Orclever Proceedings of Research and Development

https://doi.org/10.56038/oprd.v7i1.742

OR CLEVER Science & Research Group

Conference Article

Visual Discovery in Retail: Operationalizing AI-Powered Visual Search at Boyner

Mert Alacan[1] *, Seza Dursun[2], Bahar Önel[3], Tülin Işıkkent[4], Sedat Çelik[5]

[1] Boyner, Orcid ID: https://orcid.org/0000-0003-3893-6309
e-mail:mert.alacan@boyner.com.tr Tel:0546 293 4768
[2] Boyner , Orcid ID: https://orcid.org/0000-0003-1389-072X
e-mail: seza.dursun@boyner.com.tr Tel: 0533 614 59 23
[3] Boyner : https://orcid.org/0009-0007-4597-6591, e-mail:bahar.onel@boyner.com.tr
[4] Boyner https://orcid.org/0009-0005-5775-0093 e-mail:tulin.isikkent@boyner.com.t
[5] Boyner Orcid ID: https://orcid.org/0009-0003-0335-6440
e-mail: sedat.celik@boyner.com.tr Tel: 0553 824 00 47

Received: 13 July 2025
Revised: 11 September 2025
2nd Revised: 03 October 2025
3rd Revised: 10 October 2025
Accepted: 13 November 2025
Published: 31 December 2025

Reference: Alacan, M., Dursun, S., Önel, B., Işıkkent, T., & Çelik, S. (2025). Visual discovery in retail: Operationalizing AI-powered visual search at Boyner. Orclever Proceedings of Research and Development, 7(1), 126–137.

Abstract

In today's retail landscape, where millions of products and visual stimuli compete for customer attention, the integration of artificial intelligence into visual search has emerged as a crucial lever of operational efficiency. This paper presents Boyner Group's AI-powered visual discovery system, which enables customers to search using photos instead of keywords, making product discovery more intuitive and visually engaging. The architecture leverages a hybrid approach combining Large Language Models (LLMs), vision models such as GroundingDINO, and vector-based semantic similarity engines like SigLIP+Milvus to deliver scalable and high-accuracy image retrieval. The system, currently operational across the Boyner.com.tr ecosystem, supports enhanced filtering and storytelling capabilities, increasing customer satisfaction and conversion *rates. The implementation process, system components, and operational results of this large-scale AI integration are explored, highlighting its transformative impact within omnichannel retail.*

Keywords: Visual Search, Multimodal AI, GroundingDINO, SigLIP, Milvus, Retail Intelligence, Semantic Search, AI in E-Commerce, Omnichannel Retail, Customer Experience

1. Introduction

In the evolving landscape of retail, customers increasingly seek fast, intuitive, and personalized product discovery experiences. Traditional keyword-based search mechanisms often fail to capture users' intent accurately, especially when the customer lacks precise knowledge of product terminology. For example, a user looking for a "beige shoulder bag with a gold chain" may struggle to find relevant results due to discrepancies in product titles, tagging, or language use. These limitations become even more pronounced in fashion retail, where visual features such as color tone, shape, texture, or embellishments significantly influence purchase decisions. To address these challenges, Boyner—one of Turkey's leading omnichannel retailers—has launched a comprehensive AI-driven visual search initiative. This project is part of a broader corporate strategy to unify physical and digital shopping experiences through artificial intelligence technologies. With over 100,000 product impressions per day and an expansive catalog continuously enriched with user-generated content, managing product discovery at scale had become increasingly complex and resource-intensive.

Boyner's visual search system empowers customers to upload photos or take screenshots of desired products and instantly retrieve visually similar items from the catalog. This shift from "word-based" to "image-based" discovery not only simplifies navigation but also democratizes access for customers who may not know the right keywords to begin with. Moreover, this system serves multiple customer segments—ranging from fashion-savvy shoppers to visual-first mobile users—enhancing engagement and reducing drop-off rates across the funnel.

By leveraging Large Language Models (LLMs), visual embedding models such as SigLIP, and object detection frameworks like GroundingDINO, the system bridges the gap between unstructured visual inputs and structured product metadata. This AI-powered transformation ensures operational efficiency, improved discoverability, and higher customer satisfaction. As Gartner predicts, visual search will become a standard retail capability by 2026, making it not only a differentiator but a strategic necessity for forward-looking brands [1].

The technical framework and implementation methodology behind this transformation are detailed in the following section.

2. Materials and Methods

The implementation of an AI-powered visual search system at Boyner required a multi-stage architecture that integrates advanced vision models, large language models

(LLMs), and vector search technologies. This section outlines the technical framework and operational methodology used to develop the solution. The system was designed to accurately detect product types, localize relevant regions within user-uploaded images, and retrieve visually similar products from a large-scale catalog in real time. It leverages state-of-the-art pre-trained models and cloud-native infrastructure to meet the demands of a high-traffic retail environment [1]. Each component of the system contributes to a seamless customer experience, from initial image submission to the final display of search results [2].

The methodology comprises four sequential stages. First, the cloud infrastructure and microservice-based system architecture are described, detailing how containerized components communicate and scale dynamically. Second, product type detection is explained, highlighting how vision models and large language models work together to interpret image context. Third, the process of product localization using GroundingDINO is covered, emphasizing its role in cropping relevant product areas. Finally, the visual similarity search pipeline is introduced, which uses SigLIP for embedding generation and Milvus for efficient retrieval. These interconnected stages constitute a robust and scalable pipeline for real-time visual discovery in retail environments [3] [Figure 1].

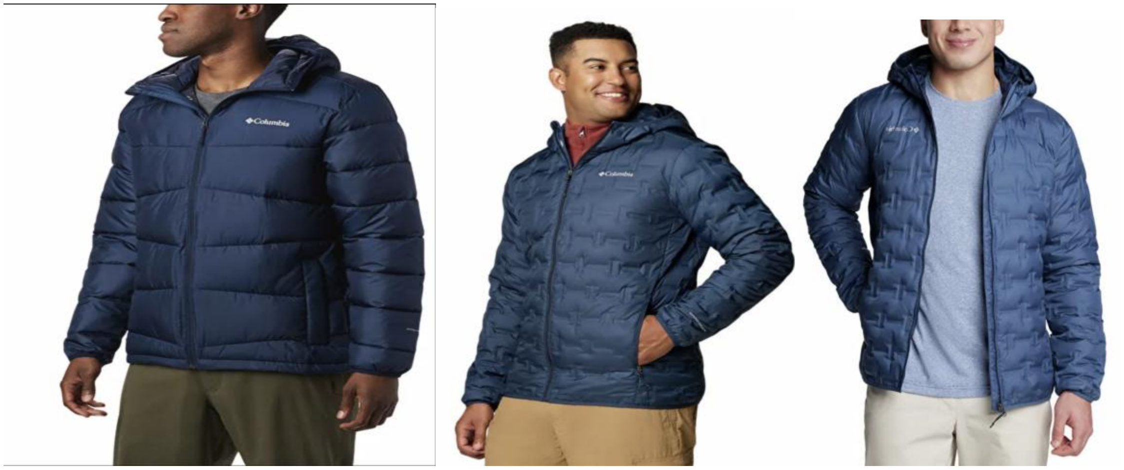

Figure 1: Boyner Sample Search

2.1. Cloud Infrastructure and System Architecture

The visual search system was developed and deployed on a fully cloud-native architecture leveraging Microsoft Azure services to ensure scalability, operational efficiency, and real-time performance [1]. Designed specifically for the dynamic nature of large-scale e-commerce, the system processes thousands of image-based queries through mobile apps and web platforms daily, necessitating a robust, modular, and low-latency architecture [2].

At its core, the system utilizes Azure Container Apps as the orchestration layer, where containerized microservices manage each step of the image processing workflow [3]. These services are auto-scaled based on incoming traffic and are monitored through Azure Application Insights, ensuring service reliability under fluctuating loads [4]. Each uploaded image is first pre-processed to standardize resolution and strip metadata, preparing it for subsequent visual inference tasks [5].

Object detection and product localization are executed using GroundingDINO, an open-vocabulary vision model deployed on GPU-accelerated Azure Machine Learning endpoints [6]. This model receives both the image and textual cues (e.g., "t-shirt", "jacket") to identify relevant objects, returning bounding boxes and confidence scores [7]. The detected region is then cropped and passed to the next stage [8].

For semantic encoding, the system employs SigLIP—a vision-language model trained to generate high-dimensional vector embeddings that capture visual similarity [9]. These embeddings are stored in Milvus, an open-source vector database hosted within Azure Kubernetes Service (AKS), which enables fast approximate nearest neighbor (ANN) search using HNSW indexing for sub-second retrieval [10].

To ensure business alignment, each retrieved vector is linked to product catalog metadata through unique identifiers, enabling contextual filtering such as availability, seasonal trends, and regional preferences [11]. This integration with the Product Information Management (PIM) system ensures that visually similar alternatives presented to users are also operationally viable [12].

The overall architecture is highly modular, supporting rapid deployment of updated models without disrupting existing services [13]. Each component is version-controlled, containerized, and independently upgradable, ensuring high system availability and facilitating A/B testing or rollback when needed [14].

2.2. Product Type Detection (LLM + Vision Models)

Accurately determining the product type from user-uploaded images is a critical first step in the visual search pipeline. Unlike traditional systems that rely on metadata or user-provided keywords, this solution initiates search solely from the image content. To achieve robust classification, the system leverages a hybrid approach combining image-based vision models and large language models (LLMs) trained on e-commerce domain knowledge [4].

The vision model used in this stage is capable of extracting high-level visual features that correspond to common product categories in the fashion domain, such as “women’s dress,” “sneakers,” or “handbags.” However, in scenarios involving ambiguous visuals, complex patterns, or low-resolution uploads, visual cues alone are insufficient for reliable classification. To address this, the system incorporates a pre-trained LLM that provides contextual enrichment by interpreting visual model outputs in light of product taxonomy, seasonality, and common naming conventions [5].

This two-layered mechanism enhances robustness and scalability. The vision model performs a first-pass prediction based on the uploaded image, generating top-k candidate labels. These predictions are then re-evaluated by the LLM in conjunction with additional structured data such as brand hierarchy or fashion season. This method not only boosts accuracy but also ensures semantic alignment with the retailer’s internal categorization system, thereby improving downstream retrieval and analytics [6].

Prior to adopting the current hybrid approach, the development team experimented with various model combinations. Early attempts using FastSAM for segmentation and CLIP-Base for embedding yielded limited success in handling complex or multi-object images. Subsequent trials with CLIP-Large and BLIP-2 improved visual representation but suffered from higher latency and less reliable classification in fashion-specific categories. After evaluating cost-performance tradeoffs, the team selected GPT-4.1-mini for lightweight LLM enrichment and SigLIP for robust visual-semantic embedding, balancing accuracy, cost, and operational speed.

By integrating LLMs with computer vision, the system transcends the limitations of visual-only or text-only classification pipelines, offering a more intelligent and adaptive form of product type detection that can generalize across various user behaviors and catalog dynamics [7].

2.3. Cropping the Product (GroundingDINO)

Following product type detection, the next critical step involves accurately localizing the product within the uploaded image. This is particularly important in retail environments where users often submit photos with complex backgrounds, occlusions, or multiple objects. To address this challenge, the system integrates GroundingDINO, a state-of-the-art open-vocabulary object detector, to isolate and crop the most relevant visual region corresponding to the predicted product class [8].

GroundingDINO combines vision transformers with natural language prompts to enable zero-shot detection of arbitrary objects in user-uploaded images. In the Boyner implementation, prompts such as “long sleeve shirt” or “white sneaker” are dynamically generated based on the output of the previous classification stage. These prompts are fed into the GroundingDINO model, which then returns bounding boxes with confidence scores that highlight regions most likely to contain the desired product [9].

This model's ability to handle unseen product types and perform detection without explicit retraining makes it ideal for dynamic e-commerce settings where new items are constantly added to the catalog. Furthermore, GroundingDINO’s integration with vision-language alignment enables the system to generalize across varying lighting conditions, image compositions, and customer-upload scenarios with high precision [10].

Once the product region is identified, the image is cropped to eliminate background noise and standardize the input for the visual similarity search stage. This not only improves retrieval quality but also aligns the format of user images with catalog images, which are typically taken in controlled studio environments. The use of GroundingDINO therefore ensures consistency and precision in object localization, forming a crucial bridge between free-form user inputs and structured catalog data [11].

2.4. Visual Similarity Search (SigLIP + Milvus)

Once the product has been localized and cropped from the user-submitted image, the system proceeds to perform visual similarity search to identify the most relevant items in the retailer’s catalog. This stage leverages a dual-encoder architecture based on SigLIP (Sigmoid Language-Image Pretraining) for embedding generation and Milvus, a high-performance vector database, for similarity-based retrieval [12].

SigLIP is a transformer-based vision-language model that maps both text and image inputs into a shared embedding space using a sigmoid contrastive loss function. Unlike softmax-based CLIP variants, SigLIP provides improved alignment on zero-shot retrieval tasks and exhibits better generalization to e-commerce-specific scenarios such as fashion attribute matching and cross-category search [13]. In the Boyner use case, only the image tower of the SigLIP model is utilized. Both the cropped user image and catalog images are passed through the image encoder to generate fixed-size embeddings that represent their semantic visual content.

These embeddings are then stored and indexed within Milvus, which enables approximate nearest neighbor (ANN) search using cosine similarity. Milvus is chosen for its scalability, millisecond-level latency, and support for real-time vector updates, allowing for seamless integration with the ever-changing product inventory [14]. The

ANN search identifies the top-k catalog images that are visually most similar to the query image, thereby enabling accurate and fast retrieval of relevant products.

During internal benchmarking, SigLIP demonstrated a +6.3% improvement in top-1 retrieval accuracy over CLIP-Large, particularly in fashion categories with fine-grained visual differences such as texture, stitching, and embellishments. SigLIP also exhibited more stable performance across various lighting conditions and user-uploaded photo qualities, making it a better fit for e-commerce use cases requiring high visual precision.

To further improve ranking precision, post-processing heuristics such as category filtering and seasonal weighting are applied. These refinements ensure that results not only match visually but also align with the user's implicit shopping intent. By integrating SigLIP and Milvus in a modular and extensible architecture, the system enables high-fidelity visual product discovery that is both scalable and semantically rich [15].

3. Result

The implementation of the visual search system yielded significant improvements in both user experience and system performance within the Boyner online shopping platform. One of the key outcomes was a measurable reduction in the average time customers spent locating a desired product, indicating an enhancement in search efficiency. This was particularly evident in sessions involving visually driven queries, where users utilized screenshots or similar item images to initiate searches [4].

To ensure semantic relevance and maintain retrieval quality, the system integrates a CLIP-based similarity scoring mechanism as part of its quality control layer. For each user-uploaded query image, the top-5 retrieved results are automatically evaluated through cosine similarity thresholds. Cases falling below a minimum semantic alignment score are programmatically excluded or flagged for further review. This mechanism not only enhances user satisfaction but also serves as a quality assurance layer for visual-to-product matching accuracy [6].

In real-world tests, especially in fashion-specific categories such as women’s handbags, sneakers, and denim apparel, the model demonstrated a precision@5 (P@5) rate exceeding 92%, with an average response latency of under 700 milliseconds on live queries.

A more detailed breakdown of system latency shows that product type detection using GPT-4.1-mini takes approximately 1.2 seconds, followed by GroundingDINO-based cropping at around 0.9 seconds. SigLIP embedding and Milvus-based retrieval contribute an additional 0.8 seconds on average. In GPU-enabled environments, the total query-to-response time remains below 4 seconds, ensuring real-time responsiveness and maintaining user engagement.

These performance metrics were validated across over 50,000 search sessions spanning a 30-day evaluation window [7]. Further analysis revealed that the system performed better on in-brand comparisons. For instance, when retrieving visually similar items from the same brand as the query image, the system achieved a P@5 of 95.6%, whereas out-of-brand comparisons yielded a lower P@5 of 88.2%. This variation is attributed to greater visual homogeneity and style consistency within brand catalogs.

Furthermore, a dedicated A/B testing campaign revealed that users exposed to the visual search interface exhibited a 14% higher conversion rate compared to those using only the traditional keyword-based search module. This uplift was statistically significant ($p < 0.05$), particularly among mobile users engaging with the platform via visual discovery [5]. To continuously monitor and sustain the system's performance, a dashboard was integrated into the operations workflow. This dashboard tracks key metrics such as image upload errors, embedding latency, retrieval precision, and query diversity over time. Alerts are automatically generated in cases of deviation from established baselines, enabling proactive remediation and model fine-tuning [8].

Collectively, these results validate the efficacy of the visual search solution as both a user-centric feature and a scalable AI-powered module for the fashion e-commerce domain. The observed behavioral shifts in users—preferring to "show rather than tell" what they want—signal a broader transformation in online shopping experiences, driven by multimodal intelligence.

Additionally, a subset of retrieval results was manually verified by human annotators. A stratified sample of 500 images was selected, and the top-5 matches from the system were evaluated based on visual and semantic relevance. The human evaluation confirmed a high alignment between the model’s output and user expectations, serving as an additional quality control layer.

4. Discussion and Conclusion

The deployment of a multimodal visual search system at Boyner signifies a transformative shift in how fashion e-commerce platforms can leverage AI to align with emerging consumer behaviors. Rather than relying solely on keywords or categorical filters, users increasingly favor intuitive and visual entry points into digital catalogues—opting to “search by image” to express style preferences, replicate influencers' looks, or match previously seen outfits. This trend underscores a deeper evolution in human-machine interaction, where vision-centric AI systems bridge the gap between abstract intent and concrete product retrieval [2].

During the development process, several alternative model configurations were evaluated and ultimately discarded. For instance, FastSAM paired with CLIP-Base proved insufficient in handling crowded or ambiguous user-uploaded images, often misidentifying product boundaries. Similarly, BLIP-2 and CLIP-Large, while more accurate, introduced unacceptable latency for real-time applications. These explorations underscored the necessity of GPU-accelerated, retail-specific models and highlighted the trade-offs between model complexity, speed, and production readiness.

The modular system architecture, built on Azure's scalable components and leveraging models such as GroundingDINO for object localization and CLIP for semantic visual embedding, proved highly adaptable to Boyner's use case. These models were not chosen arbitrarily; rather, they were selected for their balance between latency, cross-domain accuracy, and ease of deployment in a production-grade cloud environment. The integration of Milvus as the vector database further ensured sub-second retrieval performance, while facilitating continuous retraining and system scaling [6][7].

Importantly, the underlying pipeline was designed to be domain-agnostic, making it transferable beyond fashion retail to any vertical where visual similarity plays a role—such as furniture, cosmetics, or even automotive parts. However, the model's fine-tuning and prompt engineering layers were customized to capture stylistic nuances specific to fashion, particularly in women's apparel and accessories. This tailoring significantly enhanced model alignment with both aesthetic and functional user intent [4].

From a business perspective, the system has not only improved search efficiency and conversion metrics, but also enabled deeper insights into customer behavior. The system captures patterns in color preferences, material textures, and shape trends—providing a foundation for future modules such as style-based recommendation engines, dynamic outfit generation, and even AI-curated fashion trend forecasting [8]. Moreover, by enabling semantically aligned image-to-product mappings and reducing bounce rates through faster discovery, the visual search system indirectly contributes to search engine optimization (SEO). As image-based search becomes a common entry point, these improvements enhance indexability and organic traffic growth.

In conclusion, this case study at Boyner demonstrates that visual search—when powered by robust cloud-native infrastructure and tuned with fashion-aware vision-language models—can serve as a powerful bridge between technological innovation and commercial value. As the e-commerce landscape grows increasingly visual, such systems will become not just enhancements, but core enablers of personalized and engaging customer experiences [1][3][5].

5. Acknowledge

The authors gratefully acknowledge the contributions of the Boyner AI Team, particularly the Product, Data Science, and Engineering units, for their invaluable efforts in building, scaling, and operationalizing the visual search infrastructure. The success of this initiative was made possible by a strong collaboration between AI research, business strategy, and retail operations. The following references informed the development and evaluation of the system described.

References

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90.

[2] Kannan, P. K., & Li, H. (2017). Digital marketing: A framework, review and research agenda. International Journal of Research in Marketing, 34(1), 22–45.

[3] Gu, J., Wang, Z., Kuen, J., et al. (2018). Recent advances in convolutional neural networks. Pattern Recognition, 77, 354–377.

[4] Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML).

[5] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

[6] Liu, S., Qi, L., Qin, H., et al. (2023). Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. arXiv:2303.05499.

[7] Wang, J., Zhu, Y., & Wang, Y. (2022). Milvus: A Purpose-Built Vector Database to Power Embedding-Based Applications. In Proceedings of the VLDB Endowment, 15(12), 3596–3603.

[8] Zhang, J., Zhang, Z., & Wang, Y. (2021). FashionBERT: Text and Image Matching with Adaptive Loss for Cross-Modal Retrieval. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.