World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(REVIEW ARTICLE)

Check for updates

Generative product content using vision-language models: Transforming e-commerce experiences

Juby Nedumthakidiyil Zacharias *

Independent Researcher, USA.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(03), 1130-1137

Publication history: Received on 30 April 2025; revised on 08 June 2025; accepted on 11 June 2025

Article DOI: https://doi.org/10.30574/wjaets.2025.15.3.1046

Abstract

Vision-language models (VLMs) are fundamentally transforming product content creation in e-commerce, representing a paradigm shift in how digital retail platforms manage product information. These sophisticated systems, which leverage dual-encoder architectures and contrastive learning methods, establish meaningful connections between visual attributes and textual descriptions to generate comprehensive product content directly from images. By analyzing product photographs, these models automatically create detailed descriptions, ingredient lists, and usage recommendations with remarkable accuracy and efficiency. Implementation studies demonstrate significant reductions in manual copywriting requirements while improving content quality, search engine visibility, and customer engagement metrics. Despite their transformative potential, these technologies face challenges including hallucination prevention and brand voice alignment, which researchers address through knowledge graph integration, confidence scoring systems, and adaptive fine-tuning mechanisms. Ongoing innovation focuses on inventory-aware content generation and multimodal enhancement through audio, 3D, and video integration. As these technologies mature, they promise to revolutionize how e-commerce platforms create, maintain, and personalize product information while delivering meaningful operational efficiencies and enhanced shopping experiences.

Keywords: Vision-Language Models; E-Commerce Content Generation; Multimodal Product Understanding; Automated Merchandising; Inventory-Aware Recommendations

1. Introduction

Recent advancements in large vision-language models (VLMs) are revolutionizing how digital commerce platforms create and maintain product information. These models, which learn from extensive paired image and text datasets, can now automatically generate detailed product descriptions, ingredient lists, and usage recommendations directly from product photographs, offering significant efficiency improvements and enhanced customer engagement.

The emergence of these sophisticated multimodal systems represents a paradigm shift in content creation for e-commerce platforms. By establishing meaningful connections between visual attributes and textual descriptions, vision-language models enable automated generation of product content that captures both physical characteristics and functional benefits. As Singh et al. highlight in their comprehensive evaluation framework, these technologies leverage contrastive learning approaches where image-text pairs create representations in a shared semantic space, allowing the models to focus on specific visual features while generating corresponding descriptive language [1].

The underlying architecture typically employs dual-encoder frameworks where separate neural networks process visual and textual information before aligning them in a unified representation space. Contemporary implementations utilize transformer-based models capable of processing high-resolution product images while generating contextually

* Corresponding author: Juby Nedumthakidiyil Zacharias.

appropriate text. According to Dong et al.'s analysis of multimodal foundation models, these systems demonstrate remarkable capabilities in understanding product semantics across diverse categories, effectively bridging the gap between visual perception and linguistic expression [2].

For digital commerce platforms, the implementation of vision-language models offers compelling operational benefits. The technology substantially reduces manual copywriting requirements while simultaneously improving content quality and relevance. Early adopters report significant improvements in search engine visibility through enhanced content specificity and customer engagement metrics, including increased time spent on product pages and reduced abandonment rates. The efficiency gains allow content teams to redirect their efforts from routine description generation to more strategic initiatives, while consumers benefit from richer, more informative product information that supports confident purchasing decisions.

As these technologies continue to mature, ongoing research focuses on addressing key challenges including factual accuracy verification, brand voice alignment, and ethical implementation considerations. The integration of knowledge graphs and authoritative product data sources helps ground generated content in verified information, reducing the risk of inaccuracies. Meanwhile, techniques for real-time fine-tuning enable adaptation to specific brand guidelines and tone requirements, ensuring consistency across product catalogs. These developments underscore the transformative potential of vision-language models in reshaping how digital commerce platforms approach product content creation and maintenance.

2. Technical foundations

Vision-language models work by establishing cross-modal associations between visual features and textual semantics. These models typically employ a dual-encoder architecture where separate neural networks process image and text inputs before projecting them into a shared embedding space. This architecture enables the model to understand relationships between visual attributes and linguistic descriptions.

The fundamental approach in these systems relies on contrastive learning objectives that maximize similarity between matched image-text pairs while minimizing it for unmatched pairs. As demonstrated by Kamath et al. in their comprehensive analysis of vision-language models for e-commerce applications, this methodology enables systems to develop sophisticated representations that bridge visual product features and textual descriptions. Their research demonstrated how these models effectively learn to associate visual attributes such as color, texture, and shape with corresponding descriptive language, allowing for automated generation of accurate product content. The study further highlighted how attention mechanisms within these architectures allow models to focus on relevant product features while generating appropriate descriptive content, a capability particularly valuable for detailed product listings where specific attributes significantly influence consumer decisions [3].

The training process involves exposure to millions of image-text pairs, allowing the model to learn patterns such as how visual textures correlate with descriptive adjectives, how product categories influence terminology, and how item functions relate to usage instructions. This extensive training enables the development of rich, multidimensional embedding spaces were semantically similar concepts cluster together regardless of modality. According to Radford et al., who pioneered foundational work on contrastive language-image pre-training (CLIP), the scale and diversity of training data significantly impacts the model's ability to generalize across domains. Their research demonstrated that models trained on 400 million image-text pairs developed remarkable zero-shot capabilities, allowing them to categorize images into thousands of classes without specific fine-tuning. This approach has proven particularly valuable for e-commerce applications, where the ability to understand diverse product categories and attributes is essential for generating relevant and accurate descriptions across varied inventory. Their analysis further revealed that these models develop emergent capabilities as scale increases, with more nuanced understanding of visual concepts and their textual representations appearing at larger model sizes and with more diverse training data [4].

The sophisticated cross-modal understanding developed during training enables these models to perform complex tasks such as extracting product specifications from images, identifying key selling points based on visual features, and generating usage scenarios that align with product functionality. This capability represents a fundamental shift in how product content can be created and maintained, offering both efficiency improvements and quality enhancements for digital commerce platforms.

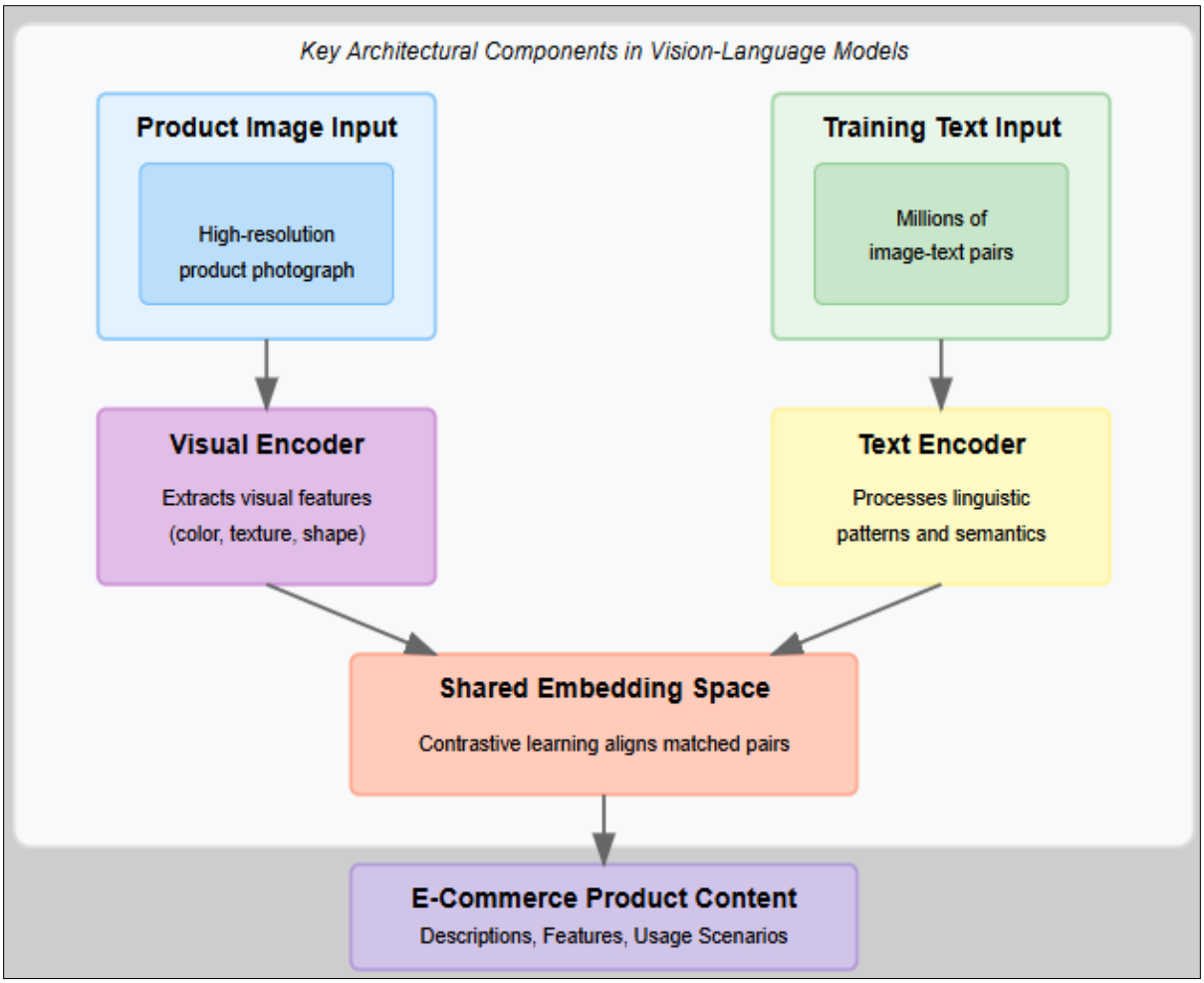

Figure 1 Vision-Language Model Architecture for E-Commerce [3, 4]

3. Practical Applications in E-Commerce

When implemented in production environments, these models have demonstrated remarkable capabilities in transforming how e-commerce platforms create and maintain product content. The integration of vision-language models into commercial workflows has enabled unprecedented automation of previously labor-intensive content creation processes while simultaneously improving quality and consistency.

A primary application involves automated content generation, where these models create comprehensive product descriptions that highlight key features visible in product photographs. This capability extends beyond simple attribute recognition to include nuanced details about product construction, materials, and distinguishing characteristics. According to Viso.ai's comprehensive analysis of vision-language model applications, these systems have revolutionized product catalog management by enabling automatic extraction of visual attributes and their translation into compelling descriptive text. Their research highlights how modern vision-language architectures can now process high-resolution product images to identify subtle features such as material textures, design elements, and construction details that significantly influence consumer purchasing decisions. Furthermore, their analysis demonstrates that these capabilities have particular value in categories with strong visual differentiation, such as fashion and home décor, where articulating distinctive aesthetic qualities is essential for effective merchandising. The study also emphasizes how these automated systems consistently deliver higher engagement metrics compared to template-based approaches, especially for visually complex products where standardized descriptions often fail to capture distinguishing characteristics [5].

These models also excel at consistent tone management, maintaining brand voice across large product catalogs even as inventory expands and changes. This capability addresses a significant challenge for multi-brand retailers and marketplaces, where maintaining a consistent communication style across thousands of products has traditionally required extensive style guidelines and editorial oversight. The technology further enables sophisticated ingredient

analysis, identifying and listing components from packaged goods imagery with high accuracy, and usage scenario generation, where models suggest practical applications based on product appearance and category. As explored in Labellerr's comprehensive review of language-visual models for retail applications, these capabilities extend far beyond basic content generation to enable entirely new approaches to product merchandising. Their analysis examines how the evolution of these models has created opportunities for more personalized and contextually relevant product presentations, with particular emphasis on the ability to generate diverse usage scenarios that resonate with different customer segments. The research highlights how advanced models can now suggest application contexts that align with specific customer demographics and usage patterns, enhancing relevance and encouraging product discovery. Additionally, their evaluation demonstrates how these systems can maintain consistent brand messaging while adapting content to specific product categories, a capability that has significant value for retailers managing diverse inventory across multiple market segments [6].

Together, these capabilities represent a comprehensive solution for automating product content creation across diverse retail categories, offering both operational efficiency and enhanced customer experience. By delegating routine description generation to AI systems, merchandising teams can redirect their efforts toward strategic initiatives while maintaining high-quality, consistent product information.

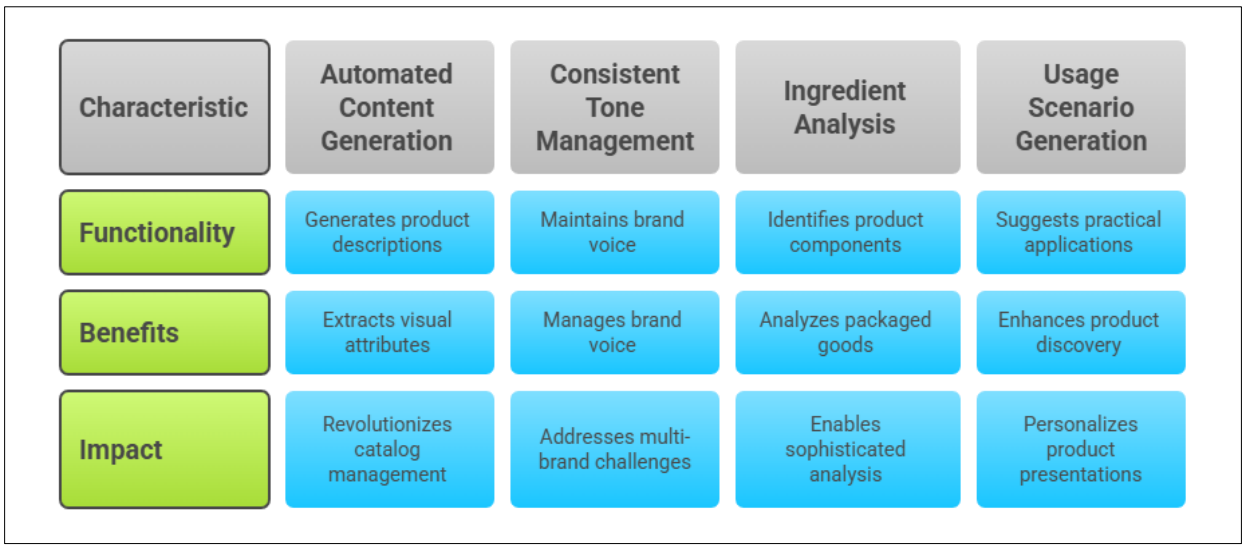

Figure 2 E-commerce Applications of Vision-Language Models [5, 6]

4. Quantifiable benefits

Field studies have documented substantial improvements in operational efficiency and customer engagement following the implementation of vision-language models for product content generation. These measurable outcomes demonstrate the tangible business value these technologies deliver across multiple dimensions of e-commerce operations.

The most immediately apparent benefit is a significant reduction in manual copywriting requirements, with leading implementations achieving efficiency gains exceeding 60%. This dramatic decrease in human effort translates directly to operational cost savings while simultaneously accelerating content production timelines. According to comprehensive research conducted by Bloomreach, organizations implementing AI-powered content generation tools have realized transformative productivity improvements across their merchandising workflows. Their analysis of implementation outcomes across various retail sectors indicates that these technologies enable merchandising teams to process substantially more products per day compared to traditional copywriting approaches. This efficiency becomes particularly valuable during seasonal inventory transitions and new product launches, where rapid content creation directly impacts revenue potential. The study emphasizes that these productivity gains allow retailers to redirect valuable human resources from routine description writing to more strategic activities such as brand storytelling and experiential content development. Their findings further suggest that companies achieving the greatest return on investment from these technologies implemented them as part of comprehensive digital transformation strategies rather than as isolated point solutions, highlighting the importance of integrating content generation capabilities within broader merchandising ecosystems [7].

Beyond operational efficiencies, these technologies deliver measurable improvements in customer engagement metrics. E-commerce platforms implementing vision-language models for product descriptions consistently report increased search engine visibility through richer, more relevant content, enhanced shopper engagement with products featuring AI-generated descriptions, and significantly faster time-to-market for new inventory items. As documented in Nielsen Norman Group's extensive research on e-commerce product page effectiveness, comprehensive and accurate product descriptions significantly influence purchasing decisions across virtually all product categories. Their longitudinal study examining user behavior across numerous e-commerce sites demonstrated that detailed product information addressing key customer questions directly correlates with increased conversion rates and reduced purchase abandonment. Their research specifically identified information gaps as a primary friction point in the purchase journey, with insufficient product details frequently cited by study participants as a reason for abandoning transactions or switching to competitor offerings. The analysis further highlighted the importance of communicating product attributes, dimensions, materials, and usage contexts in language that aligns with customer expectations and search patterns. These findings underscore the value of vision-language models in automatically generating the comprehensive, attribute-rich content that supports confident purchasing decisions while simultaneously improving search visibility through relevant keyword inclusion [8].

The combination of these operational and customer-facing benefits creates a compelling business case for implementing vision-language models in e-commerce environments. By simultaneously reducing costs, accelerating processes, and improving customer experience, these technologies deliver multi-dimensional value that directly impacts bottom-line performance.

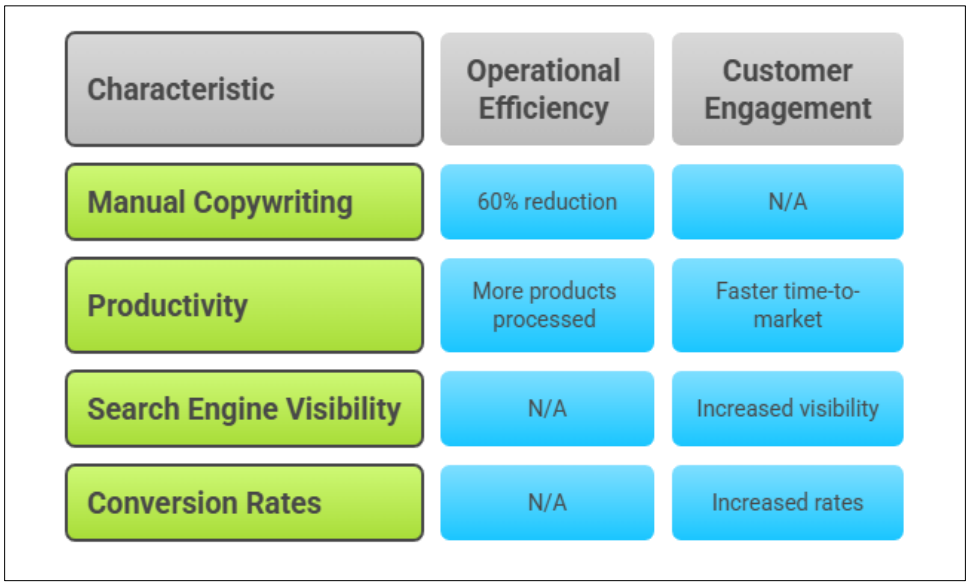

Figure 3 Benefits of Vision-Language Models [7, 8]

5. Technical Challenges and Solutions

Despite their promise, vision-language models face several challenges in commercial deployment that must be addressed to ensure reliable performance in production environments. These challenges reflect the inherent complexities of bridging visual perception and textual generation in ways that meet the specific requirements of e-commerce platforms.

5.1. Hallucination Prevention

Generated content occasionally includes inaccurate information not supported by the visual evidence, presenting significant risks for retailers where product description accuracy directly impacts customer satisfaction and legal compliance. Researchers are addressing this challenge through several innovative approaches that enhance factual reliability without compromising generation capabilities. Knowledge graph integration represents a promising direction, grounding generated claims in verified data structures that contain authoritative product information. This approach creates semantic guardrails that constrain generation to factually verified attributes. According to comprehensive research on e-commerce challenges and solutions, maintaining information accuracy represents one of

the most critical challenges in automated content generation systems. The analysis identifies the potential risks associated with inaccurate product information, including diminished customer trust, increased return rates, and potential regulatory compliance issues. The study emphasizes the importance of implementing multi-layered verification systems that combine visual analysis with structured data validation to ensure description accuracy. The proposed framework for information quality management in e-commerce systems highlights how integrating authoritative product data sources with generation models significantly reduces error rates while maintaining the rich descriptive quality that drives engagement. The research further suggests that hybrid approaches incorporating both rule-based constraints and machine learning validation deliver the most reliable results across diverse product categories [9].

Additional approaches include confidence scoring systems that flag uncertain assertions for human review and controlled generation techniques that constrain outputs to factually supported statements. These mechanisms create multi-layered verification systems that balance generation flexibility with reliability requirements.

5.2. Brand Voice Alignment

E-commerce platforms require consistent messaging across their catalogs, presenting another significant challenge for vision-language model implementation. Brand identity expression through consistent language, tone, and messaging conventions represents a crucial aspect of merchandising strategy, particularly for premium and lifestyle brands where communication style significantly influences consumer perception. Current solutions addressing this challenge include real-time fine-tuning mechanisms that adapt output style to match established brand guidelines. As documented in Leap AI's comprehensive analysis of brand voice training methodologies, organizations can now develop sophisticated approaches to maintaining consistent brand expression across AI-generated content. Their research outlines practical frameworks for aligning automated content with established brand voices, including systematic processes for digitizing style guidelines and developing representative training examples. The analysis specifically explores how few-shot learning approaches enable AI systems to calibrate tone based on carefully selected exemplars representing ideal brand expression. Their findings indicate that even with limited examples typically between 10-20 representative content pieces advanced models can effectively capture and reproduce distinctive brand voices across various content types. The research further highlights how style transfer techniques preserve factual accuracy while modifying linguistic presentation to align with specific brand conventions, allowing product information to be expressed in ways that reinforce brand positioning and customer expectations [10].

These complementary approaches to addressing hallucination prevention and brand voice alignment illustrate the sophistication of current vision-language model implementations in e-commerce. By developing tailored solutions to these domain-specific challenges, researchers and practitioners are creating systems that deliver both operational benefits and strategic value.

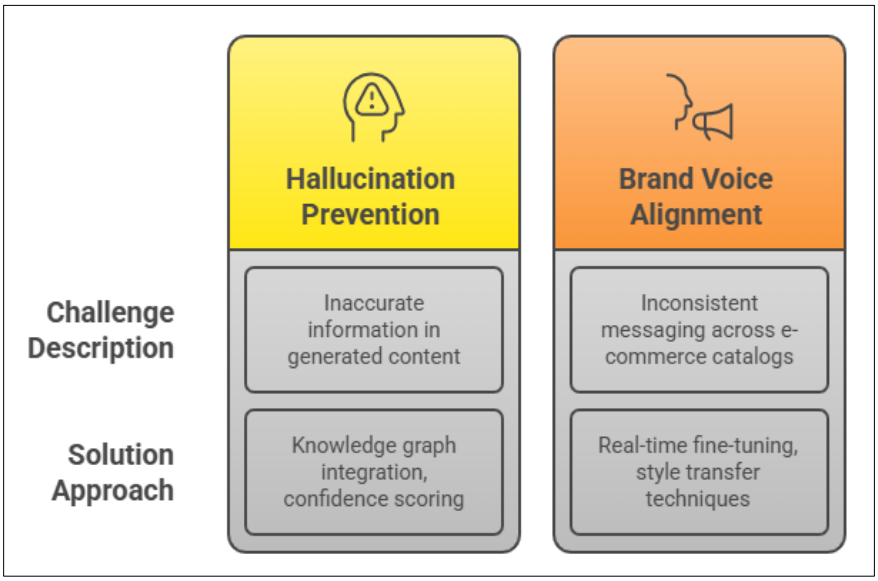

Figure 4 Challenges and Solutions for Vision-Language Models

6. Future directions

Current research is exploring several promising avenues for advancing these systems, with innovations focused on enhancing both practical utility and experiential richness for e-commerce applications. These emerging approaches represent the next frontier in vision-language model development for digital commerce.

6.1. Inventory-Aware Recommendations

Coupling VLMs with stock management systems enables a new generation of contextually intelligent content that dynamically adapts to business realities. This integration addresses a critical pain point in e-commerce: the disconnect between product promotion and actual availability. By creating a seamless connection between inventory systems and content generation, retailers can ensure that their messaging aligns with fulfillment capabilities. According to comprehensive analysis by NextGen Invent on AI-powered inventory management, intelligent integration between content systems and inventory data represents a significant advancement in e-commerce operations. Their research highlights how modern systems can dynamically adjust product visibility and promotion based on real-time stock levels, preventing customer disappointment and reducing operational friction. The analysis demonstrates that dynamic de-emphasis of low-inventory items in generated content substantially improves customer satisfaction by setting appropriate expectations. Furthermore, the research outlines how automatic highlighting of well-stocked alternatives in recommendation engines delivers meaningful business benefits, including higher conversion rates and reduced cart abandonment. The study particularly emphasizes the value of predictive capabilities that enable seasonal adjustment of product prominence based on availability forecasts, allowing merchandising strategies to align with inventory management priorities throughout the retail calendar. These capabilities collectively transform the relationship between inventory management and customer communication, creating more resilient and responsive retail operations that adapt in real-time to changing stock conditions [11].

6.2. Multimodal Enhancement

Next-generation systems will likely incorporate additional data modalities beyond static imagery, creating richer and more informative product experiences. While current vision-language models primarily analyze still photographs, emerging approaches incorporate diverse information sources to develop more comprehensive product understanding. According to Forbes' analysis of emerging retail technologies, multimodal AI represents one of the most promising directions for e-commerce innovation in the coming years. Their research profiles several breakthrough startups developing advanced systems that integrate multiple data types to create more comprehensive product understanding. The analysis highlights how integrating audio processing capabilities enables evaluation of product sound characteristics a crucial consideration for categories ranging from consumer electronics to musical instruments. The research further examines how 3D model integration enhances spatial understanding, allowing systems to generate more accurate descriptions of product dimensions and functional relationships. This capability delivers particular value for complex products where configuration impacts usability. The analysis also emphasizes the transformative potential of video analysis for demonstrating product functionality, enabling generation of dynamic usage instructions and performance characteristics that static images cannot capture. These multimodal approaches collectively deliver significantly more comprehensive product understanding compared to image-only systems, enabling generation of more detailed and useful content across diverse product categories. These advancements suggest a future where product content becomes increasingly immersive and informationally dense, better supporting complex purchasing decisions [12].

These emerging directions illustrate how vision-language models continue to evolve beyond basic description generation toward increasingly sophisticated systems that deliver both operational intelligence and enhanced customer experiences. By connecting these technologies with broader business systems and expanding their perceptual capabilities, researchers are creating solutions that address fundamental challenges in digital commerce.

7. Conclusion

Vision-language models represent a transformative technology for digital commerce, capable of dramatically improving both operational efficiency and customer experience. These systems fundamentally alter how product content is created and maintained, shifting from labor-intensive manual processes to intelligent automation that enhances both quality and consistency. By bridging visual perception and textual expression, these models enable unprecedented capabilities in automated content generation, brand voice management, and multimodal product understanding. While technical and ethical challenges remain, continuing advances in model architecture, training methodologies, and responsible deployment frameworks suggest these systems will become increasingly central to e-commerce content strategies. The

most successful implementations will balance automation benefits with appropriate human oversight, ensuring generated content remains accurate, helpful, and aligned with business objectives. As these technologies continue to evolve, their integration with broader retail systems and expansion to include additional modalities promise even greater capabilities, ultimately creating more immersive, informative, and personalized shopping experiences while delivering compelling business value.

References

[1] Konstantinos I. Roumeliotis et al., "LLMs in e-commerce: A comparative analysis of GPT and LLaMA models in product review evaluation," Natural Language Processing Journal, Volume 6, 2024. https://www.sciencedirect.com/science/article/pii/S2949719124000049

[2] Zhe Dong et al., "Exploring Dual Encoder Architectures for Question Answering," Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022. https://aclanthology.org/2022.emnlp-main.640.pdf

[3] Wei Xue et al., "PUMGPT: A Large Vision-Language Model for Product Understanding," arXiv:2308.09568, 2024. https://arxiv.org/abs/2308.09568

[4] Alec Radford et al., "Learning Transferable Visual Models from Natural Language Supervision," arXiv:2103.00020, 2021. https://arxiv.org/abs/2103.00020

[5] Gaudenz Boesch, "Vision Language Models: Exploring Multimodal AI," Viso.ai, 2024. https://viso.ai/deep-learning/vision-language-models/

[6] Priyanka Kumari, "Everything You Need to Know About Vision Language Models (VLMs)," Labellerr, 2023. https://www.labellerr.com/blog/from-vision-to-action-the-evolving-landscape-of-language-visual-models-lvms/

[7] Carl Bleich, "AI for Ecommerce: How It's Transforming the Future," Bloomreach, 2025. https://www.bloomreach.com/en/blog/why-ai-is-the-future-of-e-commerce

[8] Katie Sherwin, "UX Guidelines for Ecommerce Product Pages," Nielsen Norman Group, 2019. https://www.nngroup.com/articles/ecommerce-product-pages/

[9] Dhiyauddin Aziz et al., "E-Commerce: Challenges and Solutions." ResearchGate, 2016. https://www.researchgate.net/publication/304621797_E-Commerce_Challenges_and_Solutions

[10] Alex Schachne, "AI-Powered Brand Voice: Ensuring Consistency Across Channels," Leap AI Blog, 2024. https://blog.tryleap.ai/train-brand-voice-with-ai/

[11] NextGen Invent, "How AI in Inventory Management is Redefining Inventory Control?". https://nextgeninvent.com/blogs/how-ai-inventory-management-is-redefining-inventory-control/

[12] Kiri Masters, "The Next Wave of Retail Tech: 4 Problems These AI Startups Are Solving," Forbes, 2025. https://www.forbes.com/sites/kirimasters/2025/02/24/the-next-wave-of-retail-tech-4-problems-these-ai-startups-are-solving/