ESP-JETA
ESP Journal of Engineering & Technology Advancements
ISSN: 2583-2646 / Volume 6 Issue 1 January 2026
Paper Id: JETA-V6I1P102 / Page No: 12-17

Original Article

Revolutionizing Online Shopping: The Power of Multimodal Search in E-Commerce

Nitin Patki
Engineering Project Manager at Apple Inc. USA.

Received Date: 20 November 2025 Revised Date: 17 December 2025 Accepted Date: 07 January 2026

Abstract: *For two decades, the atomic unit of e-commerce has been the keyword. However, as AI evolves from a tool that helps humans find products to an agent that buys them on our behalf, the keyword is no longer sufficient. This article examines the watershed moment of multimodal search—the transition where digital catalogs stop being static lists of text and become dynamic ecosystems that can "see" and "listen." We break down the physics of meaning behind vector embeddings, the battle between "build vs. buy" search strategies, and the rise of 3D spatial indexing. Ultimately, we demonstrate why the retailers that thrive in the next decade will be those that re-architect their systems not just to be visible to humans, but to be intelligible to the algorithm.*

Keywords: *Multimodal AI Search, E-commerce, Visual Search Technology, Voice Search for Retail, Vision-Language Models (VLMs), CLIP Model E-commerce, Semantic Search Algorithms, Neural Search Engine.*

I. INTRODUCTION

The digital commerce landscape is currently navigating a profound structural transition, moving away from the deterministic rigidity of lexical search toward the probabilistic, context-aware capabilities of multimodal intelligence. For over two decades, the fundamental atomic unit of e-commerce intent has been the keyword—a text string matched against a database index. While this mechanism served the initial era of digital retail, it fundamentally fails to capture the nuance, ambiguity, and visual nature of human desire. The emergence of high-dimensional vector embeddings, Vision-Language Models (VLMs) such as CLIP (Contrastive Language-Image Pretraining), and the nascent paradigm of Agentic AI are collectively rewriting the architectural rules of information retrieval.

This article offers an exhaustive technical and strategic analysis of this transformation. It explores the mathematical foundations of vector spaces where text, images, and audio converge into a unified semantic plane, enabling cross-modal retrieval that mimics human cognitive association. It details the infrastructure required to support these systems—from specialized vector databases like Pinecone and Weaviate to the complex data pipelines necessary for real-time indexing. Furthermore, it examines the "machine gaze," a new optimization frontier where product assets must be refined not just for human appeal but for algorithmic readability.

II. MULTIMODAL SEARCH & IT'S LIMITATIONS

A. What is Multimodal Search

Multimodal search is a sophisticated technology that applies artificial intelligence (AI) and machine learning (ML) to understand and interpret multiple input modes—such as voice, images, and standard text—to generate more accurate search results.

Users can take a picture of something and use that image to search for similar items or even speak in a different language and receive relevant results. The more users input queries, the more personalized responses can be.

For example, when searching for "milk" over time, AI will recognize that the user frequently adds the same type of milk to their cart and recommend those results first.

B. The Problem: Limitations of Lexical Search

- Outdated Architecture: Traditional e-commerce search relies on inverted indexes (like Apache Lucene), which match specific keywords from a user's query to tokens in a product catalog.
- Vocabulary Mismatch: This approach fails when a user's terminology differs from the merchant's (e.g., searching "crimson trainers" vs. "red running shoes").
- Business Impact: This rigidity leads to a high volume of null results (up to 5% of queries), causing a direct loss of high-intent traffic. It also fails to process conceptual or "vibe-based" queries (e.g., "garden party outfit").

C. The Solution: Multimodal Search

- Bridging the Gap: Multimodal search addresses the "semantic gap"—the difficulty of describing visual concepts in

text—by allowing image, voice, and hybrid inputs.

- Understanding Intent: Powered by vector embeddings, this architecture shifts the focus from "what the user said" to "what the user meant."
- Geometric Precision: It transforms vague human concepts into precise mathematical relationships, allowing the system to infer context and visual attributes without relying on exact keyword matches.

D. Benefits of Multimodal Search for E-Commerce

With multimodal search, organizations can expect increased sales, improved efficiency, and lower costs. Many of these services are automated, leading to better cost structures and operational efficiencies.

Implementing additional search types provides more data points, providing deeper insights into customer shopping patterns, personalization, and contextualization. AI enables employees to switch their focus from search data analysis to other tasks by automating repeatable tasks.

Voice and conversational search use has transformed customer care. Customer service associates can use the multimodal search features to quickly locate and accurately describe products to assist customers in person, on the phone, or online.

Multimodal search allows e-commerce companies to stay on top of the competition. Search startups are developing multimodal search options using large language models so consumers can shop in different modals (e.g., product images). One multimodal search provider saw positive results immediately, with a conversion rate five times higher than the standard e-commerce rate, demonstrating that return users activate even more searches than on their first visit.

III. THE PHYSICS OF MEANING: VECTOR EMBEDDINGS AND HIGH DIMENSIONAL SPACES

A. Mathematical Foundations of Vector Retrieval

At the core of multimodal search lies the vector embedding. An embedding is a representation of a discrete piece of information (a word, a sentence, an image) as a vector of real numbers in a continuous high-dimensional vector space.

Unlike traditional database rows where attributes are discrete (Color: Red), embeddings capture semantic relationships through geometric proximity. The dimensionality of these vectors can range from hundreds to thousands, for instance, Google's multimodal models often generate 1,408-dimension vectors. In this 1,408-dimensional space, every dimension represents a latent feature of the data—abstract concepts like "formality," "warmth," or "texture"—learned by the model during training.

The retrieval mechanism relies on calculating the distance between the query vector and the product vectors. The most common metric is Cosine Similarity, which measures the cosine of the angle between two non-zero vectors. The formula is:

$$Cosine\ Similarity(A, B) = \frac{A.B}{\|A\|\,\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\,\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Where A is the query vector and B is the product vector. A result of 1 indicates the vectors are identical in orientation (perfect semantic match), while 0 indicates orthogonality (no relation), and -1 indicates opposition. Because the magnitude of the vectors (length) is normalized, the focus remains purely on the orientation, which represents the semantic "direction" of the content.

It is crucial to note that the dot product of embeddings is not a calibrated probability, it is a raw similarity score. Different models and use cases will have different score distributions, meaning a "0.8" similarity in one domain might imply a stronger match than "0.8" in another. Therefore, thresholding strategies must be dynamic rather than fixed.

B. CLIP and the Unification of Modalities

The breakthrough that enabled true multimodal search was the development of Vision-Language Models (VLMs), most notably OpenAI's CLIP (Contrastive Language-Image Pretraining).

Prior to CLIP, computer vision models were typically trained on classification tasks—predicting a label from a fixed set of categories (e.g., ImageNet's 1,000 classes). This limited the model to recognizing only what it had been explicitly taught. CLIP, however, was trained on 400 million pairs of images and text scraped from the internet. The training objective was contrastive: the model learned to pull the vector representation of an image and its corresponding text caption closer together in the embedding space, while pushing mismatched pairs apart.

This process forces the model to learn a shared vector space for both modalities. The embedding for an image of a "golden retriever catching a frisbee" converges with the text embedding for the sentence "a dog playing fetch."

The strategic advantage of this architecture is Zero-Shot Learning. Because the model has learned to associate visual features with language generally, it can recognize and retrieve concepts it was never explicitly trained on. An e-commerce system using CLIP can retrieve "minimalist desk lamps" even if the product catalog has no tags for "minimalist," simply because the model understands the visual characteristics associated with that word. This eliminates the need for exhaustive manual tagging, which has historically been a bottleneck in Digital Asset Management (DAM).

IV. BUILDING THE MULTIMODAL ENGINE

Designing a multimodal search engine requires a specialized technology stack that differs significantly from traditional relational database architectures. The system must be capable of high-throughput vector calculations, low-latency retrieval, and seamless integration with existing product catalogs.

A. The Vector Database Landscape

The foundation of the multimodal stack is the Vector Database (or Vector Search Engine). These systems are optimized to store high-dimensional vectors and perform Approximate Nearest Neighbor (ANN) searches efficiently. The market has bifurcated into fully managed cloud-native solutions and flexible, open-source or hybrid options.

Pinecone has emerged as a leader for enterprise-grade, high-scale applications. It offers a fully managed, serverless architecture that abstracts away the infrastructure management. Pinecone is particularly favored for its consistent performance, maintaining sub-50ms latency (p99) even at billion-vector scales. Its architecture separates storage from compute, allowing retailers to handle massive spikes in query volume (e.g., during Cyber Monday) without over-provisioning storage. However, it is a proprietary, closed-source system, which may introduce vendor lock-in concerns.

Weaviate presents a contrasting philosophy. It is an open-source vector search engine that supports hybrid deployment models—cloud, on-premise, or embedded. This makes it the preferred choice for organizations with strict data sovereignty requirements (e.g., European retailers adhering to GDPR) or those who wish to keep their proprietary embedding models within their own firewalls. Weaviate utilizes a modular architecture that supports "pluggable" vectorizers and rerankers. Crucially, Weaviate has native support for Hybrid Search—the ability to combine vector search scores with traditional keyword (BM25) scores in a single query execution. This is vital for e-commerce, where users expect exact matches for specific SKUs or model numbers that fuzzy vector search might miss.

Chroma targets the developer experience, focusing on simplicity and ease of integration with Python-based ML workflows. It is often used for prototyping or smaller-scale applications where the overhead of a complex distributed system is unnecessary. However, for massive-scale production e-commerce, its performance metrics (p99 latency of ~89ms) currently lag the dedicated enterprise solutions.

B. The Indexing and Retrieval Pipeline

The operational flow of a multimodal search engine involves several distinct stages, moving from data ingestion to the final search result.

a) Ingestion and Vectorization:

As product data (images, descriptions) is ingested, it is passed through the chosen embedding models. For text, this might be a BERT-based model or OpenAI's text-embedding, for images, a CLIP or ResNet model. APIs like Google's Multimodal Embeddings API can return 1408-dimension vectors for both text and images, placing them in the same semantic space. It is critical that the image and text vectors share the same dimensionality and semantic alignment, otherwise, cross-modal retrieval is impossible.

b) Indexing:

The generated vectors are loaded into the vector database. To ensure fast retrieval, the database builds an index. The most common algorithm is HNSW (Hierarchical Navigable Small World) graphs. HNSW creates a multi-layered graph structure that allows the search algorithm to quickly traverse from broad regions of the vector space to the specific neighborhood of the query vector.

c) Hybrid Retrieval (The "Best of Both Worlds"):

Pure vector search can be imprecise with specific constraints (e.g., finding a "Canon EOS R5" specifically, not just a "high-end Canon camera"). Therefore, modern architectures employ a hybrid approach. The query is processed in parallel:

- Dense Retrieval: The query is vectorized and searched against the vector index to find semantically similar items.
- Sparse Retrieval: The query is tokenized and searched against a traditional inverted index (BM25) to find exact keyword matches.

d) Fusion and Reranking:

The results from the dense and sparse streams are merged, often using Reciprocal Rank Fusion (RRF), which

normalizes the scores from both systems and combines them. The top candidates (e.g., the top 50) are then passed to a Reranker. This is a more computationally intensive model (often a Cross-Encoder) that evaluates the specific pair of Query + Product to assess relevance with high precision. The reranker can also incorporate business logic, such as boosting high-margin items or demoting out-of-stock products, before presenting the final list to the user.

V. BUILD V/S BUY STRATEGIES

E-commerce leaders face a strategic choice: adopt a comprehensive "black box" platform from a hyperscaler or build a custom "glass box" solution using specialized components.

A. Hyperscalers: Google Vertex AI Search for Retail

Google offers a compelling, fully managed proposition with Vertex AI Search for Retail. This solution leverages Google's deep graph of product knowledge and its advanced Gemini and PaLM models. It excels in "understanding" user intent without extensive manual tuning. A key differentiator is its out-of-the-box Conversational Commerce capability, which allows for multi-turn interactions (e.g., user searches for "dress," then refines with "show me longer ones"). It also deeply integrates with Google Merchant Center, simplifying data ingestion for existing Google ad customers.

However, this convenience comes with opacity. Retailers have limited visibility into why a specific product ranked #1, and the data resides within Google's ecosystem, raising potential competitive concerns for some large retailers.

B. Specialized Search Engines: Algolia and Bloomreach

Algolia has positioned itself as the speed leader. Its NeuralSearch product combines its famous milliseconds-fast keyword engine with vector-based "Neural Hashing." Algolia focuses on the "search-as-you-type" experience, aiming for <20ms latency. It is highly developer-friendly, offering extensive APIs and UI kits, and provides transparent controls for merchandising (e.g., pinning products, boosting brands). This contrasts with Google's black-box approach, giving merchants more control over their digital shelf.

Bloomreach takes a broader "Commerce Experience" approach, integrating search deeply with a Customer Data Platform (CDP). Its engine doesn't just look at the query; it looks at the user issuing the query. If a user has a history of buying discount items, Bloomreach's algorithms can adjust the search results to favor lower-priced items, even for a generic query like "shirt." This personalization layer is a key value add for increasing Customer Lifetime Value (CLV).

C. Case Studies in Strategic Implementation

Real-world deployments highlight how different strategies yield results:

- eBay: Leveraging its massive scale (20 billion images), eBay built a custom multimodal stack using BERT for text and ResNet-50 for images. They employed a specific training technique called "triplet loss" to teach the model to distinguish between closely related items. A key focus was spotting mismatches between images and descriptions (e.g., a listing for a "PS5" with a picture of a generic controller). This implementation drove a 15.9% increase in Click-Through Rate (CTR) and a massive 31.5% jump in Purchase-Through Rate, proving that accurate visual-text alignment directly drives sales.
- Focal: This platform utilized the Google Product Taxonomy (GPT) to structure its vector space. By mapping vectors to a predefined tree of 5,500 categories, Focal optimized for precision in categorization. Their system achieved an 8% boost in "leaf precision" (correctly identifying the specific sub-category, like "women's leather gladiator sandals" rather than just "shoes"). This approach is particularly powerful for handling "long-tail" queries where specific attributes are crucial.
- The Iconic: This fashion retailer implemented multimodal search to address the "null results" problem. By allowing the semantic engine to find "similar" items when exact matches were unavailable, they reduced null searches from 5% to nearly 0%. This retention of traffic that would otherwise have bounced contributed to a 2.6% increase in revenue, illustrating that search improvements often flow directly to the bottom line.

VI. OPTIMIZING FOR THE MACHINE GAZE: DATA HYGIENE AND SEO

For multimodal search to function effectively, the underlying product data must be optimized not just for human eyes, but for the "machine gaze." AI models "see" differently than humans; they parse pixel-level tokens and semantic text structures, and they are sensitive to noise that humans ignore.

A. Image Hygiene and Computer Vision

Images are the primary data payload for visual search. Technical hygiene is critical.

- Resolution and Artifacts: Aggressive compression (like low-quality JPEGs) creates visual artifacts. To a computer vision model, these artifacts act as "noise" that can lead to hallucinations. A blurry logo might be misclassified as a different object, or a texture might be misread. High-resolution source files are essential for accurate vectorization.
- OCR readability: Models often use Optical Character Recognition (OCR) to read text on product packaging (e.g.,

"Gluten-Free," "100% Cotton"). If the packaging text is obscured by glare, low contrast, or stylized fonts, the model misses this critical attribute. Packaging photography should be treated as a machine-readability feature, ensuring clear, high-contrast text visibility.

- Emotional Alignment: Advanced models like CLIP assess the "sentiment" or "vibe" of an image. If a retailer sells "fun summer dresses" but uses high-fashion photography models with moody, neutral expressions, the semantic vector of the image might drift toward "somber" or "formal," causing it to rank lower for "fun" queries. Visual sentiment must align with the intended search intent.

B. Structured Data and Schema.org

To assist search engines (both internal and external like Google) in understanding the product, structured data is non-negotiable.

- Product Schema: Retailers must implement robust Schema.org/Product markup. This includes standard fields like name, description, SKU, and offers.
- 3D and AR Schema: With the rise of 3D commerce, the 3DModel schema type is becoming essential. Using properties like encoding Format to specify glTF or USDZ files signals to search engines that the product has interactive assets available.
- Shipping Details: As conversational queries like "boots that arrive by Friday" increase, the shipping details property in schema becomes critical. It allows the search engine to deterministically answer logistics questions without needing to parse unstructured text on a shipping policy page.
- Merchant Listings: Google's "Merchant Listing" markup is specifically designed for pages where a transaction can occur. It requires precise pricing, availability, and return policy information, directly influencing eligibility for rich product results in Google Search.

VII. The Third Dimension: 3D Assets, AR, and Spatial Indexing

The definition of "multimodal" is rapidly expanding to include 3D spatial data. This is particularly transformative for the "Home & Garden," "Furniture," and "Fashion" verticals, where the ability to visualize a product in 3D is a strong driver of conversion and a reducer of returns.

A. The Format War: glTF v/s USDZ

Two dominant file formats have emerged for 3D commerce, each championed by different tech giants.

- glTF (GL Transmission Format): Often called the "JPEG of 3D," glTF is an open standard developed by the Khronos Group. It is optimized for efficient transmission and loading on the web. It supports Physically Based Rendering (PBR) materials, which ensures realistic lighting and textures. It is the preferred format for Android devices and web-based 3D viewers.
- USDZ (Universal Scene Description): Developed by Pixar and championed by Apple, USDZ is a zero-compression archive format. It is the native format for AR Quick Look on iOS devices. If a retailer wants an iPhone user to be able to place a virtual sofa in their living room directly from Safari, they must provide a USDZ file.

Most sophisticated 3D pipelines now maintain a "master" asset (often in high-fidelity formats like FBX or OBJ) and automatically transcode it into optimized glTF and USDZ derivatives for delivery.

B. Optimization for Search Performance

3D assets are heavy. An unoptimized 3D model can be tens of megabytes, destroying page load speed (Core Web Vitals) and negatively impacting SEO.

- Geometry Optimization: Retailers must use Draco compression, a library for compressing geometry (vertices and polygons). It can reduce glTF file sizes by significantly without visible loss of quality.
- Texture Optimization: Textures should be compressed (using WebP or JPEG) and "atlased" (combined into a single large image file) to reduce the number of HTTP requests required to load the model.
- Level of Detail (LOD): Just as maps load details as you zoom in, 3D engines should serve lower-polygon versions of a model when it is small on screen, loading the high-fidelity version only when the user interacts with it.

C. Spatial Indexing and the USD Search API

The frontier of search is not just finding a 3D model but searching within it. NVIDIA's USD Search API allows developers to index massive libraries of OpenUSD assets. Using AI, it can "read" the 3D scene without manual tagging. A user could query "find all scenes with a mid-century modern chair," and the system analyzes the geometry and metadata of the indexed 3D files to return relevant results. This capability is crucial for B2B marketplaces selling 3D assets to game developers and architects.

Furthermore, Spatial Search is emerging as a consumer capability. By combining AR with search, a user can scan their living room with a LiDAR-equipped phone. The search engine can analyze the space—detecting dimensions, lighting conditions, and existing color palettes—and then recommend furniture that fits both physically and aesthetically. This moves search from "keyword matching" to "environmental fitting".

VIII. STRATEGIC ROADMAP FOR MULTIMODAL ADOPTION

Phase	Technology Focus	Key Metrics	Strategic Goal
1. Foundation	Vector DB, CLIP Embeddings, Hybrid Search	Null Result Rate, MRR (Mean Reciprocal Rank)	Fix "semantic gap" and broken searches. Eliminate zero-result pages.
2. Expansion	Visual Search, Multi-language, Reranking	Click-Through Rate (CTR), Add-to-Cart Rate	Enable "search by image" and improve relevance for long-tail queries.
3. Immersion	AR/3D Indexing (glTF/USDZ), Spatial Search	Interaction Time, Return Rate Reduction	Bridge online/offline gap; reduce returns via accurate 3D visualization.
4. Agentic	API-first Catalog, Agent Protocols, Automated Negotiation	API Transaction Volume, Customer LTV	Prepare for disintermediated, automated consumption by AI agents.

Table 1 : Strategic Roadmap for Multimodal Adoption

IX. CONCLUSION

The transition to multimodal search represents a watershed moment in the history of e-commerce. It is the moment when the digital catalog stops being a static list of text strings and becomes a dynamic, semantic ecosystem that can see, listen, and understand.

The theoretical advantages are clear: vector embeddings dissolve the "semantic gap," VLMs like CLIP automate the understanding of visual assets, and hybrid search architectures bridge the divide between specific and conceptual intent. The financial case is equally robust, with early adopters reporting significant lifts in conversion and revenue by virtually eliminating the "null search" failure state.

However, the operational bar is rising. Success requires a sophisticated infrastructure of vector databases and GPU compute, a rigorous approach to data hygiene that prioritizes the "machine gaze," and a strategic readiness for the agentic future. As AI evolves from a tool that helps humans find products to an agent that buys them on our behalf, the retailers that thrive will be those that re-architect their systems to serve this new, non-human consumer. The future of e-commerce is not just about being visible, it is about being intelligible to the algorithm.

X. REFERENCES

[1] Multimodal embeddings concepts - Image Analysis 4.0 - Foundry Tools | Microsoft Learn, accessed December 28, 2025.
[2] THE ICONIC Case Study | Google Cloud.
[3] Designing Multimodal AI Search Engines for Smarter Online Retail.
[4] Get multimodal embeddings | Generative AI on Vertex AI - Google Cloud Documentation
[5] Pinecone vs Weaviate vs Chroma 2025: Complete Vector Database Comparison| Performance, Pricing, Features - Aloa.
[6] Building a cost-effective image vector search engine with CLIP - Wasim Lorgat.
[7] Conversational Commerce agent overview | Vertex AI Search for commerce.
[8] Image SEO for multimodal AI - Search Engine Land.
[9] How to Optimize 3D Models for the Web: The Complete Guide | by echo3D - Medium.