries like ffmpeg or Tesseract, providing a standardized JSON schema output for every ingested asset.

3.2. Stage 2: Representation and Embedding

Once ingested, data must be converted into a format suitable for machine reasoning. This is the domain of Embeddings. The core mechanism of multimodal representation is the high-dimensional vector. Modern embedding models, such as Amazon Nova Multimodal Embeddings or Google's multimodal gecko models, are designed to project text, images, and video into a single semantic space [52]. The dimensionality of these vectors typically ranges from 768 to 4096 dimensions. Higher dimensionality captures more nuance but increases storage and search costs.

A key innovation is the move towards Cross-Modal Embeddings that do not require an intermediate text step. Early multimodal systems often captioned an image and then embedded the caption. True multimodal models embed the image directly, capturing visual features (texture, lighting, spatial composition) that text descriptions often miss [52].

One of the most complex aspects of multimodal representation is Chunking. A vector represents a single semantic unit. A 20-page PDF or a 1-hour video cannot be compressed into a single vector without massive information loss. Therefore, data must be segmented as follows [57]:

1. Text Chunking: Moves beyond fixed-size windows to Semantically Coherent Chunking (sentence/paragraph level) and Adaptive Chunking (using LLMs or "Late Chunking" to preserve document-level context).
2. Image Chunking: Discusses Patch-based (fixed grids), Object-centric (using detectors like YOLO), and Scene Graph approaches which preserve the relationships between visual elements.
3. Audio Chunking: Covers Silence-based segmentation and Speaker-based (diarization) chunking, highlighting that splitting audio during a speaker's turn can blend semantic signals.
4. Video Chunking: Explores Shot-based and Action-based detection to ensure temporal segments capture complete events rather than fragmented frames.
5. Cross-Modal Alignment: The "frontier" of chunking, which focuses on synchronizing disparate data types (e.g., ensuring a text description stays linked to the specific video timestamp it describes).

3.3. Stage 3: The Vector Storage Infrastructure

The explosion of vector data necessitates specialized storage infrastructure. For example, LanceDB offers a serverless, embedded architecture. It utilizes the Lance columnar format, which is optimized for machine learning data. Unlike traditional databases that require a running server, LanceDB can run in-process (like SQLite) and query data directly from S3. This "Multimodal Lakehouse" approach drastically simplifies the stack for developers, as the vector index and the raw data live together [38].

 doi:10.20944/preprints202604.1121.v1

4. Retrieval-Augmented Generation

4.1. What is RAG?

Retrieval-Augmented Generation (RAG) is an architectural framework in artificial intelligence that enhances the capabilities of generative foundation models by integrating an external retrieval mechanism. It has become the standard pattern for grounding LLMs. In a multimodal context, RAG involves retrieving relevant images/video clips and feeding them into the context window of an LMM. While traditional foundation models generate responses based solely on the patterns learned during their training phase, RAG-enabled systems first search an external database or corpus to find relevant information related to a specific query. This retrieved content is then provided as additional context to the generative model, which synthesizes the final output. This process ensures that the AI's responses are not only based on its parametric memory, but are also grounded in verifiable, up-to-date, or domain-specific evidence, thereby significantly reducing the risk of "hallucinations" [33,50,63].

A critical challenge in modern AI is the gap between raw, unstructured multimodal data—such as high-resolution images, audio recordings, and video streams—and the sophisticated "understanding" or reasoning capabilities of foundation models. RAG acts as the functional bridge between these two domains. By transforming raw multimodal artifacts into searchable semantic representations (often through multimodal embeddings), RAG allows foundation models to "see" and "hear" specific data points within a vast database. Instead of the model attempting to memorize every detail of a multimodal corpus, the RAG bridge selectively feeds the model only the most relevant snippets of data required for a specific reasoning task, ensuring a streamlined and context-aware inference process [35,54,73].

While foundation models possess extensive "internal" knowledge—a static repository of information encoded within their billions of parameters during pre-training, this knowledge is often fixed in time and lacks the granularity required for specific, data-intensive tasks. To alleviate the issue, RAG provides the necessary "external" context drawn from a multimodal database, encompassing text, audio, video, and image data. This external context grounds the AI's reasoning in the actual data being managed. For example, in medical imaging or radiology, a model might use its internal knowledge of anatomy but rely on RAG to retrieve specific, similar cases from a multimodal database to provide a grounded and evidenced interpretation of a new scan. This distinction transforms the model from a generalist into a specialized agent capable of reasoning over private or highly dynamic datasets [54,65].

Traditionally, data management systems have functioned as passive storage units where information is retrieved using specific "row" identifiers or keyword matches. RAG fundamentally transforms these systems into active retrieval systems. By utilizing semantic vector search, RAG enables users and AI agents to query a database based on "concepts" or "semantic meaning" rather than exact string matches. This shift allows for the discovery of relationships across different modalities—such as finding a video segment that semantically matches a textual description of an event—without the need for manual tagging or rigid indexing. Consequently, data management evolves from a simple archival process into a dynamic, semantic layer that actively supports the reasoning and decision-making capabilities of AI systems [43].

4.2. Limitations of Standard RAG

However, this conventional or standard RAG approach has significant limitations.

1. Context Window Bottlenecks: Although powerful AI models can handle many tokens, filling the context window with massively retrieved chunks ("Context Stuffing") can increase noise, degrade reasoning performance and increase latency/cost [53]. The "Lost in the Middle" phenomenon means models often forget information buried in the center of a large prompt [44].

2. Textual Dominance: The internal "space" where MLLMs store information is overwhelmingly biased toward text. Even when processing images, the model's neurons focus mostly on linguistic

 doi:10.20944/preprints202604.1121.v1

concepts, which suppresses the fine-grained visual details necessary to distinguish one image from another during retrieval [17].

3. The "Alignment" Trap: MLLMs spend most of their computational "energy" trying to bridge the gap between images and text (making them speak the same language). While this is great for generating captions, it homogenizes the embeddings. This means different images start to look mathematically similar to the model, reducing the "discriminative power" needed to pick the exact right match from a list [17].

4.3. The Rise of Agentic RAG

To overcome these limitations, the industry is shifting toward Agentic Architectures. Agentic RAG is characterized primarily by autonomous decision-making and goal-oriented behavior. Unlike older RAG, which follows a rigid, linear sequence of fetching and then generating, Agentic RAG acts as a reasoning-capable assistant that can dynamically plan its approach. Key characteristics include the ability to break complex queries into sub-tasks, use diverse tools—such as web search or specialized APIs via the Model Context Protocol (MCP)—and iteratively refine its results until a sufficient answer is reached [36].

Naive or standard RAG paradigms rely on static, "one-shot" retrieval. If the initial search fails to find the right information, the system produces an incomplete or hallucinated response. Agentic RAG uses Planning and Routing to decompose complex queries into smaller sub-tasks. It can decide in real-time whether to query a structured SQL database, a vector store, or a live web search tool based on the specific needs of the prompt [58].

To combat the "garbage in, garbage out" problem, Agentic RAG implements ReRanking. It generates multiple candidate answers from diverse contexts and uses a specialized Ranking Agent to evaluate and pick the most accurate response, rather than just relying on the top-k vector search results. Further, while older systems are confined to a static knowledge base, the agentic version can identify gaps in its own information. If a document index lacks current data, the agent can autonomously decide to use a web search tool or a live API to supplement its context. More importantly, agentic RAG introduces "reflection," allowing the system to check its own work. If a retrieved chunk is irrelevant or a tool call fails, a Fallback Agent can switch strategies mid-stream, a level of flexibility that traditional, sequential pipelines cannot achieve [36,58]. Figure 1 illustrates the workflow of both the standard/simple version and the advanced version of RAG.

 doi:10.20944/preprints202604.1121.v1

Figure 1. Standard/Naive and Advanced Agentic RAG Frameworks.

4.4. Relationship Between RAG and Conventional Databases

The growing capabilities of multimodal artificial intelligence (AI) systems and the rapid advancement of retrieval-augmented generation (RAG) architectures do not necessarily imply a diminishing role for traditional structured databases, such as SQL-based systems. On the contrary, these developments underscore the continued importance of structured data management as a foundational component of trustworthy AI systems.

Large language models (LLMs), when used in isolation, operate by generating responses based on probabilistic patterns learned from vast corpora of text. Although these models demonstrate remarkable fluency and general knowledge, they are prone to factual inaccuracies and hallucinations, particularly when responding to queries involving proprietary, domain-specific, or time-sensitive information [41]. RAG addresses this limitation by enabling LLMs to retrieve relevant information from external, authoritative knowledge sources at inference time. Retrieved passages from databases, document repositories, or knowledge bases are injected into the model's prompt, thereby grounding generation in verified context rather than relying solely on internal model parameters.

This grounding mechanism substantially improves factual accuracy and reliability, making RAG-based systems suitable for enterprise and mission-critical applications, including financial reporting, internal policy support, and clinical documentation, where creativity is undesirable and errors carry significant consequences [48]. In such contexts, the integrity of the underlying data pipeline becomes as important as the sophistication of the AI model itself.

It is at this juncture that structured data management approaches, particularly SQL-based systems, reassert their relevance. If an AI model functions as the cognitive engine of an intelligent system, the data pipeline serves as the infrastructure that ensures information is delivered accurately, securely, and consistently. Within this pipeline, SQL plays a central role in enforcing constraints, validating transactions, and mediating access to trusted data sources. In RAG workflows, data pipelines must both prepare data—through cleaning, normalization, and governance—and retrieve it efficiently during inference. Modern SQL platforms increasingly support hybrid querying paradigms that combine traditional relational filters (e.g., temporal constraints, identifiers, and access controls) with semantic retrieval mechanisms based on vector embeddings [48].

Although platforms such as SQL Server continue to evolve in their native support for vector search, they are clearly moving toward tighter integration with AI workloads, often within architectures that pair relational databases with specialized vector or search services. The central issue is not whether SQL systems replace vector databases, but rather that SQL anchors semantic retrieval within an auditable, secure, and well-governed data foundation. This anchoring is essential for maintaining transparency, compliance, and accountability in enterprise AI systems.

As multimodal AI systems become increasingly capable of interpreting images, audio, and video, the relevance of SQL does not diminish. Organizations will continue to require reliable "sources of truth" for operational, regulatory, and financial data. Domains such as healthcare, finance, and public administration demand data that are consistent, traceable, and ACID-compliant. ACID compliance refers to the guarantee that database transactions uphold atomicity, consistency, isolation, and durability, ensuring that operations remain correct, predictable, and auditable even under concurrent access and system failures [24].

While analysts may rely on multimodal models to interpret medical images or summarize customer service interactions, they will still depend on SQL databases to manage patient histories, billing records, contracts, and access permissions with precision and regulatory compliance.

As AI agents become more autonomous, the role of data professionals is likely to shift from manual query construction via SQL and other conventional tools toward higher-level system design. Practitioners will increasingly act as AI orchestrators, responsible for defining schemas, constraints, views, and retrieval strategies that guide intelligent systems toward safe, reliable, and interpretable behavior. Rather than fading into obsolescence, SQL and structured data management approaches

 doi:10.20944/preprints202604.1121.v1

will remain indispensable tools for imposing order, governance, and clarity in an increasingly probabilistic AI landscape.

5. The Model Landscape: A Comparative Analysis

The "brain" of the multimodal pipeline is the Foundation Model. Currently the landscape is dominated by a fierce competition between Google, OpenAI, and Anthropic, each with distinct architectural philosophies.

5.1. Video Analysis: The Battle for Context

Video is the ultimate stress test for multimodal models due to its temporal dimension and massive data volume. Google Gemini is currently the leader for long-form video understanding. Its defining feature is its 1 million token context window. This allows users to feed entire movies, hour-long surveillance tapes, or massive codebases into the prompt without sophisticated chunking or subsampling. In the Video-MME benchmark, which tests temporal reasoning and long-term event tracking, Gemini achieved a state-of-the-art accuracy of 75.0%. It significantly outperforms competitors in "Needle-in-a-Haystack" retrieval, capable of finding a specific visual event (e.g., "a red balloon floating past") in hours of footage with >99% recall [18]. Armed with this multimodal capability, Gemini is ideal for archival search, media asset management, and security analytics where "skipping frames" is not an option.

Another formidable contender in this area is OpenAI GPT. While its context window (128k) is smaller than Gemini's, it excels in short-horizon tasks. It scores 71.9% on Video-MME [18]. However, its strength lies in its World Knowledge and integration with tools. For generic tasks ("What is happening in this 30-second clip?"), it is often faster and more conversational than Gemini. Open AI GPT is ideal for real-time monitoring, customer-facing video bots, and applications where latency is more critical than analyzing hour-long duration [14].

5.2. Document and Visual Reasoning: Precision and Logic

For tasks involving the extraction of data from complex documents (invoices, P&IDs, scientific charts), the ability to reason visually is paramount. As such, Claude has emerged as the "Engineer's Choice" for visual reasoning. It dominates the ANLS (Average Normalized Levenshtein Similarity) benchmark for Document Visual QA with a score of 95.2%, surpassing GPT-4o (92.8%) [10]. Specifically, Claude 3.5 Sonnet is uniquely capable of understanding the structure of visual information. It can look at a screenshot of a software UI and write the exact React code to reproduce it (the "Artifacts" workflow). It is less prone to "OCR hallucinations" than other models. Taking the above into account, Claude is one of the best for IDP (Intelligent Document Processing), automating software engineering workflows, and scientific data extraction [5].

5.3. Audio Intelligence: Beyond Transcription

Gemini is good at Native Audio Understanding. Unlike most "Speech-to-Text" models (like Whisper) that transcribe audio to text and then analyze the text, Gemini processes audio natively. This means it "hears" the audio; it can detect sarcasm, hesitation, emotion, and background noises (e.g., "The speaker sounded anxious and there was a siren in the background"). This information is lost in transcription [22]. Further, it can perform text-content-aware speaker detection without an external diarization model. This approach is beneficial to call center sentiment analysis, security audio monitoring, and complex meeting summarization [6]. Nonetheless, Whisper (OpenAI) remains the gold standard for pure transcription cost-efficiency. It is robust to noise and accents but lacks the reasoning layer of Gemini. For pure "records of conversation," Whisper is sufficient; for "analysis of conversation," Gemini is superior [6]. Table 2 summarizes functionalities and use cases of major frontier models in terms of multimodality.

 doi:10.20944/preprints202604.1121.v1

Table 2. Comparative Analysis of Frontier Multimodal Models.

Model	Primary Strength	Key Benchmark	Ideal Use Case
Google Gemini	Long-form Video / Audio	Video-MME (75.0%)	Media Archival, Security Analytics
OpenAI GPT	World Knowledge / Speed	Video-MME (71.9%)	Real-time Monitoring, Chatbots
Anthropic Claude	Visual Reasoning / Precision	ANLS DocVQA (95.2%)	IDP, Code Generation, Scientific Data

6. Governance and Operational Challenges

The transition to probabilistic Systems of Reasoning introduces risks that deterministic Systems of Record never faced. The major issue is that scalability in multimodal systems is non-linear. As a result, the user might suffer from the consequences below:

1. Indexing Costs: Storing billions of 4096-dimension vectors requires immense RAM. The "Curse of Dimensionality" means that as dimensions increase, distance metrics (like Euclidean distance) become less meaningful, requiring more expensive computations to find true nearest neighbors [13].

2. Latency vs. Recall: Approximate Nearest Neighbor (ANN) algorithms (like HNSW) trade accuracy for speed. In high-stakes domains (e.g., legal discovery), missing a relevant document due to index approximation is unacceptable. Systems must balance this by using Re-Ranking—retrieving a larger set of candidates quickly (ANN) and then using a heavy cross-encoder model to re-rank them accurately [20].

3. Data Gravity: Moving terabytes of video archives from on-premise storage to the cloud for embedding is often cost-prohibitive due to egress fees. This is driving Edge AI, where models (like Gemini Nano) run locally to filter and embed data at the source, sending only the vectors (metadata) to the cloud [29].

6.1. Interpretability, Hallucination, and Independent Modality Verification

The "Black Box" nature of embeddings is a major barrier to adoption in regulated industries. For example, why did the model think this video of a "handshake" was similar to "corruption"? The embedding space does not provide an explanation. Techniques like Concept-Based Explanations are emerging to probe these spaces, but they are in their infancy. In multimodal RAG, models can "hallucinate" details that are not in the retrieved context, especially when the visual input is cluttered or low-resolution. Studies show that hallucination rates in citation generation can be significant. Implementing "Citation Evaluation" loops—where a second model verifies that the generated claim is actually supported by the retrieved pixel/text data—is becoming a standard architectural pattern [2,37].

To mitigate the risk of stochastic errors and "hallucinations" in complex AI systems, researchers increasingly advocate for Independent Modality Verification (IMV) as a framework for high-assurance decision-making. Under an IMV architecture, disparate data streams—or "modalities"—operate as isolated sensory inputs to validate a single proposition, ensuring that the system's confidence is grounded in multi-source consensus rather than a single, potentially corrupted pipeline [7,30].

The core strength of IMV lies in its structural decoupling. In traditional multimodal fusion, features are often integrated early in the process ("Early Fusion"), which allows a failure in one input to propagate through the entire neural network. In contrast, IMV utilizes a Late Fusion or Decision-Level Fusion approach. This ensures that the manipulation or "noise" affecting one modality does not automatically bias the other, providing a robust defense against adversarial attacks and sensor failure [56].

As specified in the three-component model, a typical IMV workflow consists of [3,9]:

 doi:10.20944/preprints202604.1121.v1

1. Modality A (e.g., Facial Recognition): Captures and processes visual biometric data.
2. Modality B (e.g., Voice Recognition): Captures and processes auditory/linguistic data.
3. Verification Logic: A symbolic or rule-based controller that evaluates the outputs.

For the system to authorize a decision, both modalities must independently exceed a predefined confidence threshold and reach a congruent conclusion. This creates a "redundant safety" mechanism similar to those found in aerospace and medical diagnostics.

Despite its security benefits, IMV introduces significant computational overhead [42]. Running parallel inference pipelines increases latency and hardware requirements, which may be prohibitive for edge computing devices with limited power budgets. Furthermore, the implementation of IMV is governed by the principle of Adaptive Security. For low-risk applications—such as unlocking a mobile device for non-sensitive tasks—the "convenience-to-security ratio" favors single-factor authentication. High-rigor IMV is typically reserved for "High-Stakes AI" environments, such as autonomous driving or financial transactions, where the cost of a false positive is catastrophic [46,60]. The workflow of IMV is illustrated in Figure 2.

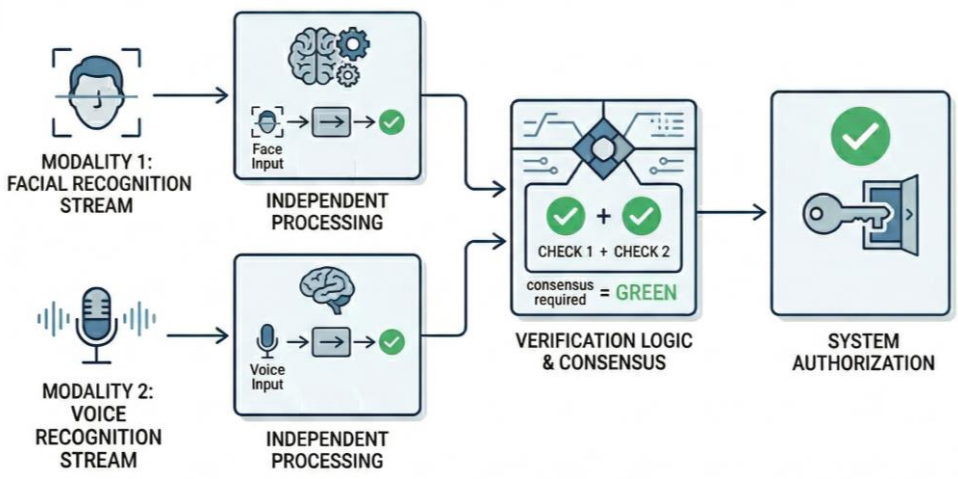

Figure 2. Independent Modality Verification (IMV) Framework.

6.2. Governance: PII, Bias, and Unlearning

It is a misconception that vectors are anonymous. Research has shown that original text (personally identifiable information [PII] including social security numbers or names) can be reconstructed from their embeddings via "inversion attacks" [62]. To rectify the situation, organizations must implement PII Masking (using tools like Microsoft Presidio) before the data are embedded [11]. Additionally, Private Information Retrieval (PIR) protocols are being explored to allow querying vector databases without revealing the query itself.

Further, Models like CLIP inherit biases from their internet-scale training data. They might associate images of doctors primarily with men. To remediate the problem, modern algorithms allow engineers to identify specific dimensions in the vector space that correlate with protected attributes (gender, race) and "prune" or re-weight them, reducing bias without retraining the entire model [2].

Another issue is the Right to Be Forgotten (Unlearning) demanded by Europe's GDPR. If a user revokes consent, their data must be removed. Deleting the file is easy; deleting the influence of that file on a trained model is hard. Proposed solutions include SISA (Sharded, Isolated, Sliced, and Aggregated) training. By training the model on isolated shards of data, a deletion request triggers the retraining of only one small shard, making "Exact Unlearning" computationally feasible (European Data Protection Supervisor, 2025).

 doi:10.20944/preprints202604.1121.v1

7. Future Horizons

7.1. World Models and Any-to-Any Training

The field is evolving rapidly beyond static perception. Put it bluntly, mastering Python, SQL, Oracle or basic LLMs is no longer sufficient to manage multimodal data in the age of diverse big data. While current models predict the next token, future "World Models" will predict the state of the world. To actualize the vision of world models, Yann LeCun founded Advanced Machine Intelligence (AMI) Labs, focusing on the Joint-Embedding Predictive Architecture (JEPA) [51]. Unlike LLMs that predict pixels or words, JEPA predicts the state of the world in an abstract representation space. These models are trained primarily on video and sensory data rather than just text, allowing them to learn "common sense" and basic physics. Doing so can enable them to ignore "noise" (like the exact texture of a leaf) and focus on "signal" (like the trajectory of a falling object). They will understand physics, causality, and object permanence, allowing them to simulate outcomes ("What happens if I drop this glass?") rather than just describing pixels [39,45]. This approach is crucial for robotics and autonomous systems [47]. In addition, newer approaches are moving away from models trained on fixed pairs (image-text) to models trained on unified streams of data. "Any-to-Any" models will be able to translate between any modalities directly (e.g., humming a melody to generate a violin score) without passing through a text bottleneck, leading to more fluid and human-like interaction.

7.2. Neuro-Symbolic Integration and GraphRAG

Another trend in data management for unstructured data is the reintegration of symbolic AI. While Neural Networks (Deep Learning) are excellent at perception (identifying a cat in a video), they are poor at strict logic and rule adherence. Neuro-Symbolic AI combines the best of connectionism and symbolism [26], which is manifested in GraphRAG. Instead of just retrieving vector chunks, GraphRAG approach uses an LLM to extract nodes (entities) and relationships from the data to build a Knowledge Graph. Retrieval then involves traversing this graph. Because the relationships are explicit, the reasoning behind an answer is traceable and thus explainability is greatly enhanced [13,27]. For example, in a financial investigation, vector search might find documents mentioning "Company A" and "Company B." A Knowledge Graph can reveal that "Director X" sits on the board of both, a relationship that might be three hops away and invisible to vector similarity. GraphRAG combines the "fuzzy" search of vectors with the "structured" pathfinding of graphs.

7.3. Personal Intelligence (PI) and Resource Integration

Google's Personal Intelligence (PI), often referred to as the "PT" (Personal Technology) layer of the Gemini ecosystem, represents a paradigm shift from general-purpose chatbots to context-aware, individualized reasoning systems. PT is not a data management system per se; nonetheless, PI is powered by the Gemini 3 architecture—a unified, multimodal transformer model designed to process information across multiple Google accounts and applications, which can be used as data sources [21]. By granting PI access to an ecosystem that includes Gmail, Google Drive, Google Photos, and YouTube, the system builds a "knowledge graph." This allows the AI to reason across silos, connecting a flight confirmation in Gmail to a presentation slide in Drive and a captured photo in Google Photos. Unlike previous iterations of AI that treated each query as an isolated event, PT functions as a continuous memory-based assistant that "remembers" the user's professional context, research interests, and digital history [23].

The transition from PT to Research Data Intelligence (RDI) is a significant game-changer for researchers. Traditionally, managing complex datasets required specialized technical skills in relational databases, such as SQL, or non-relational structures, such as NoSQL. PT eliminates this barrier by allowing researchers to simply upload multimodal data—including Word documents, PDFs, PowerPoint slides, spreadsheets, PNG/JPEG images, and MP3/MP4 files—directly to Google

 doi:10.20944/preprints202604.1121.v1

Drive. In many instances, even the act of "uploading" is automated. For example, photos taken during fieldwork are synchronized to Google Photos, and a researcher's YouTube viewing history—such as watching documentaries on global policy—is automatically indexed as a searchable context [21]. This creates a frictionless "living dataset" where the data management is handled by the ecosystem's native synchronization rather than manual entry.

The true power of PT lies in its native multimodal AI capabilities. While conventional systems rely on human-generated labels, tags, or text-based transcripts to "understand" media, Gemini utilizes a unified transformer architecture that "sees" image pixels, "listens" to audio waves, and "watches" video frames directly [4]. By converting visual and auditory signals into mathematical tokens—the same language it uses for text—the model achieves a much higher level of nuance. It can identify the architectural style of a building in a photo or the frustrated tone in an interview subject's voice without needing a text-based description first. This leads to significantly higher accuracy in retrieval and analysis compared to traditional OCR or basic speech-to-text methods [21].

8. Discussion: Use Case

8.1. Textual Data and Ethics Research

The first author of this article is a professor of Data Science and AI Ethics. In this case, PT facilitates the management of dense philosophical and technical literature. The professor can store thousands of pages regarding the major schools of moral philosophy—such as Eudaimonist, Deontological, and Utilitarian ethics—within a dedicated Google Drive folder. By querying PT, the researcher can instantly synthesize how these frameworks apply to modern "black box" algorithmic transparency. Furthermore, these data can be connected to NotebookLM, an AI-native research tool that allows for even deeper grounded analysis. In this workflow, PT acts as the retrieval engine that fetches relevant passages across multiple files, which NotebookLM then uses to generate citations and structured summaries for an academic paper or grant proposal [23].

8.2. Audio Data and Qualitative Fieldwork

In the realm of qualitative research, PT revolutionizes the analysis of interviews, journals, and other forms of open-ended data. Traditionally, researchers had to manually transcribe audio or rely on third-party services like Otter.ai, Sonix, or Descript. These transcripts were then exported to specialized qualitative software like MAXQDA or IBM Modeler for coding and textual analytics. In the RDI framework, researchers can simply upload MP3 recordings to Google Drive. Then PT can analyze the audio natively, identifying themes, emotional cues, and speaker intent without the intermediate step of generating a separate text file. This is particularly useful for linguistic research or psychological studies where the way something is said (prosody and pitch) is as important as the words themselves [21].

8.3. Visual Data and Ethnography

For sociologists and artists, PT transforms the smartphone into a powerful ethnographic tool. A researcher studying the architectural styles of Honolulu's Chinatown can simply walk through the neighborhood taking photos. Because the device is linked to Google Photos, the images are automatically uploaded and indexed by PT before the researcher even returns to their desk. PT can then be asked to "Find all photos of buildings with Richardsonian Romanesque lava rock" or "Summarize the evolution of signage in the historical district." This automates the "coding" process that previously took weeks of manual labor, allowing the researcher to focus on higher-level cultural synthesis. [23]. Appendix A shows an example of how Google's PT is utilized to retrieve and analyze visual data.

 doi:10.20944/preprints202604.1121.v1

8.4. Video Data and Technical Analysis

Video data management becomes exponentially more efficient when integrated with YouTube and PT. If a researcher wants to perform a comparative study of hardware—comparing GPU, TPU, and LPU architectures—they can search for and view relevant technical videos on YouTube. Because the viewing history is recorded by PT, the AI can "watch" these movies on behalf of the user to provide a technical summary of the performance benchmarks mentioned across different channels. Additionally, a researcher could upload recorded technical demonstrations to Drive; PT can then identify specific moments where a system failure occurred or summarize the visual steps of a complex software installation, bridging the gap between instructional video and technical documentation [21]. Appendix B shows an example of using PT to retrieve and analyze video data from YouTube.

8.5. Systems of Reasoning

The integration of Google's Personal Intelligence (PI) into the enterprise data stack represents the ultimate maturation of the Big Data promise. By leveraging foundational, cross-account reasoning, organizations can finally unlock the 80% of their data that has remained "dark" and unstructured. However, this is not a "plug-and-play" upgrade. It requires a fundamental re-architecture of the data platform—one that transitions from passive storage to Research Data Intelligence (RDI). This involves embracing ObjectRef storage, vector indices, and agentic workflows within a rigorous new governance framework that accounts for the probabilistic and opaque nature of neural intelligence. The winners of the next decade will be those who successfully build Systems of Reasoning: platforms that do not just store data, but perceive, understand, and act upon the rich, multimodal reality of the world through a unified digital consciousness.

9. Conclusion

With the rapid advancement of AI algorithms and the growing availability of data across multiple modalities, traditional approaches are increasingly constrained and, in some cases, approaching their practical limits. While multimodal AI and retrieval-augmented generation (RAG) are still evolving, they have already demonstrated sufficient maturity to be considered viable, production-ready technologies. In contrast, paradigms such as neuro-symbolic AI and world models remain largely exploratory, holding promise but not yet achieving the same level of operational readiness. Against this backdrop, it is both timely and strategic to consider extending Google's Personal Intelligence (PI) paradigm to the enterprise level by leveraging corporate accounts as unified repositories for diverse data sources. Currently PI is for personal Google accounts and is not presently available for Workspace business, enterprise, or education users. Nonetheless, in the near future it is conceivable that the same technology can be scaled up to enterprise intelligence.

10. Future Directions

Concretely, PI's ability to bridge traditionally siloed platforms—such as Gmail, Drive, Photos, and YouTube—can be reimagined within an enterprise context to form an integrated knowledge graph. This enables a "Research Data Intelligence" framework in which context is preserved and continuously enriched across multimodal interactions, rather than fragmented across systems. For long-form video archives, Google Gemini 1.5 Pro stands out as a practical solution due to its extensive context window, which allows for sustained analysis of lengthy content without the need for complex segmentation or sliding-window techniques. For legal and technical document processing, Anthropic Claude is particularly well suited, given its strong visual reasoning capabilities and comparatively low hallucination rates within structured outputs such as "Artifacts," making it a reliable option for tasks that demand high precision and interpretability.

From an infrastructure perspective, transitioning toward a multimodal lakehouse architecture is advisable. Platforms such as BigQuery with Vertex AI or vector-native systems like LanceDB enable tight coupling between raw data and vector representations, aligning with the PI model of continuous

 doi:10.20944/preprints202604.1121.v1

and automated data synchronization. This approach reduces the dependency on traditional ETL pipelines, which often introduce latency, redundancy, and operational complexity. At the same time, the challenge of data gravity—where large datasets are costly to move—can be mitigated by adopting edge AI strategies. By relocating computation closer to where data is generated, organizations can process and filter information locally, transmitting only compact representations such as embeddings to centralized systems.

Equally important is the need for robust governance and regulatory compliance. Techniques such as SISA-style data sharding and systematic PII masking provide a scalable foundation for privacy-preserving architectures. Drawing inspiration from the implicit "air gaps" within PI—such as the separation between public-facing YouTube content and private Studio data—organizations can design systems that enforce strict data boundaries while still enabling powerful cross-modal intelligence. This balance between accessibility and control is critical for meeting GDPR and other regulatory requirements.

This discussion is by no means exhaustive, particularly given the rapid pace of innovation in AI. New models, architectures, and best practices continue to emerge on a near-weekly basis. As such, practitioners and researchers alike should actively monitor developments, experiment with evolving techniques, and remain adaptable in order to effectively navigate the shifting landscape of multimodal AI.

Author Bio: Dr. Chong Ho Yu is Professor and Program Director of Data Science and Artificial Intelligence at Hawaii Pacific University. He has a Ph.D. in Measurement, Statistics, and Methodological Studies, and a Ph.D. in Philosophy of Science (Arizona State University, 2007). He is a three-time winner of the SAS faculty scholarship (2016, 2017 SAS Global Forum and 2017 Western Users of SAS Software Conference). In addition, he also won the Distinguished SAS Educator Award in 2021. His research interests include, but are not limited to, text mining, model comparison, ensemble methods, machine learning, data visualization, multimodal AI, philosophical aspects of research methodologies, and STEM education. He is the President of South California Chapter of American Statistical Association. Nino Miljkovic is a Business Analyst and Data Scientist with a Master of Science in Data Science from Hawaii Pacific University, where he graduated with Distinction in 2025. His academic work focused on explainable artificial intelligence in medical imaging, applying techniques such as Grad-CAM, LIME, and SHAP to enhance transparency in deep learning models. Professionally, he has experience in data analysis, machine learning, and decision-support systems, and currently works closely with executive leadership on forecasting and strategic analytics. His research interests include multimodal AI, explainable AI, machine learning systems, the application of AI in high-stakes domains, and economic game theory. Miljkovic has also conducted research on sustainable artificial intelligence and cloud-based systems. He is actively developing projects in retrieval-augmented generation (RAG), CNN-based image analysis, and neural forecasting. Zhaoyang Wang is an architect in artificial intelligence whose work centers on multimodal AI systems and the analysis of unstructured data, with a Master of Science in Data Science from Hawaii Pacific University. Her research interests include multimodal foundation models, retrieval augmented generation (RAG), and the design of AI systems for high stakes and real-world applications. She is particularly interested in how largescale AI models can integrate heterogeneous data modalities—such as text, images, and video—to support reasoning, interpretability, and decision-making.

Author Contributions: Conceptualization: Chong Ho Yu; Validation: Chong Ho Yu, Nino Miljkovic, Zhaoyang Wang; Resources: Chong Ho Yu, Nino Miljkovic, Zhaoyang Wang; Writing – Original Draft Preparation: Chong Ho Yu; Writing – Review & Editing, Chong Ho Yu, Nino Miljkovic, Zhaoyang Wang; Supervision: Chong Ho Yu; Project Administration: Chong Ho Yu.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or available.

Acknowledgments: During the preparation of this manuscript/study, the authors used ChatGPT and Google Gemini for the purposes of brainstorming, information gathering, proofreading, and use case illustration. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

 doi:10.20944/preprints202604.1121.v1

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
RAG	Retrieval-Augmented Generation
LLM	Large Language Model
SQL	Structural Query Language

Appendix A: Example of Using Google's PT to Retrieve and Analyze Visual Data

Prompt: "Please access my account on photos.google.com. Please select only the photos taken in Chinatown, Honolulu, view those photos and summarize their architectural style."

Gemini's answer:

Based on your images, the architectural style of Honolulu's Chinatown can be summarized as a complex, multicultural hybrid of 19th-century Western commercial forms and traditional Asian cultural motifs. Here are the key themes I identified:

1. 19th-Century Western Commercial Architecture

Many of the buildings you photographed (particularly the two-story brick and stone structures) are in the Italianate or Renaissance Revival styles.

- Key Features: You captured high, narrow windows with arched lintels, decorative cornices along the rooflines, and sturdy masonry facades—some using the local "blue stone" (volcanic basalt).
- Significance: These reflect the rebuilding of Chinatown after the devastating fires of 1886 and 1900, showing the influence of Western developers during the territorial era.

2. Traditional Asian Vernacular & Symbolic Motifs

Interspersed with the Western frames are clear markers of the neighborhood's cultural identity.

- Roofs and Ornamentation: I see photos of the classic upturned eaves (swallowtail roofs) and green glazed tiles that characterize traditional Chinese architecture, often found on temples or association buildings.
- Color Palette: There is a recurring use of "luck and prosperity" colors—deep reds for pillars and gold for decorative inscriptions and signage.
- The Dragon Pillars: Your photos of the ornate dragon-wrapped pillars at major intersections or market entrances are iconic examples of modern traditionalism.

3. The "Mixed-Use" Market Vernacular

A significant portion of your photos focuses on the open-air market entrances and street-level storefronts.

- Function-Driven Design: These feature large folding metal gates and recessed entryways, designed to facilitate the high-traffic flow of the produce and fish markets.
- Aesthetic: It's a "lived-in" style where the architecture serves as a backdrop for vibrant, hand-painted signage and the display of goods, creating a dense, textured urban environment.

4. Historical "Time-Capsule" Elements

You also caught some of the faded signage and ghost murals on the sides of older buildings. These provide a visual timeline of the neighborhood's economic history, showing how storefronts have transitioned from early 20th-century saloons to modern-day boutiques and galleries (see Figure A1).

 doi:10.20944/preprints202604.1121.v1

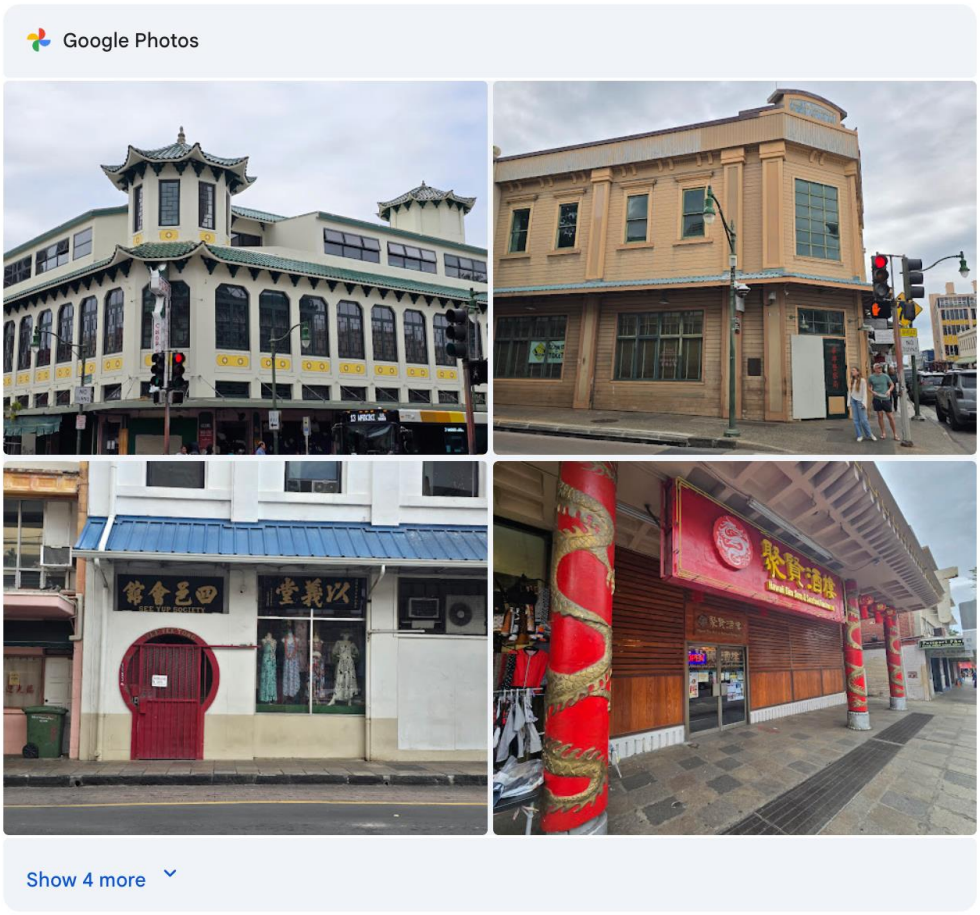

Figure A1. Visual data of architecture in Chinatown, Honolulu.

Appendix B: Example of Using PT to Retrieve and Analyze Video Data from YouTube

Prompt: "Access the YouTube videos about comparing Nvidia GPU and Google TPU. I want to know how many videos favor TPU, how many favor GPU, and how many are neutral? Please show one example of each category (i.e., the link to the video)."

Gemini's answer:

Based on a representative sample of 10 prominent videos from late 2024 and 2025, the breakdown of stances is as follows:

- Favoring Google TPU (3 videos): These videos often frame the TPU as an "Nvidia-killer," focusing on "ruthless efficiency" and the breakthrough in performance-per-dollar for massive LLM training.
- Favoring Nvidia GPU (2 videos): These prioritize the "CUDA moat" and architectural flexibility, arguing that GPUs remain the superior choice for the broader AI market (beyond just language models).
- Neutral/Comparative (5 videos): These provide balanced technical deep dives, framing the choice as a strategic "Implementation Detail" based on whether a project needs R&D flexibility or production-scale efficiency.

Representative Examples and Analysis

1. Favoring Google TPU: "Google TPU v5p: Faster and More Powerful than Nvidia H100 GPU" This video presents the TPU v5p as a direct and superior challenger to Nvidia's flagship hardware. It emphasizes that TPUs are increasingly "on par or even superior" to the H100 in raw power and

 doi:10.20944/preprints202604.1121.v1

bandwidth for training "frontier-scale" models like Gemini 3. It views Google's vertical integration (chip + software + cloud) as a game-changer that challenges Nvidia's market dominance.

- Link: http://www.youtube.com/watch?v=VSdUaox0sfg

2. Favoring Nvidia GPU: "IT'S OVER! I Can't Stay Quiet on GOOGLE vs NVIDIA Any Longer" While the title is sensational, the analysis concludes that Nvidia's "general-purpose" architecture makes it the safer and more versatile bet for 90% of the industry. It argues that while TPUs win on "token generation" for Google and Meta, GPUs are the only viable hardware for robotics, protein folding, and self-driving cars. It frames the TPU as a niche alternative that only aims to capture 10% of the market.

- Link: http://www.youtube.com/watch?v=ZZ2nWg1QhR4

3. Neutral/Balanced: "TPU vs GPU: The Ultimate Guide to the AI Hardware War" This video provides a highly technical comparison, describing GPUs as "Swiss Army Knives" for messy R&D and TPUs as "Industrial Apple Corers" for ruthless production efficiency. It posits that the "war" is not zero-sum, but rather a diversification of the supply chain. It concludes that with modern compilers like JAX and XLA, the "hardware choice" is becoming less of a technical prison and more of a boardroom decision based on unit economics.

- Link: http://www.youtube.com/watch?v=THruaU8V6gg

Appendix C: Example of Using PT to Retrieve and Analyze Video Data on Google Drive

Prompt: "In the folder "Unit 5" of "6500" in my Google drive, please summarize the video about how quantum technology would affect security of cloud computing."

Gemini's answer:

The video in your "Unit 5" folder, titled "The Quantum Dawn: When Encryption Meets Its Greatest Challenge," presented by Alex Yu, explores how the transition from classical to quantum computing creates a fundamental shift in cloud security.

The Quantum Challenge to Encryption

The core message is that current digital security—protecting everything from financial transactions to confidential cloud files—relies on mathematical problems that are currently "unbreakable" by classical supercomputers. Quantum technology changes this landscape through several key mechanisms:

- Breaking Mathematical Foundations: Current schemes like RSA and ECC rely on the difficulty of factoring large prime numbers. Shor's algorithm, designed for quantum processors, can solve these problems exponentially faster, effectively dismantling the foundation of modern encryption.
- Weakening Symmetric Encryption: Grover's algorithm can accelerate brute-force attacks, halving the effective key strength of symmetric encryption like AES.
- "Harvest-Now, Decrypt-Later": A significant emerging risk is that adversaries may collect and store encrypted cloud traffic today with the intent of decrypting it once quantum tools become sufficiently mature.

Impact on Cloud Infrastructure

Cloud computing's entire trust framework—including TLS/SSL connections, user authentication, and data-at-rest encryption—is built on these vulnerable cryptographic methods. The video warns that once quantum computing reaches full maturity, these frameworks could essentially collapse, making current data protection methods obsolete.

 doi:10.20944/preprints202604.1121.v1

The Race Toward Quantum-Resistant Security

To mitigate these risks, the video discusses two primary defensive paths:

- Post-Quantum Cryptography (PQC): Developing new encryption standards (like CRYSTALS-Kyber and CRYSTALS-Dilithium) that use "lattice-based puzzles"—complex geometric structures believed to be resistant to quantum attacks.
- Quantum Key Distribution (QKD): A defense strategy that uses the laws of physics rather than mathematics. Because observing a quantum particle changes its state, any attempt to eavesdrop on a message is immediately detectable.

The video concludes that while the "Quantum Dawn" poses a threat to digital privacy, it also offers a "Quantum Renaissance" for breakthroughs in areas like medicine and materials science, provided humanity can adapt its security frameworks fast enough.

References

1. Acharya, K., Sharifi, I., Lad, M., Sun, L., & Song, H. Integrating neurosymbolic AI in advanced air mobility: A comprehensive survey. arXiv, 2025. https://doi.org/10.48550/arxiv.2508.07163
2. Adewumi, T., Alkhaled, L., Gurung, N., van Boven, G., & Pagliai, I. Fairness and Bias in Large Multimodal Models: A Survey. arXiv.2024. https://arxiv.org/html/2406.19097v2
3. Alharbi, B., & Alshanbari, H. S. Face-voice based multimodal biometric authentication system via FaceNet and GMM. PeerJ. Computer science, 2023, 9, e1468. https://doi.org/10.7717/peerj-cs.1468
4. Anil, R. et al. Gemini: A family of highly capable multimodal models. arXiv, 2023. https://arxiv.org/abs/2312.11805
5. Anthropic. Claude 3.5 Sonnet. Anthropic News. Available online: https://www.anthropic.com/news/claude-3-5-sonnet (accessed on 21 June 2024).
6. Appaca. (2025). GPT-4o Audio vs Gemini 1.5 Pro: A Detailed Comparison. Appaca Resources. https://www.appaca.ai/resources/llm-comparison/gpt-4o-audio-vs-gemini-1.5-pro
7. Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., & Shou, M. Z. Hallucination of multimodal large language models: A survey. arXiv, 2024. https://arxiv.org/pdf/2404.18930
8. Barry, C. What is data gravity and why does it matter? Barracuda. Available online: https://blog.barracuda.com/2026/01/28/what-is-data-gravity-and-why-does-it-matter- (access on 28 January 2026).
9. Bengio, S., Marcel, C., Marcel, S., & Mariéthoz, J. Confidence measures for multimodal identity verification. IDIAP Research Report, 2002. https://publications.idiap.ch/attachments/reports/2001/rr01-38.pdf
10. Chaithanya, J. Claude 3.5 Sonnet vs. GPT-4o: Which is superior? Elephas Blog. Available online:https://elephas.app/blog/claude-sonnet-vs-gpt4o (access on 25 March 2026)
11. Chen, J., & Bian, V. Safe RAG with HydroX AI and Zilliz: PII Masking for Responsible GenAI. Zilliz Blog, 2024.https://zilliz.com/blog/safe-rag-with-hydrox-ai-and-zilliz-pii-masking-for-responsible-genai
12. Denzin, N. K., & Lincoln, Y. S. (2018). The SAGE handbook of qualitative research (5th ed.). SAGE Publications.
13. Dong, D. The limitations of vector-based retrieval in RAG. 2025. https://writer.com/blog/vector-based-retrieval-limitations-rag/
14. Erich, H. (2024, November 8). Gemini 1.5 Pro vs ChatGPT 4o: Choosing the Right Model. PromptLayer Blog. https://blog.promptlayer.com/gemini-1-5-pro-vs-chatgpt-4o-choosing-the-right-model/
15. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118. https://doi.org/10.1038/nature21056
16. European Data Protection Supervisor. Machine Unlearning. TechSonar, 2025. https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/machine-unlearning_en
17. Feng, H., Sheng, Z., Qiang, M., & Zhang, W. Generative giants, retrieval weaklings: Why do multimodal large language models fail at multimodal retrieval? arXiv, 2025. https://arxiv.org/abs/2512.19115

 doi:10.20944/preprints202604.1121.v1

18. Fu, C. et al. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. arXiv preprint, 2025. https://arxiv.org/pdf/2405.21075
19. Gao, Y.; Ji, Z.; Zhao, G. A Multimodal retrieval-augmented generation system for intelligent question answering. In Proceedings of the 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Prague, Czech Republic, 9–12 October 2022; pp. 104-110.
20. Gong, W. G. Reranker optimization via geodesic distances on k-NN manifolds. arXiv, 2026. https://arxiv.org/abs/2602.15860
21. Google Cloud. Audio understanding with Gemini. Google Cloud Vertex AI Documentation. https://cloud.google.com/vertex-ai/generative-ai/docs/multimodal/audio-understanding (Accessed in March 2025)
22. Google. Gemini 1.5: Our next-generation model. Google Blog. https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/ (Accessed in February 2026)
23. Google. (2026). Personal Intelligence: Help that's made for you. https://gemini.google/overview/personal-intelligence
24. Gray, J., & Reuter, A. Transaction processing: Concepts and techniques. Morgan Kaufmann, USA, 1993.
25. Han, J., Haihong, E, Le, G., and Du, J. Survey on NoSQL database. 2011 6th International Conference on Pervasive Computing and Applications, Port Elizabeth, 2011, pp. 363-366, doi: 10.1109/ICPCA.2011.6106531.
26. Hitzler, P., Eberhart, A., Ebrahimi, M., Sarker, M. K., & Zhou, L. Neuro-symbolic approaches in artificial intelligence. National Science Review, 9(6), 2022. https://doi.org/10.1093/nsr/nwac035
27. Hunger, M. What is GraphRAG? Neo4j. https://neo4j.com/blog/genai/what-is-graphrag/ (Accessed on March 24, 2026)
28. Husserl, E. The crisis of European sciences and transcendental phenomenology (D. Carr, Trans.). Northwestern University Press, USA, 1936/1970.
29. IBM, Inc. (2026). What is edge AI? https://www.ibm.com/think/topics/edge-ai
30. Iwanowski, M., & Gahbler, M. Multiple large AI models' consensus for object detection — A survey. MDPI Applied Sciences, 2025. https://www.mdpi.com/2076-3417/15/24/12961
31. Jain, A. K., & Flynn, P. J. Handbook of biometrics. Springer, 2019.
32. Jain, A. K., Ross, A., & Nandakumar, K. Introduction to biometrics. Springer, 2011.
33. Ji, Z., et al. Survey of hallucination in natural language generation. ACM Computing Surveys, 2023, 55(12), 1-38.
34. Jin, Y., Li, J., Gu, T. et al. Efficient multimodal large language models: a survey. Visual Intelligence. 2025, 3, Article 27. https://doi.org/10.1007/s44267-025-00099-6
35. Kim et al. Pixel-grounded retrieval for knowledgeable large multimodal models. arXiv, 2026, arXiv:2601.19060v1.
36. Kumar, A. Hands-on Agentic RAG. Medium, 2025. https://abvijaykumar.medium.com/hands-on-agentic-rag-1-2-cdf375ad7e7a
37. Lakera Team. LLM Hallucinations in 2026: How to Understand and Tackle AI's Most Persistent Quirk. Lakera.ai, 2026. https://www.lakera.ai/blog/guide-to-hallucinations-in-large-language-models
38. LanceDB. Designed for multimodal: Built for scale. LanceDB, 2025. https://lancedb.com/
39. LeCun, Y. A path towards autonomous machine intelligence. OpenReview, 2022. https://openreview.net/pdf?id=BZ5a1r-kVsf
40. Levandoski, J., Casto, G., Deng, M., Desai, R., Edara, P., Hottelier, T., Hormati, A., Johnson, A., Johnson, J., Kurzyniec, D., McVeety, S., Ramanathan, P., Saxena, G., Shanmugam, V., & Volobuev, Y. BigLake: BigQuery's evolution toward a multi-cloud lakehouse. Proceedings of the 2024 International Conference on Management of Data (SIGMOD '24), 2024 https://research.google/pubs/biglake-bigquerys-evolution-toward-a-multi-cloud-lakehouse/
41. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 2020, 33, 9459–9474.

 doi:10.20944/preprints202604.1121.v1

42. Li, P., Wang, X., Huang, K., Huang, Y., Li, S., & Iqbal, M. Multi-model running latency optimization in an edge computing paradigm. MDPI Sensors, 2022. https://www.mdpi.com/1424-8220/22/16/6097
43. Liu, B., Ning, Y., He, S., Guo, F., Jia, S., Zhu, L. A Multimodal Retrieval-Augmented Generation System for Intelligent Question Answering. In: Yang, Y., Huang, M., Pan, X., Zhang, J., Chen, J., Zhang, LJ. (eds) Cognitive Computing - ICCC 2025. ICCC 2025. Lecture Notes in Computer Science, vol 16156. Springer, Cham, 2026. https://doi.org/10.1007/978-3-032-06310-6_1
44. Liu, N. F., et al. Lost in the middle: How language models use long contexts. arXiv reprint, 2023. https://arxiv.org/abs/2307.03172.
45. Maes, L., Le Lidec, Q., Scieur, D., LeCun, Y., & Balestriero, R.. LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels. arXiv, 2026. https://arxiv.org/pdf/2603.19312v1
46. Mahmood, A., & Szabolcsi, R. A systematic review on risk management and enhancing reliability in autonomous vehicles. MDPI Machines, 2025. https://www.mdpi.com/2075-1702/13/8/646
47. Meta. (2024). V-JEPA: The next step toward advanced machine intelligence. Meta AI Blog. https://ai.meta.com/blog/v-jepa-yann-lecun-ai-model-video-joint-embedding-predictive-architecture/
48. Microsoft. Microsoft SQL Server. 2025. https://www.microsoft.com/en-us/sql-server
49. Microsoft. Multimodal search in Azure AI Search. Microsoft Learn. 2025. https://learn.microsoft.com/en-us/azure/search/multimodal-search-overview
50. Mortaheb, M.; Khojastepour, M.A.A.; Chakradhar, S.; et al. RAG-Check: Evaluating Multimodal Retrieval Augmented Generation Performance. arXiv 2025, arXiv:2501.03995.
51. O'Brien, C. Yann LeCun's AMI Labs launches with $1.03 billion to build AI that understands the real world. The French Tech Journal, 2026. https://www.frenchtechjournal.com
52. Poccia, D. Amazon Nova Multimodal Embeddings now available in Amazon Bedrock. AWS News Blog. Available Online: https://aws.amazon.com/blogs/aws/amazon-nova-multimodal-embeddings-now-available-in-amazon-bedrock/ (Accessed on October 28, 2025)
53. Raymond, B. Level Up Your GenAI Apps: What's Next for RAG. Unstructured.io Blog. Available Online: https://unstructured.io/blog/level-up-your-genai-apps-what-s-next-for-rag (Accessed on June 4, 2025)
54. Samanta, H. Grounded Multimodal Retrieval-Augmented Drafting of Radiology Impressions Using Case-Based Similarity Search. arXiv 2026, arXiv:2603.17765v1.
55. Schutz, A. The phenomenology of the social world (G. Walsh & F. Lehnert, Trans.). Northwestern University Press, USA, 1967.
56. Shao, Z., Dou, W., & Pan, Y. Dual-level Deep Evidential Fusion: Integrating multimodal information for enhanced reliable decision-making in deep learning. Information Fusion, 2023. https://www.sciencedirect.com/science/article/pii/S1566253523004293
57. Shashanka, B. R., Charan, M., & Banu, F. S. Chunking strategies for multimodal AI systems. arXiv, 2025. https://arxiv.org/pdf/2512.00185
58. Singh, A., Chen, J., Zhang, Y., & Wang, W. Y. Agentic retrieval-augmented generation: A survey on agentic RAG. arXiv preprint, 2025. arXiv:2501.09136.
59. Su, J., & Soni, G. Introducing BigQuery ObjectRef: Supercharge your multimodal data and AI processing. Google Cloud Blog. Available Online: https://cloud.google.com/blog/products/data-analytics/new-objectref-data-type-brings-unstructured-data-into-bigquery (Accessed on June 25, 2025)
60. Szczuko, P., Harasimiuk, A., & Czyżewski, A. Evaluation of decision fusion methods for multimodal biometrics in the banking application. Sensors, 2022. https://pmc.ncbi.nlm.nih.gov/articles/PMC8951111/
61. Talukdar, W., Hu, J., Mascarenhas, K., & Zhang, L. Unleashing the multimodal power of Amazon Bedrock Data Automation. AWS Machine Learning Blog. Available Online: https://aws.amazon.com/blogs/machine-learning/unleashing-the-multimodal-power-of-amazon-bedrock-data-automation-to-transform-unstructured-data-into-actionable-insights/ (Accessed on March 20, 2025)
62. Tonic.ai. Sensitive data in text embeddings is recoverable. Tonic.ai Blog. Available Online: https://www.tonic.ai/blog/sensitive-data-in-text-embeddings-is-recoverable (Accessed on November 16, 2025)
63. Wang, Y.; Zhang, Z. Visual-RAG: Benchmarking Text-to-Image Retrieval Augmented Generation for Visual Knowledge Intensive Queries. arXiv 2025, arXiv:2502.16636v2.

 doi:10.20944/preprints202604.1121.v1

64. Xia, P. et al. MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models. arXiv 2024, arXiv:2410.13085v2.
65. Xie, J., Zhou, X., & Cheng, L. Edge computing for real-time decision making. International Journal of Advanced Computer Science and Applications, 2024, 15(7), 598-607.
66. Yu, C. H. Philosophical foundations of quantitative research methodology. University Press of America, USA, 2006.
67. Yu, C. H. Are positive trait attributions for the deceased caused by fear of supernatural punishments?: A triangulated study by content analysis and text mining. Journal of Psychology and Christianity, 2015, 34, 3-18.
68. Yu, C. H. Data mining and exploration: From traditional statistics to modern data science. CRC Press, USA, 2022.
69. Yu, C. H. Philosophical foundations of classical statistic, Bayesian statistic, and data science/machine learning. In Ulf Liebe (Ed.). Philosophical principles of quantitative research. Edwards Elgar Publishing, USA, in press
70. Yu, C. H., DiGangi, S., & Jannasch-Pennell, A. Using text mining for improving student experience management. In P. Tripathi & S. Mukerji (Eds.), Cases on innovations in educational marketing: Transnational and technological strategies, 2011, 196-213. IGI Global.
71. Yu, C. H., Jannasch-Pennell, A., & DiGangi, S. Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. Qualitative Report, 2011, 16, 730-744. http://www.nova.edu/ssss/QR/QR16-3/yu.pdf
72. Yu, C. H., Jannasch-Pennell, A., & DiGangi, S. Enhancement of student experience management in higher education by sentiment analysis and text mining. International Journal of Technology and Educational Marketing, 2018, 8, 16-33. DOI: 10.4018/IJTEM.2018010102.
73. Zheng, X, et al. Retrieval augmented generation and understanding in vision: A survey and new outlook. arXiv 2025, arXiv:2503.18016.