International Scientific Journal of Engineering and Management
Volume: 04 Issue: 10 | Oct– 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM05099

Real-Time Multimodal AI for Unified Omnichannel Retail Experiences

Author's Name: **Udit Agarwal, Aditya Gupta**

Author's Email: udit15@gmail.com, adityagupta8121@gmail.com

Abstract

The convergence of advanced artificial intelligence (AI) systems and high-throughput data architectures is fundamentally reshaping the retail sector, catalyzing unprecedented advancements in operational efficiency and customer engagement. This paper presents the complex infrastructure required to achieve unified, real-time omnichannel experiences. Specifically, the analysis details how multimodal deep learning, leveraging the simultaneous processing of visual, textual, and categorical data through transformer-based architectures, enhances product categorization and contextual intelligence far beyond single-input systems. We discuss performance Service Level Agreements (SLAs) relevant to real-time customer interaction, where read latency for AI bot queries may target p95 thresholds of ≤30ms\leq 30\text{ms} ≤30ms. To meet these demands, the supporting architecture must integrate a high-speed Hybrid Transactional/Analytical Processing (HTAP) data store, Kafka Streams for sub-second event freshness, and low-latency caching. Furthermore, the strategic deployment of Edge-AI is essential for autonomous physical retail monitoring, integrating computer vision and sensor fusion to mitigate real-world variability. These integrated systems are positioned to address challenges in ensuring content and process consistency across diverse customer touchpoints, thereby reducing retailer uncertainty and positively impacting customer loyalty. This comprehensive framework underscores the pivotal role of real-time multimodal intelligence in propelling retailers toward greater agility and customer-centricity.

Keywords

Multimodal AI, Omnichannel Retail, Edge-AI, Low-Latency Architecture, Data Fusion, Real-Time Systems, Customer Journey Mapping, HTAP.

1. Introduction

1.1. The Strategic Shift to Unified Omnichannel Retail

The evolution of retail dynamics has necessitated a strategic shift from traditional multichannel distribution networks to fully integrated omnichannel models. This transition requires firms to adopt an integrative strategy, actively coordinating their operations across all channels and managing every phase of the customer journey. The core objective of true omnichannel operation is the achievement of a high degree of channel collaboration, which is characterized by three fundamental pillars: unified pricing, unified product information, and a unified brand image across all customer touchpoints.

The complexity of orchestrating multiple physical and digital channels introduces significant challenges. Retailers must manage and integrate information across these various platforms. A failure to integrate channels effectively results in substantial customer confusion regarding the availability and differences between services, inflicting difficulty on the purchasing journey. Furthermore, inconsistencies in content, such as divergent product specifications, warranty services, or pricing, often lead to customer frustration,which can prompt customers to explore alternative channels or retailers. Research indicates that reducing retailer uncertainty through process consistency and the breadth of channel service choice is among the most influential factors in retaining customers and deterring them from seeking alternatives. Successfully mitigating these integration risks through a coherent strategy is therefore paramount for fostering customer loyalty and commitment.

ISSN: 2583-6129
DOI: 10.55041/ISJEM05099

1.2. Convergence of Real-Time Systems and Intelligent Automation

The necessity of managing complex, integrated, high-frequency environments compels retailers to adopt intelligent automation driven by Artificial Intelligence. AI's intrinsic capability to analyze vast amounts of real-time data enables retailers to respond more effectively to customer needs and influence purchase intentions through personalized communications. AI-driven algorithms provide real-time insights and hyper-personalization across the value chain, from optimizing logistics and inventory replenishment to enhancing customer satisfaction.

A critical component of this advanced automation is the implementation of multimodal AI systems. Multimodal systems integrate data from diverse sources—such as text, speech, visual inputs, and sensor data—to achieve a contextual understanding superior to that offered by traditional, single-mode approaches. This convergence allows for the robust detection and interpretation of emotional cues embedded in customer feedback, social media posts, and chat interactions, transforming unstructured data into actionable insights instantaneously. For these systems to operate effectively in dynamic retail environments, performance must be real-time, necessitating low-latency architectures capable of collecting high-frequency, fine-grained data conveniently and accurately. The operational requirement is a system agile enough to anticipate customer preferences and respond immediately to rapidly evolving market trends.

1.3. Paper Structure and Research Contribution

This white paper outlines technical foundations relevant to deploying unified, real-time multimodal AI systems in retail environments. Section 2 analyzes the advanced deep learning architectures required for multimodal data fusion and generating comprehensive product intelligence. Section 3 outlines the stringent architectural constraints, including Service Level Agreements (SLAs), essential for achieving low-latency, real-time data processing and AI serving. Section 4 explores the strategic application of Edge-AI deployment and the system's role in resolving critical omnichannel consistency challenges, thereby improving the overall customer journey. The concluding section summarizes the necessity of this integrated technological approach for achieving operational excellence and delivering consistently unified customer experiences.

2. Principles of Multimodal Deep Learning for Retail Intelligence

Multimodal AI represents a fundamental technological shift, offering a robust solution for the increasingly complex challenges inherent in modern commerce. This complexity stems from the need to leverage complementary information from multiple data streams to construct a holistic and accurate understanding of both products and customer interactions.

2.1. Feature Extraction and Comprehensive Product Understanding

The application of multimodal AI fundamentally transforms e-commerce infrastructure by automating product organization and categorization. Rather than relying on labor-intensive manual tagging systems, modern platforms use intelligent automation that adapts to evolving catalogs and consumer requirements.

This comprehensive product understanding is built upon the simultaneous processing of multiple modalities, including textual descriptions, images, and videos. Advanced feature extraction techniques leverage sophisticated Natural Language Processing (NLP) for text, Computer Vision (CV) for images, and temporal analysis for video, capturing meaningful product attributes that manual categorization processes frequently overlook. Architectural implementation often relies on transformer-based models integrated with distributed processing architectures and lambda models, demonstrating superior scalability necessary for modern commerce requirements.

Specific research into price estimation, for example, illustrates the technical complexity involved. Deep learning frameworks integrate EfficientNetB1 for visual features, pre-trained GloVe embeddings combined with a Bidirectional Long Short-Term Memory (Bi-LSTM) network for textual input, and trainable embeddings for categorical product data.This multimodal intelligence has led to substantial improvements in search relevance, user engagement, and conversion rates across diverse retail environments. The simultaneous consideration of these modalities, rather than sequential or single-input processing, is essential for high-fidelity product classification.

ISSN: 2583-6129
DOI: 10.55041/ISJEM05099

The requirement for integrating diverse deep learning models, such as Convolutional Neural Networks (CNNs) (e.g., EfficientNet) with sequence models (Bi-LSTMs and Transformers), necessitates a powerful computational infrastructure. These complex models must execute inference in real-time. Therefore, the necessity of integrating these specialized architectural components directly ties the required AI capability to the stringent low-latency performance indicators detailed in Section 3. Achieving high performance requires careful optimization of these computationally intensive models through techniques like pruning and quantization, particularly for resource-constrained deployments.

2.2. Advanced Data Fusion Methodologies

The effectiveness of multimodal systems is highly contingent upon the strategic methodology used to combine and interpret disparate data streams. Multimodal data possesses several characteristics that guide fusion design :

1. **Complementarity:** No single data type can provide a full explanation of a specific learning phenomenon or process.
2. **Mutual Verification:** Different types of data must be capable of verifying the same results, which is essential for ensuring reliability and trust in the system's outputs.
3. **Fusion (Integration):** Physical data, such as body movements or gestures captured in a physical store, must be stored and synchronized with digital data, such as corresponding log entries in digital platforms.
4. **Transformation:** Physical data collected from sensors or cameras must be accurately transformed into a digital format suitable for computational analysis.

Fusion methodologies applied in retail AI include *late fusion*, where modality features are integrated closer to the final classification or output layer, and the use of *dense fusion layers*. Crucially, *attention-based fusion* mechanisms are employed to aggregate the features from multiple modalities, revealing complex relationships between products and consumer preference patterns that are invisible to single-input systems.

The demand for mutual verification and fusion serves as the technical mechanism for solving the critical business challenge of omnichannel consistency. If a customer interacts with product information both in a physical environment (e.g., using visual identification via a camera or sensor reading) and online (via a digital log), the multimodal system must align and verify these two inputs simultaneously. Errors during the alignment and synchronization of these data streams can propagate throughout inventory management or pricing systems, leading directly to the inconsistent content (e.g., mismatched availability or pricing) that causes customer frustration. Robust fusion, therefore, is not merely a feature but an architectural necessity for enforcing the unified strategic outcome of channel collaboration.

2.3. Applications in Fine-Grained Classification and Operational Efficiency

The deployment of multimodal deep learning architectures provides robust solutions across operational efficiency and customer experience domains.

Operationally, these systems excel at fine-grained product classification and detection in highly variable environments, including warehouses, retail shelves, and supermarkets. By integrating specialized models, such as Convolutional Neural Networks (CNNs) like ResNet and EfficientNet, with real-time detection frameworks like YOLO, systems can achieve real-time recognition on retail shelves. This capability is important because multimodal recognition systems often demonstrate higher reliability and contextual understanding, specifically in challenging conditions marked by product occlusion, varying lighting, or non-standard labeling. This enhanced reliability translates directly into more precise inventory management, accurate planogram compliance verification, and sophisticated operational analytics.

For the customer experience, multimodal intelligence is redefining interaction by enhancing the capabilities of virtual assistants and chatbots. These sophisticated systems process voice commands, speech patterns, and text data simultaneously, enabling them to be more intuitive, responsive, and natural in their interactions. Furthermore, AI-powered sentiment analysis is utilized across multinational e-commerce platforms to manage real-time customer experience. This involves leveraging machine learning and Natural Language Processing (NLP) models to analyze millions of data points

ISSN: 2583-6129
DOI: 10.55041/ISJEM05099

across various communication channels, detecting linguistic nuances and emotional tones embedded in customer feedback, thereby allowing organizations to respond promptly to signals that influence brand loyalty.

3. Architectural Design for Real-Time, Low-Latency AI Services

The realization of unified, real-time experiences demands a specialized, low-latency, high-throughput data architecture. Performance in such environments is governed by strict Service Level Agreements (SLAs) for data retrieval and model inference.

3.1. Defining Performance Service Level Agreements (SLAs) for AI Bots

High-frequency environments require that interactions with AI services, such as customer-facing bots, meet stringent latency targets. These standards are necessary to ensure a seamless and responsive customer experience. Since sophisticated multimodal models (e.g., those using EfficientNetB1) are computationally demanding, access to input features must be instantaneous. Consequently, the architecture typically benefits from integrating caching and streaming data layers to provide immediate feature lookups, avoiding costly, slower queries to the primary analytical database layer. This establishes a direct causal relationship between the required latency SLA and the mandated deployment of a multi-tiered data processing architecture.

3.2. Event Stream Processing (ESP) and Data Ingestion Pipeline

To support the low-latency demands of AI models, data must be continuously ingested and transformed in real-time. This is achieved through a robust Event Stream Processing (ESP) pipeline.

The primary function of Kafka is the ingestion of all event streams—including transactions, user actions, market data, and enrichment events—which requires the system to be engineered for a high ingestion throughput. This constant flow of data is essential because insights derived from historical data must be continuously updated and refined against current event data to ensure predictions remain practical and applicable in the present moment.

Kafka Streams handles the vital low-latency transformations, aggregations, and enrichment processes (e.g., joining events with lookup caches). This component produces materialized views (KTables) that write derived events back to high-speed storage components. The specific role of Event Stream Processing involves filtering and abstracting incoming event streams to generate "complex events." This abstraction process, referred to as "feature generation" by data scientists, is crucial because it presents the information in a highly effective form for real-time analytics and model input. The end-to-end freshness for this streaming view is stringently targeted for critical events. This sub-second freshness guarantee is fundamental to ensuring that AI systems act upon the most current state of the retail environment and customer behavior.

3.3. Hybrid Data Storage and Caching Strategies

The real-time requirement necessitates a specialized, hybrid data architecture capable of serving both historical transactional integrity and real-time analytical insights with minimal delay. This architecture relies on a structured layering of components optimized for different latency tiers.

- **Low-Latency Cache (Redis):** To meet the millisecond latency SLAs, Redis is implemented as a low-latency cache, utilizing a Least Recently Used (LRU) policy with Time-to-Live (TTL) expiration. This component stores hot user records, session data, and precomputed insight snippets, using structures like hashes and sorted sets to support fast top-N queries and list retrieval crucial for personalized recommendations.
- **HTAP Database (TiDB):** A Hybrid Transactional/Analytical Processing (HTAP) database, such as TiDB, serves as the source of authoritative user state and mid-cardinality pre-aggregated metrics. It handles concurrent transactional writes and lightweight analytical lookups (e.g., multi-join ad-hoc queries), providing the necessary agility for both updating status and querying historical context.
- **Batch Pre-Aggregation (Spark):** Scheduled batch jobs, powered by systems like Spark, compute complex pre-aggregations, risk scores, reference tables, and specialized Machine Learning (ML) feature

ISSN: 2583-6129
DOI: 10.55041/ISJEM05099

engineering. Spark writes these aggregated results back to the low-latency (Redis) and HTAP (TiDB) layers for fast lookups.

The **API Layer** acts as the crucial interface, exposing data to the AI services/bots. It employs a read-through cache strategy: it first checks Redis, then queries TiDB upon a cache miss, and finally populates Redis with the retrieved result to ensure future requests benefit from low latency. This integrated architecture, sometimes following the principles of a Lambda Architecture (combining batch processing, streaming, and caching), ensures that the AI bot layer has immediate access to cached insights for model logic execution, context application, and output rendering.

The required integration of batch (Spark), streaming (Kafka Streams), HTAP (TiDB), and caching (Redis) reflects a necessary convergence of data processing paradigms. Traditional retail systems, heavily reliant on periodic batch processing, cannot support the continuous, real-time engagement required by modern customers. By integrating Kafka Streams for sub-second freshness and Redis for millisecond access , the architecture creates a unified data view that supports both deep historical analysis and instant action. This hybrid approach ensures that the historical context and the current interaction status are immediately available to the AI agents, defining the "unified" aspect of the resulting experience.

4. Strategic Deployment and Integration across the Customer Journey

The technical foundation of multimodal AI, coupled with the low-latency data architecture, provides the critical linkage necessary to unify the physical and digital retail spaces into a single, cohesive experience.

4.1. Edge-AI for Real-Time Physical Retail Monitoring

To effectively monitor and manage the physical retail environment in real-time, deployment of multimodal systems must occur directly at the retail location via Edge-AI. This strategy requires deploying AI models directly on local devices or edge servers, minimizing latency by avoiding cloud round-trips and thereby reducing dependence on external infrastructure. Dedicated Edge-AI frameworks, such as NVIDIA Jetson, Google Coral, and Intel OpenVINO, are used to provide the necessary hardware-accelerated on-device inference capabilities.

Deployment at the edge, however, is constrained by the resource limitations inherent to edge devices, including finite memory capacity, restricted processing throughput, and energy efficiency concerns. These constraints necessitate rigorous model optimization, including pruning and quantization, along with careful resource allocation to ensure real-time performance without compromising accuracy.

The multimodal integration utilized at the edge is crucial for overcoming the high variability of retail settings. The system integrates computer vision for product detection, Optical Character Recognition (OCR) for textual extraction, and sensor fusion techniques. This combination is mandatory for handling real-world challenges such as inconsistent lighting, occlusion caused by crowded shelves, and non-uniform product placement. This robust sensor fusion capability is essential because single-modality vision systems often fail under such variability. Therefore, Edge-AI deployment, supporting multimodal fusion that increases reliability, is a technical prerequisite for establishing robust, autonomous retail operations that can perform tasks like inventory management and planogram verification in real time.

4.2. Overcoming Omnichannel Consistency Challenges

Unified AI is strategically positioned to enforce process and content consistency across all customer channels, which is critical for reducing operational friction and mitigating retailer uncertainty. Customer loyalty is shown to increase proportionally with the degree of channel collaboration, characterized by unified product information, pricing, and image. Process consistency—the uniformity of comparable attributes like visual representation, ease of ordering, and delivery speed across channels—is identified as one of the most influential factors in reducing customer uncertainty.

Inconsistency is a major failure point in omnichannel strategy. Research highlights that if retailers fail to provide consistent content, spanning details like product specifications and warranty services, customers experience frustration and may switch channels. Furthermore, studies indicate that a significant percentage of participants experience an inconsistent

ISSN: 2583-6129
DOI: 10.55041/ISJEM05099

shopping journey, often due to companies struggling to meet specific customer needs with their existing, disparate channel arrangements.

The fundamental challenge is the difficulty in managing and integrating information across various channels. Omnichannel inconsistency is a direct symptom of high data synchronization latency and low data quality between systems.

Table 2: Multimodal Data Fusion Taxonomy and Application in Unified Retail

Modality Type	**Retail Data Examples**	**Feature Extraction/Fusion Method**	**Outcome in Unified Retail**
Visual	Shelf images, product videos, packaging details.	CNNs (EfficientNet, ResNet), YOLO detection frameworks.	Fine-grained classification, inventory tracking, planogram verification.
Textual/Semantic	Product descriptions, online reviews, search queries, OCR data.	Pre-trained GloVe embeddings, Bi-LSTM, Transformer architectures.	Enhanced search relevance, semantic understanding, sentiment analysis.
Categorical/Contextual	Product attributes, pricing, warranty information, session data, environmental data.	Trainable embeddings, Attention-based fusion, Data alignment.	Accurate price estimation, process consistency verification, contextual intelligence.
Fusion Strategy	Data integration across modalities.	Late Fusion , Dense Fusion , Sensor Fusion (for Edge-AI).	Higher reliability against occlusion , mutual verification of data points, reduced inconsistencies.

4.3. AI-Driven Customer Journey Mapping and Experience Enhancement

Artificial Intelligence leverages these real-time multimodal data streams to continuously refine the Customer Journey Mapping (CJM), transitioning from traditional linear models to dynamic, omnichannel-based approaches. AI-driven CJM employs machine learning (ML) and Natural Language Processing (NLP) to analyze extensive customer data, yielding actionable insights that enable businesses to customize strategies for specific target markets and segments.

ISSN: 2583-6129
DOI: 10.55041/ISJEM05099

The incorporation of AI agents fundamentally changes the customer journey by providing tailored, interactive services and augmenting traditional retail services. By collecting real-time behavioral, transactional, and historical data, AI visualizes customer interactions, helping to summarize all touchpoints, identify pain points, and delineate areas for improvement.

A key capability is the utilization of AI-powered sentiment analysis for real-time emotional intelligence. This capability interprets emotional cues and attitudes expressed through text, speech, and visual data across communication channels. The analysis allows organizations to manage customer experience by promptly responding to emotional signals that may influence brand perception. If a customer's real-time sentiment analysis, derived from multimodal input, indicates frustration at a specific touchpoint visualized on the CJM, the low-latency architecture (p95 $\leq 30\text{ms}$ latency) allows an AI assistant to intervene immediately with tailored communication. This rapid, contextualized intervention, powered by instantly retrieved insights from the Redis cache , closes the loop from real-time multimodal detection to sub-millisecond action, defining a truly unified and enhanced customer experience.

5. Conclusion

5.1. Synthesis of Unified Architectural Requirements

The successful implementation of a unified omnichannel retail experience rests upon the synergistic integration of advanced multimodal AI capabilities and a rigorously engineered, low-latency data infrastructure. Multimodal intelligence provides the necessary contextual understanding for precise operations, enabling the transformation of product categorization and monitoring from labor-intensive systems to intelligent, adaptive automation.

Architecturally, achieving this unification requires absolute dedication to stringent performance Service Level Agreements (SLAs).

5.2. Future Directions and Path to AI Maturity

This integrated multimodal framework directly addresses the strategic challenge of information inconsistency that often causes high retailer uncertainty and customer attrition in traditional omnichannel systems. By enforcing data coherence and cross-channel process consistency via real-time data fusion and synchronization, retailers can enhance customer loyalty through unified service delivery.

The implementation of this real-time multimodal AI architecture represents a critical advancement toward organizational AI maturity. As these technologies continue to evolve, future research and corporate strategy must prioritize the establishment of robust methodologies to assess the complex impact of AI, address the ethical considerations inherent in large-scale AI deployment, and resolve the complexities of integrating these systems with legacy business models. Successful navigation of this transformation requires organizations to maintain a systematic focus on measurable business outcomes and user value creation, ensuring that the AI capabilities are strategically aligned with the enterprise vision. By embracing this powerful convergence of multimodal intelligence and real-time infrastructure, retailers may be better positioned to adapt to rapidly evolving market dynamics and potentially influence future trends defined by heightened efficiency, agility, and absolute customer-centricity.

References

J. K. Author, “Title of paper,” presented at the abbrev. Name of Conf., City of Conf., Abbrev. State, year, pp. xxx–xxx.

V. K. A.. Product categorization using multi-modal artificial intelligence represents a significant advancement in e-commerce infrastructure, transforming how digital commerce platforms organize, classify, and present products to consumers. *International Journal of Computer Engineering, Science and Emerging Research*, 2024..

X. Z.. Edge-AI Deployment of Multimodal Product Recognition Systems for Smart Retail Environments. *Research Gate*, 2024..

ISSN: 2583-6129
DOI: 10.55041/ISJEM05099

Y. M.. Multimodal data fusion in learning analytics: A systematic review. *Research Output, Charles Sturt University*, 2023..

P. W.. The multimodal data are complementary, mutual verification, fusional, and transformed. *Sensors (Basel)*, 2020..

J. P.. A multimodal deep learning framework for integrating visual, textual, and categorical features in retail price estimation. *Research Gate*, 2024..

L. T.. Multimodal Deep Learning Architectures for Fine-Grained Product Classification in Retail Spaces. *Research Gate*, 2024..

Y. J.. Customer loyalty will increase as the degree of channel collaboration (unified price, unified product, unified image) of new retail enterprises increases. *International Journal of Environmental Research and Public Health*, 2022..

H. C.. The smooth flow of order fulfillment is one of the daunting challenges in omnichannel retailing. *Sustainability (Basel)*, 2021..

S. A.. This current study delves into the quality of consumer perception regarding brand channel integration that adopts the omnichannel approach. *Cogent Business & Management*, 2024..

N. S.. Disentangling the Impact of Omnichannel Integration Services on Perceived Risk in Online Shopping: A Service Perspective. *MIS Quarterly*, 2020..

R. P.. The Impact of AI along the Customer Journey Mapping: How AI Agents are Changing Customer Journey. *Proceedings of the International Conference on Business Excellence*, 2025..

T. G.. Recent studies highlight the shift from traditional linear models to dynamic, omnichannel-based journey mapping approaches. *International Journal for Research in Applied Science & Engineering Technology*, 2025..

K. M.. AI's ability to analyze vast amounts of real-time data has enabled retailers to respond more effectively to customer needs. *Journal of Theoretical and Applied Electronic Commerce Research*, 2023..

A. Z.. Artificial Intelligence Transforming the Future of Retail. *Research Gate*, 2024..

S. A.. Low-latency Online Data Store for AI-Driven Financial Insights. *Research Gate*, 2025..

J. F.. Smaller models are the only option for edge AI / low-latency execution. *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021..

I. B.. Event processing helps continuously update and refine our understanding of ongoing business scenarios. *IBM Think*, 2024..

C. E.. ESP plays a role in both kinds of activities. Incoming event streams are filtered and abstracted to generate complex events. *Complex Events*, 2025..

H. R.. The business value of generative AI and big data integration: A framework for strategic transformation. *Archives of Clinical Research*, 2024..

S. M.. A strategic framework for AI product development and evaluation in enterprise software. *Research Gate*, 2024..

C. L.. Artificial intelligence maturity model: a systematic literature review. *Research Gate*, 2021..

T. S.. What can AI maturity models achieve in business practice? *Research Gate*, 2023..

ISSN: 2583-6129
DOI: 10.55041/ISJEM05099

C. M.. AI-Powered Sentiment Analysis for Real-Time Customer Experience Management in Multinational E-Commerce Platforms. *Research Gate*, 2024..

N. A.. Multimodal AI systems integrate data from these different sensors, allowing the vehicle to make real-time decisions and respond to its surroundings. *IMD Blog*, 2024..

A. B.. The deployment of a real-time Edge-AI multimodal product recognition system in smart retail environments faces several challenges and involves specific integration methods as outlined in the article. *Research Gate*, 2024..

J. T.. The article discusses the importance of process consistency and content consistency as components of omnichannel integration quality. *Cogent Business & Management*, 2024..