ADVANCES IN CYBER-PHYSICAL SYSTEMS

Vol. 10, No. 1, 2025

CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING (CLIP) IN E-COMMERCE: APPLICATIONS, METHODOLOGIES, AND PERFORMANCE

Oleksandr Khainas[1]***, Nataliia Melnykova***[1, 2]***, Solomiia Fedushko***[1, 3]

[1] *Lviv Polytechnic National University, 12, S. Bandery str., Lviv, 79013, Ukraine,*
[2] *Uniwersytet Rolniczy im. Hugona Kołłątaja, al. Adama Mickiewicza 21, 31-120 Kraków, Poland,*
[3] *Comenius University Bratislava, Odbojárov 10, 820 05 Bratislava, Slovakia*
Authors' e-mails: *oleksandr.y.khainas@lpnu.ua, nataliia.i.melnykova@lpnu.ua,*
solomiia.s.fedushko@lpnu.ua, solomiia.fedushko@fm.uniba.sk

https://doi.org/10.23939/acps2025.01.100

Submitted on 16.04.2025

Abstract: **This article thoroughly examines the architecture and applications of the Contrastive Language-Image Pre-training (CLIP) model within the e-commerce domain, focusing on key tasks such as visual search, product recommendation, and attribute extraction. The article also provides an in-depth analysis of the methodologies used for CLIP's adaptation to e-commerce tasks and the relevant datasets employed. By highlighting the unique capabilities of the CLIP model, such as its ability to perform zero-shot learning and contrastive pre-training, this article underscores its potential impact on the industry while also acknowledging its limitations, including the "domain gap" and the need for adaptation strategies. Furthermore, the article explores the future research directions for enhancing CLIP's performance in specialized e-commerce contexts and compares it with other traditional and multimodal AI techniques.**

Index terms: **CLIP, multimodal AI, E-commerce, visual search, product recommendation, attribute extraction, contrastive learning, zero-shot learning, catalog management.**

I. INTRODUCTION

The exponential growth of the e-commerce sector presents both immense opportunities and significant operational challenges. Platforms grapple with vast, ever-expanding product catalogs, demanding sophisticated solutions for efficient product discovery, robust personalization, and streamlined catalog management [1]. A primary hurdle is the inherent "semantic gap" – the difficulty traditional systems face in accurately mapping the intent behind user queries, whether textual or visual, to relevant products represented through diverse data modalities like images and textual descriptions [1]. This gap hinders user experience and limits the effectiveness of search and recommendation engines.

The unique ability of CLIP to connect visual and textual data makes it theoretically well-suited to tackle the core data integration and semantic understanding challenges prevalent in modern e-commerce, where product information is inherently multimodal [2]. This article aims to provide a comprehensive scientific review of the application of the CLIP model within the e-commerce domain. It will delve into the model's fundamentals, examine the methodologies used for its adaptation and deployment in e-commerce tasks, evaluate its reported performance, discuss its limitations, and explore its potential impact on the industry [2]. The subsequent sections will cover a literature review and problem statement, the scope and objectives of this review, the methodologies employed, a discussion of results across key application areas, and concluding remarks with future research directions.

II. LITERATURE REVIEW AND PROBLEM STATEMENT

The CLIP model architecture is characterized by its dual-encoder structure [3]. It comprises an image encoder, typically a Vision Transformer (ViT), and a text encoder, usually based on the Transformer architecture. Both encoders function independently to process their respective inputs (images or text snippets) and project them into a shared, fixed-dimensional embedding space. The key innovation lies in the pre-training methodology used to align these two modalities within the shared space.

CLIP employs contrastive learning on a massive scale. During pre-training, the model is presented with large batches of (image, text) pairs, many sourced from the internet (the original OpenAI model used 400 million pairs [3]). The objective is to learn representations such that the cosine similarity between the embeddings of correctly matched image-text pairs is maximized, while the similarity between embeddings of incorrect (mismatched) pairs within the batch is minimized. This is typically achieved by optimizing a symmetric cross-entropy loss (also referred to as multi-class N-pair loss) over the similarity scores computed across all possible pairings in a batch. This process effectively teaches the model to associate visual concepts with their corresponding natural language descriptions.

While CLIP offers a novel approach, various other techniques have been employed in e-commerce AI.

Traditional machine learning and other deep learning models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been used for tasks such as product classification, text-based recommedation, and named entity recognition for attribute extraction [4]. More recently, other multimodal models inspired by architectures like Bidirectional Encoder Representations from Transformers (BERT) or employing different fusion strategies have also been explored [5]. However, these methods often struggle with the sheer scale and heterogeneity of e-commerce data or lack the powerful zero-shot generalization capabilities demonstrated by CLIP.

However, a significant challenge arises from the "domain gap". CLIP models pre-trained on general web data may not optimally perform on specialized e-commerce tasks without adaptation. The visual styles (e. g., studio photography, specific product angles) and specialized terminology (e. g., fine-grained fashion attributes, technical specifications) common in e-commerce differ from the general pre-training data distribution [1]. Studies in related fields like manufacturing quality control have shown that while zero-shot performance might be limited, few-shot or fine-tuned adaptation can yield robust results. This suggests that similar adaptation strategies are likely necessary to unlock CLIP's full potential within specific e-commerce contexts.

III. SCOPE OF WORK AND OBJECTIVES

The purpose of the review paper is to systematically analyze the application of the CLIP model and its derivatives in the e-commerce sector. It aims to consolidate methodologies, critically evaluate performance benchmarks, and summarize common datasets and evaluation metrics used in research on CLIP-based e-commerce applications.

The primary application areas investigated are visual search and cross-modal retrieval, product recommendation, product attribute value extraction, AI-generated content and product design, and catalog management.

The objectives of this scientific article are to systematically analyze the diverse methodologies employed to adapt and utilize CLIP for specific e-commerce tasks, consolidate and critically evaluate the reported performance benchmarks of CLIP-based approaches in these tasks, identify and summarize the common datasets and evaluation metrics prevalent in research assessing CLIP for e-commerce applications, conduct a thorough discussion of the inherent limitations, practical challenges (including domain adaptation, fine-grained understanding requirements, computational overhead), and potential biases associated with deploying CLIP in real-world e-commerce systems, and synthesize the collective findings to propose potential directions for future research in this rapidly evolving field.

IV. COMPARATIVE ANALYSIS OF CLIP-BASED METHODS AND APPLICATIONS

The application of CLIP in e-commerce spans a spectrum of adaptation strategies, ranging from direct zero-shot deployment to sophisticated framework integration. Understanding these methodologies is crucial for assessing the model's practical utility.

CLIP has different adaptation strategies.

Zero-Shot Application is the simplest approach involves using a pre-trained CLIP model directly. For classification or attribute extraction, text prompts are crafted (e. g., "a photo of a {category}", "a photo of a {attribute_value}") and compared against the image embedding. For retrieval, natural language queries are encoded and compared to image or text embeddings in the catalog. While straightforward, its effectiveness can be limited by the domain gap between CLIP's general web training data and specific e-commerce contexts [1]. It often serves as a baseline for more adapted methods.

Few-Shot Learning is less documented specifically for e-commerce CLIP, adapting CLIP with a small number of labeled examples per class has shown promise in related domains like manufacturing. This suggests potential for e-commerce scenarios involving niche product categories or attributes where large labeled datasets are unavailable.

Fine-Tuning is a common strategy that involves further training the pre-trained CLIP model (or parts of it, like the final layers or projection heads) on datasets specific to the e-commerce domain. This helps the model learn domain-specific visual features and terminology. Examples include fine-tuning on fashion datasets or broader e-commerce product data. Studies often compare full fine-tuning (updating all weights) with more parameter-efficient methods like top-tuning or linear probing (only training a classifier head on frozen CLIP features).

Custom Framework Integration is included in several studies that propose novel architectures that incorporate CLIP as a core component but add specialized modules or training procedures tailored to e-commerce tasks (Table 1).

This trend from zero-shot application towards fine-tuning and bespoke framework development indicates that while CLIP provides a powerful foundation, achieving state-of-the-art performance in nuanced e-commerce tasks often requires significant adaptation to bridge the domain gap and address specific problem characteristics [1].

The integration of CLIP into e-commerce workflows has yielded promising results across various applications, demonstrating its potential to enhance both customer-facing interactions and backend operations (Table 1). However, its deployment is not without challenges.

Visual Search & Cross-Modal Retrieval is one of the most interesting applications.

CLIP fundamentally enhances search by enabling queries based on natural language descriptions that capture semantic meaning beyond keywords or by using images as queries. Studies report success in retrieving visually similar items, matching products across different seller domains with varying image styles [4], and performing text-to-image or image-to-text retrieval [1]. The use of CLIP embeddings within vector databases facilitates real-time similarity search at scale. Domain adaptation through fine-tuning, as specialized frameworks like CLIP-ITA [2] and EI-CLIP [6], significantly improves performance by addressing e-commerce specific nuances like fine-grained visual details or ambiguous entity names (e. g., brand names). Comparative studies confirm that multimodal CLIP-based approaches often outperform traditional single-modality systems [4].

Table 1

Custom CLIP frameworks comparison

Framework	Description
Contrastive language-image pretraining for category-to-image retrieval in e-commerce (CLIP-ITA) [2]	Extends CLIP by using separate encoders for category, image, title, and attributes, along with distinct projection heads, specifically for category-to-image retrieval in e-commerce
Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval (EI-CLIP) [6]	Employs causal inference principles (backdoor adjustment) and entity-aware modules (EA-learner, CE-selector) to mitigate the negative impact of confounding domain-specific entities during cross-modal retrieval
Personalized group-level preference alignment framework for diffusion models (PerFusion) [7]	Uses a CLIP-based reward model (PerFusionRM) within a larger framework involving personalized adaptive networks and group-level preference optimization for personalized text-to-image generation of fashion items
Visual zero-shot e-commerce product attribute value extraction (ViOC-AG) [8]	Aims for visual zero-shot attribute extraction using frozen CLIP image and text encoders connected to a trained text decoder, potentially enhanced by OCR and LLM outputs for correction
Multimodal attribute extraction for e-commerce (MXT) [9]	Frames attribute extraction as a multimodal question-answering task, using a T5-based generative model fused with CLIP visual features via a Multimodal Adaptation Gate

Product Recommendation is one of the most compelling applications of CLIP's shared embedding space. CLIP's shared embedding space naturally supports content-based recommendations by identifying products with similar visual or semantic features. More advanced applications include comparative recommendations. Personalized recommendations are also being explored, using CLIP embeddings potentially combined with user interaction data or preference models, as demonstrated by the PerFusion framework's use of a CLIP-based reward model. CLIP can also suggest visually similar substitutes for out-of-stock items [4]. Industrial deployments integrating CLIP-based personalization have reported substantial uplifts in key metrics like Click-through rate (CTR) and Conversion Rate (CVR).

Attribute Extraction represents a particularly fascinating domain. CLIP enables the extraction of product attributes directly from images or multimodal data (image + text), reducing reliance on structured seller inputs [9]. Zero-shot extraction is possible by formulating attributes as text prompts, while fine-tuning enhances accuracy for domain-specific attributes [10]. Visual-only approaches like ViOC-AG aim to extract attributes solely from images, potentially augmented by Optical Character Recognition (OCR) or Large Language Models (LLMs) [8]. Generative approaches like MXT frame attribute extraction as a multimodal question-answering problem [9]. These methods often demonstrate significant gains over text-only baselines, particularly when visual cues are critical [9]. A key challenge remains the accurate extraction of previously unseen attribute values in a zero-shot setting [11].

Content Generation and Product Design are among the most compelling applications. CLIP's ability to connect text and images makes it valuable for guiding generative models. Its embeddings can condition diffusion models to generate images from textual descriptions. In e-commerce, this enables applications like Alibaba's PerFusion system, which uses a CLIP-based reward model to personalize text-to-image generation for fashion items, facilitating innovative "sell it before you make it" business models where photorealistic product images are generated based on design descriptions before physical production [7].

Catalog Management (Categorization & Organization) represents a crucial application of CLIP technology. CLIP's zero-shot classification capability offers a powerful tool for automating product categorization [3]. By providing category names as text prompts, CLIP can assign categories to products based on their images or descriptions, significantly reducing manual effort and improving consistency [12]. This aids in building and indexing large catalogs, ensuring products are placed in the correct hierarchy for better navigation and discoverability. Accurate categorization driven by models like CLIP is fundamental for effective search, filtering, and overall user experience, ultimately impacting sales. CLIP-based systems can integrate with Product Information Management (PIM) systems to streamline catalog updates.

Quantitative results consistently demonstrate the benefits of applying CLIP, particularly adapted versions, to e-commerce tasks. For instance, EI-CLIP achieved relative R@1 improvements exceeding 10 % for cross-modal retrieval on Fashion-Gen by specifically addressing e-commerce entity semantics [6]. The PerFusion system reported over 13 % relative gains in both CTR and CVR in a large-scale industrial deployment for personalized product generation [7]. Frameworks like ViOC-AG [8] and MXT [9] also reported outperforming relevant baselines in attribute extraction.

Despite its successes, deploying CLIP in e-commerce involves navigating several limitations. The domain gap and adaptation requirements pose a significant challenge. The mismatch between CLIP's general pre-training data and specific e-commerce domains often necessitates fine-tuning or more complex adaptation strategies. This adaptation requires domain-specific data and computational resources, potentially negating some of the zero-shot advantages. The most impactful applications invariably involve adaptation.

CLIP can struggle with tasks requiring very fine-grained visual distinctions, such as differentiating between similar models of electronics or subtle variations in fashion items, or understanding abstract concepts. This limitation impacts its utility for detailed attribute extraction or nuanced product comparisons.

Pre-training CLIP is computationally intensive. While inference is generally faster, large-scale deployment for real-time search or recommendation across massive catalogs, or fine-tuning on large datasets, still requires significant computational infrastructure. This requirement potentially poses a barrier for smaller enterprises.

Table 2

Comparative performance highlights of CLIP-based methods

E-commerce Task	Method / framework	Dataset (s)	Key Metric	Reported Results
Cross-modal retrieval (fashion)	EI-CLIP	Fashion-Gen	R@1	+10.3 % (I->T), +10.5 % (T->I) vs. baseline [6]
Personalized product generation	Per Fusion	Alibaba (Prop.)	CTR, CVR	>+13 % relative improvement vs. human-designed [7]
Visual zero-shot attribute extraction	ViOC-AG	E-com (Prop.)	F1	Outperforms fine-tuned VLMs [8]
Multimodal attribute extraction	MXT	E-com (Prop.)	Recall	Outperforms CMA-CLIP & NER baseline at same precision [9]
Classification & retrieval	CLIP (Top-tuned)	6 E-com datasets	Accuracy, mAP	Competitive with full fine-tuning, more efficient [1]
Product matching (fashion)	Fine-tuned CLIP	Market-place (Prop.)	Precision	Outperforms single modality [4]
Category-to-image retrieval	CLIP-ITA	XMarket	R-precision	+265 % relative increase vs. CLIP-I (image only) [2]

CLIP's robustness to noisy web data is a strength, but its performance during fine-tuning or inference can still be affected by the quality and alignment of e-commerce-specific image-text data, which can be noisy, incomplete, or inconsistent.

Models trained on vast, uncurated internet data like CLIP can inherit societal biases present in the data. These biases could manifest in e-commerce as skewed recommendations, unfair representation of products associated with certain demographics, or biased attribute extraction. Auditing and mitigation strategies are essential but challenging.

Like many large deep learning models, CLIP's decision-making process lacks transparency. This "black box" nature makes it difficult to understand why a particular recommendation or classification was made, hindering debugging, trust-building, and ensuring responsible deployment (seeTable 2).

Evaluating the real-world effectiveness of CLIP-based systems, especially for subjective tasks like recommendation or generative design, beyond standard offline metrics remains challenging. Online A/B testing and user studies are crucial but resource-intensive.

V. CONCLUSION

The research reviewed indicated that while zero-shot CLIP provided a powerful baseline, realizing its full potential in the complex and specialized e-commerce environment typically required domain adaptation through fine-tuning or integration into custom-designed frameworks. These adapted approaches have showed substantial performance gains over traditional methods and baseline CLIP across various tasks and datasets. CLIP's utility was multifaceted, acting as a foundational component that can enhance operations across the e-commerce value chain, from improving customer-facing product discovery to optimizing backend catalog organization and even influencing product design itself.

However, significant challenges remain. The need for domain adaptation, limitations in fine-grained understanding, computational resource requirements, sensitivity to data quality, potential for inherited biases, and lack of interpretability are critical considerations for practical deployment. Future research should focus on developing more efficient and robust domain adaptation techniques tailored for e-commerce, improving CLIP's capacity for fine-grained analysis and reasoning about complex product relationships. Investigating and mitigating biases specific to e-commerce applications is paramount for responsible AI deployment. Furthermore, exploring the integration of CLIP with other data modalities, such as structured product specifications or user interaction logs, could unlock richer insights. Developing more comprehensive evaluation methodologies, particularly for user-centric tasks, and researching smaller, more computationally efficient CLIP-like architectures will also be crucial for broader adoption. Addressing these areas will further solidify CLIP's role as a transformative technology in the future of e-commerce.

References

[1] Czerwinska, U., Bircanoglu, C., & Chamoux, J. (2025). *Benchmarking Image Embeddings for E-Commerce: Evaluating Off-the Shelf Foundation Models, Fine-Tuning Strategies and Practical Trade-offs* [Preprint]. arXiv. DOI: https://doi.org/10.48550/arXiv.2504.07567

[2] Hendriksen, M., Bleeker, M., Vakulenko, S., Van Noord, N., Kuiper, E., & De Rijke, M. (2022, April). Extending CLIP for category-to-image retrieval in E-commerce. In *European Conference on Information Retrieval* (pp. 289–303). DOI: https://doi.org/10.48550/arXiv.2112.11294

[3] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). DOI: https://doi.org/10.48550/arXiv.2103.00020

[4] Tóth, S., Wilson, S., Tsoukara, A., Moreu, E., Masalovich, A., & Roemheld, L. (2024). End-to-end multi-modal product matching in fashion e-commerce. *arXiv preprint arXiv:2403.11593*. DOI: https://doi.org/10.48550/arXiv.2403.11593

[5] Ling, X., Peng, B., Du, H., Zhu, Z., & Ning, X. (2024). Captions Speak Louder than Images (CASLIE): Generalizing Foundation Models for E-commerce from High-quality Multimodal Instruction Data. *arXiv preprint arXiv:2410.17337*. DOI: https://doi.org/10.48550/arXiv.2410.17337

[6] Ma, H., Zhao, H., Lin, Z., Kale, A., Wang, Z., Yu, T., Gu, J., Choudhary, S., & Xie, X. (2022). EI-CLIP: Entity-Aware Interventional Contrastive Learning for E-Commerce Cross-Modal Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 18051–18061). DOI: https://doi.org/10.1109/CVPR52688.2022.01752

[7] Lin, J., Du, P., Liu, J., Li, W., Yu, Y., Zhang, W., & Cao, Y. (2025). *Sell It Before You Make It: Revolutionizing E-Commerce with Personalized AI-Generated Items* [Preprint]. arXiv. DOI: https://doi.org/10.48550/arXiv.2503.22182

[8] Gong, J., Cheng, M., Shen, H., Vandenbussche, P.-Y., Jenq, J., & Eldardiry, H. (2025). *Visual Zero-Shot E-Commerce Product Attribute Value Extraction* [Preprint]. arXiv. DOI: https://doi.org/10.48550/arXiv.2502.15979

[9] Khandelwal, A., Mittal, H., Kulkarni, S. S., & Gupta, D. (2023). Large Scale Generative Multimodal Attribute Extraction for E-commerce Attributes. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Vol. 5: Industry Track)* (pp. 305–312). DOI: https://doi.org/10.18653/v1/2023.acl-industry.29

[10] Jia, Q., Liu, Y., Xu, S., Liu, H., Wu, D., Fu, J., Vollgraf, R., & Wang, B. (2023). KG-FLIP: Knowledge-guided Fashion-domain Language-Image Pre-training for E-commerce. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)* (pp. 81–88). DOI: https://doi.org/10.18653/v1/2023.acl-industry.9

[11] Hu, J., Gong, J., Shen, H., & Eldardiry, H. (2025, April). Hypergraph-based Zero-shot Multi-modal Product Attribute Value Extraction. In *Proceedings of the ACM on Web Conference 2025* (pp. 4853–4862). DOI: https://doi.org/10.1145/3696410.3714714

[12] Cheng, Z., Zhang, W., Chou, C. C., Jau, Y. Y., Pathak, A., Gao, P., & Batur, U. (2024, November). E-commerce product categorization with LLM-based dual-expert classification paradigm. In *Proceedings of the 1st Workshop on Customizable NLP: (CustomNLP4U)* (pp. 294–304). DOI:https://doi.org/10.18653/v1/2024.customnlp4u-1.22

Oleksandr Khainas, PhD student at the Artificial Intelligence Department of Lviv Polytechnic National University. His research interests include multimodal learning, AI-powered recommendation systems, computer vision for e-commerce, and interpretable machine learning for business decision-making. Oleksandr’s practical expertise spans the implementation of large-scale ML pipelines, embedding-based image understanding, and integration of LLMs in enterprise environments.

Nataliia Melnykova, Doctor of Technical Sciences (Dr. Sc.), is an Associate Professor and the Head of the Artificial Intelligence Department at Lviv Polytechnic National University, Ukraine. Her research interests encompass artificial intelligence, machine learning, big data, data mining, decision support systems, and the application of these technologies in areas such as medical data processing, multimodal data handling (including speech and video analysis), and personalized medicine. Dr. Melnykova is a published author with numerous citations and has contributed to the organization of international scientific events.

Solomiia Fedushko, PhD, is a Senior Researcher at the Department of Information Management and Business Systems, Faculty of Management, Comenius University Bratislava, Slovakia. Her extensive research interests include Big Data analytics, artificial intelligence (AI), web mining, e-business, digital technologies in economics, social communication analysis, virtual communities, decision-making support in information technologies, and applied linguistics. Dr. Fedushko is a highly cited researcher with a substantial publication record, recognized internationally for her expertise and contributions to the field, including serving on editorial boards and organizing scientific events.