A Review on Multimodal Product Clustering and Vision Transformer Based Product Recommender System

Ssvr Kumar Addagarla[1], Upendra Kumar[2]

[1] Lincoln University College, Malaysia ; [2] Department of CSE, Institute of Engineering & Technology, Lucknow, India

Email ID : pdf.ssvrkumar@lincoln.edu.my, undera.ietlko@gmail.com

---

Abstract: In modern e-commerce, recommender systems serve as a crucial force in helping users navigate large catalogs to find the most relevant products. The majority of prevailing recommendation approaches, ranging from collaborative filtering to content-based and deep visual models, primarily capture coarse, category-level similarities while failing to recognize fine-grained product characteristics. Material, design patterns, intended usage, and compatibility in style are commonly missed, giving rise to visually similar recommendations that are not functionally or contextually complementary. Many state-of-the-art models also serve as black boxes, providing little insight into why a particular product is recommended. This review paper discusses the recent works on multimodal product clustering and vision transformer–based recommendation systems, focusing on how visual and textual modalities can be jointly leveraged for fine-grained product semantics capture. Major techniques, strengths, and weaknesses of the existing approaches are highlighted in the survey, with an emphasis on explainability and context awareness in recommendation frameworks. By consolidating current advances and open challenges, this review seeks to provide clear grounds for future research into fine-grained, interpretable multimodal product recommendation.

Keywords: Multimodal recommendation; Product clustering; Vision transformers; Explainable AI; E-commerce

---

Introduction

The burgeoning nature of the e-commerce industry has significantly altered the manner in which modern-day customers search for, compare, and purchase products. Currently, the global e-commerce market was estimated to be around USD 25.93 trillion in the year 2023, which is expected to rise to USD 83.26 trillion in the year 2030, thereby symbolizing the increasing reliance of modern-day customers on virtual platforms for their daily purchase needs. Due to the intensifying level of competition among various e-commerce platforms, the provision of authentic product information has emerged as an indispensable element that significantly impacts customer interaction and sales[1]-[4].

One of the most rapidly growing e-commerce environments is that of India. The Indian e-commerce market was estimated to be approximately 147.3 billion USD in 2024, and it is projected to expand at an impressive rate of 18.7% CAGR from 2024 to 2028 [5]. With an estimated 900 million internet users by

2025, online shopping has become an intrinsic part of people's lifestyles. Popular e-commerce sites such as Flipkart, Amazon.in, and Myntra report tens of millions of monthly active users, which gives an insight into its size and diversity that needs to be addressed by today's recommenders[6][7].

Despite such enormous reach, conversion rates on e-commerce platforms remain low, with only 2.4% of visitors actually ending up with a purchase. This difference presents a great opportunity to improve with better understanding and recommendations. Simple recommendations or traditional Recommender Systems may end up depending on gross levels of similarity or past engagement, thereby failing to identify minute differences in features or reasons why a certain item should be recommended, especially with a more crowded and similar-looking set of products[8]-[10].

In this respect, the recent surge in the use of multimodal learning techniques that combine visual and textual data has received considerable attention in the literature for its ability to learn the minute details of products. This is owing to the fact that Vision Transformer-based models present immense opportunities to learn the minute patterns in the visual data, whereas the application of multimodal clustering techniques could aid in segmenting the products in a manner beyond the category divisions. This review aims to discuss the recent trends in the application of the above two concepts.

Related work

1) Explainability in recommender systems

Explainability has moved from a "nice-to-have" to a necessity because recommendations influence purchases, trust, and long-term platform loyalty. Zhang & Chen (2020) systematically review why explanations matter (trust, transparency, debugging, fairness) and how they are typically delivered—through feature/aspect-level reasons, example-based explanations, counterfactual reasoning, or model-intrinsic interpretable designs. The survey highlights a key limitation of many explainable recommenders: explanations often remain generic (category-level) rather than fine-grained (material, pattern, style, intent), which is exactly the gap your review targets. Akhtar et al. (2024) further emphasize explainable AI in e-commerce as a practical need for reliability and user confidence, outlining explanation styles and their role in ethical/robust decision-making. Their discussion supports the motivation that "why recommended" is as important as "what recommended," especially when recommendations affect conversions and repeat purchases.

2) Multimodal learning foundations (how image + text should be fused)

Baltrušaitis, Ahuja, & Morency (2019) provide a strong multimodal machine learning taxonomy that frames multimodal systems around representation learning, fusion, alignment, and reasoning. This work is important because e-commerce recommendation is not just "adding image features to text"; it requires handling cross-modal alignment (e.g., text says "cotton," image shows polyester-like texture) and modality gaps (some items missing good images or rich descriptions).

3) Early product representation learning (embeddings + graphs for structure)

Baltescu et al. (2017) (MRNet-Product2Vec) represent an early direction where product embeddings are learned using sequence-like behavioral signals and multitask objectives, aiming to capture richer product semantics beyond sparse IDs. This line of work motivates why representation learning is central before recommendation/ranking. Xu & Ruan (2019) move toward product knowledge graph embeddings, showing that e-commerce signals can be structured into entities/relations and embedded to support relatedness and inference. This is a key stepping stone for explainability because graphs naturally support "reason paths" (attributes/relations) rather than opaque similarity scores.

4) Vision–language foundation models and CLIP-style alignment

Radford et al. (2021) introduce CLIP, demonstrating that contrastive pretraining aligns images and text into a shared embedding space, enabling strong transfer and retrieval behavior. For e-commerce, this directly supports the idea that fine-grained attributes can be better captured when visual semantics and textual semantics are trained to agree. Zhou et al. (2022) (CLIP4Rec) represent the direction of adapting CLIP-like vision–language foundations for recommendation settings, leveraging image–text contrastive alignment to improve recommendation quality, particularly when item content is crucial (e.g., new/cold items). (Public metadata around CLIP4Rec frequently describes it as CLIP-based contrastive learning for recommendation.

5) Multimodal recommendation models in practical pipelines

Jeong et al. (2024) propose a multimodal recommender combining review text and images using deep encoders and attention/co-attention style fusion. This supports the practical observation that reviews + images can capture complementary signals and improve recommendation quality beyond unimodal baselines. Oramas et al. (2024) discuss multimodal embeddings for recommendation and retrieval, showing how text embeddings can be enriched with multimodal content using contrastive learning, improving performance while keeping representations usable for retrieval-style recommendation. Ganhör et al. (2024) (SiBraR) directly addresses two real marketplace issues: cold-start and missing modality.

By using a single-branch, weight-sharing design, the model becomes robust even when one modality is absent, and it reduces the modality gap by mapping different modalities into a shared region of embedding space. MERLIN (2024) shows large-scale industry-focused multimodal/multilingual embedding for e-commerce recommendation via item associations, emphasizing scalability and robustness for real catalogs, including cold-start handling through metadata. Balachandran et al. (2025) (BiLens) argue that relying on only one modality is fragile in e-commerce (e.g., keyword stuffing in descriptions or visually ambiguous items).

Their framework integrates visual and textual information to improve recommendation relevance in realistic settings. Raj et al. (2025) further reinforce that multimodal recommenders benefit when user–item latent features are better uncovered/regularized, positioning multimodality as a way to strengthen representation quality and reduce sparsity effects.

The following Table 1 summarizes the various related findings.

Table 1 : Summary Table

Sno	Reference	Key Contributions	Limitation/Research Gap
1	Zhou, K. et al. (2022). "CLIP4Rec: A New Vision-Language Foundation Model for Recommendation." arXiv:2209.12356.	Adapted CLIP for e-commerce by aligning product image and text features; dataset: AliExpress & Amazon-Book subsets. Improved Recall@10 by 6–9%.	Limited to similarity-based retrieval; lacks explainability and fine-grained semantic clustering (e.g., style, use-case).
2	Jeong, E. et al. (2024). "A Multimodal Recommender System Using Deep Learning." Applied Sciences, 14(20), 9206.	Introduced dual-stream CNN + LSTM fusion model for fashion recommendation; dataset: DeepFashion2 (800K images). Achieved Precision@10 = 0.82.	No interpretability mechanism; only visual+text fusion; fails to provide reasoning or contextual complementarity between items.
3	Tsai, Y. et al. (2024). "XRec: Large Language Models for Explainable Recommendation." arXiv:2406.02377.	Integrated GPT-style LLM for natural language explanation generation on MovieLens and Yelp datasets; BLEU score 0.67 for human-like justification.	Domain limited (non-e-commerce); does not include multimodal embeddings or fine-grained clustering.
4	Baltrusaitis, T., Ahuja, C., & Morency, L. (2019). "Multimodal Machine Learning: A Survey and Taxonomy." IEEE TPAMI, 41(2), 423–443.	Defined taxonomy for multimodal fusion and co-learning; benchmarked across vision, audio, text datasets (AVEC, MM-IMDB).	Theoretical; lacks application to recommendation systems; no explainable framework for product-level multimodal fusion.
5	Zhang, Y., & Chen, X. (2020). "Explainable Recommendation: A Survey and New Perspectives." *Foundations and Trends in Information Retrieval, 14*(1), 1–101.	Surveyed 200+ explainable recommender systems; categorized explanation methods (model-intrinsic vs. post-hoc). Provided a taxonomy of evaluation metrics.	Lacks multimodal (image + text) integration and domain-specific fine-grained analysis; focuses mainly on text-based and collaborative models.
6	Wang, T. et al. (2024). "LLM-powered Product Knowledge Graph for Explainable E-commerce Recommendation." arXiv:2412.01837.	Built an LLM-generated Product Knowledge Graph to connect textual and visual attributes; dataset: JD.com 10M products; improved explainability BLEU +15%.	no integrated multimodal embedding or clustering mechanism.

7	Xu, D. & Ruan, L. (2019). Product Knowledge Graph Embedding for E-commerce. (arXiv preprint)	Proposed a product knowledge graph embedding framework to encode product relations; dataset: E-commerce product graph; results: improved downstream recommendation tasks.	Focuses on graph structure embedding; lacks integration of visual + text multimodal fusion and LLM-based explanation for recommendations.
8	Radford, A. et al. (2021). "Learning Transferable Visual Models from Natural Language Supervision (CLIP)." ICML.	Pretrained on 400M image–text pairs (OpenAI dataset); achieved SOTA zero-shot classification (ImageNet 76.2% top-1).	Generic vision-language model; not tuned for fine-grained product-level attributes or e-commerce domain clustering.
9	Spillo, G. et al. (2025). "Knowledge-aware Recommendations Fusing Heterogeneous Multimodal Item Embeddings." Journal of Intelligent Information Systems.	Combined graph, text, and image encoders with cross-attention for multi-domain recommendation; dataset: MovieLens + Amazon reviews. Outperformed baseline by 11% in NDCG@10.	Fusion method is complex but non-explainable; no natural-language interpretability or fine-grained context modeling.
10	Oramas, S. et al. (2024). "Multimodal Embeddings for Recommendation and Retrieval." CEUR-WS Vol. 3787.	Proposed contrastive learning-based multimodal embedding; datasets: Amazon (Fashion), Pinterest, achieving Precision@10 = 0.80.	Focused on similarity; no interpretability or complementary product pairing (lacks LLM-based explanation).

Key Contributions

The main contributions of the reviewed work can be summarized as below:

- The reviewed works collectively determine a clear shift from traditional similarity-based recommendation toward multimodal representation learning that integrates images, text, and structured knowledge. Various advanced techniques such as product embeddings, vision–language models, and multimodal fusion enable systems to capture subtle product attributes (e.g., style, material, usage intent) and support more meaningful product clustering beyond broad category labels.
- LLM based Vision Transformers, CLIP-style vision–language models, and multimodal embedding frameworks have significantly improved robustness in challenging scenarios such as cold-start, missing modalities, and large-scale catalogs. The integration of product knowledge graphs and heterogeneous embeddings further enhances contextual reasoning, enabling recommendations that reflect functional relationships rather than surface-level similarity.
- Recent studies highlight a strong drive toward explainable recommendation, leveraging aspect-level reasoning, structured knowledge graphs, and Large Language Models to generate human-understandable explanations. These approaches address the black-box nature of earlier systems

and improve transparency, trust, and user confidence—key factors for real-world e-commerce adoption and conversion improvement.

Conclusions

- This review addressed the growing limitations of existing e-commerce recommender systems that primarily rely on coarse category-level similarity or historical interaction data. Such systems often fail to capture fine-grained product attributes, functional complementarity, and contextual relevance, while also operating as black boxes without meaningful explanations. The motivation behind this survey was to examine how recent advances in multimodal learning, vision transformers, knowledge-aware modeling, and large language models can collectively improve product understanding, recommendation accuracy, and transparency in large-scale e-commerce environments.

- The survey systematically reviewed representative works that employ multimodal product representations combining visual, textual, and structured knowledge signals. The reviewed methods include product embedding learning, multimodal fusion frameworks, vision transformer and vision–language foundation models, product knowledge graph–based approaches, and LLM-powered explainable recommendation frameworks. Together, these methods demonstrate how multimodal clustering and recommendation pipelines are evolving toward richer semantic modeling and interpretable decision-making.

- The analysis reveals that multimodal approaches consistently outperform unimodal systems by capturing fine-grained product characteristics and reducing ambiguity in dense product catalogs. Vision transformer–based and CLIP-style models improve visual-semantic alignment, while knowledge-aware and LLM-based methods enable explainable recommendations that are easier for users to understand and trust. Additionally, recent frameworks show improved robustness in cold-start and missing-modality scenarios, making them more suitable for real-world deployment.

- Despite these advancements, several limitations remain. Many multimodal and LLM-based models are computationally expensive and difficult to deploy at scale. Explainability mechanisms are still often post-hoc and may not fully reflect the true decision process of the model. Moreover, standardized benchmarks for fine-grained multimodal explainability are limited. Future research should focus on lightweight and scalable architectures, intrinsic explainability, better integration of domain knowledge, and user-centric evaluation metrics that assess not only accuracy but also trust, transparency, and business impact.

References

1. Zhang, Y., & Chen, X., 2020. “Explainable Recommendation: A Survey and New Perspectives.” Foundations and Trends in Information Retrieval, 14(1), 1–101.

2. Baltescu, P. et al., 2017. “MRNet-Product2Vec: A Multi-task Recurrent Neural Network for Product Embeddings.” ECML/PKDD.
3. Zhou, K. et al., 2022. “CLIP4Rec: A New Vision-Language Foundation Model for Recommendation.” arXiv:2209.12356.
4. Jeong, E. et al., 2024. “A Multimodal Recommender System Using Deep Learning.” Applied Sciences, 14(20), 9206.
5. Tsai, Y. et al., 2024. “XRec: Large Language Models for Explainable Recommendation.” arXiv:2406.02377.
6. Baltrusaitis, T., Ahuja, C., & Morency, L. (2019). “Multimodal Machine Learning: A Survey and Taxonomy.” IEEE TPAMI, 41(2), 423–443.
7. Xu, D. & Ruan, L., 2019. Product Knowledge Graph Embedding for E-commerce. (arXiv preprint)
8. Radford, A. et al., 2021. “Learning Transferable Visual Models from Natural Language Supervision (CLIP).” ICML.
9. Spillo, G. et al., 2025. “Knowledge-aware Recommendations Fusing Heterogeneous Multimodal Item Embeddings.” Journal of Intelligent Information Systems.
10. Oramas, S. et al., 2024. “Multimodal Embeddings for Recommendation and Retrieval.” CEUR-WS Vol. 3787.
11. Ganhör, S. et al. 2024. “A Multimodal Single-Branch Embedding Network for Recommendation in Cold-Start and Missing Modality Scenarios.” arXiv:2409.17864.
12. Wang, T. et al. 2024. “LLM-powered Product Knowledge Graph for Explainable E-commerce Recommendation.” arXiv:2412.01837.
13. MERLIN Team, 2024. “MERLIN: Multimodal & Multilingual Embedding for E-commerce Recommendation.” ACM RecSys 2024.
14. Balachandran, S. et al. 2025. “BiLens: A Multimodal Framework for Enhancing E-commerce Recommendation.” Information Fusion (SCIE).
15. Raj, A. et al. 2025. “Unlocking Latent Features of Users and Items in Multimodal Recommendation.” Scientific Reports (Nature).
16. Liao, Y. et al. 2025. “Aspect-Enhanced Explainable Recommendation with Multimodal Data.” ACM TIST (accepted).
17. Park, H. & Oh, Y. 2024. “Enhancing E-commerce Recommendation with Multiple-Item Purchase Data: A Transformer-based Approach.” Elsevier Information Sciences (in press).
18. Akhtar, R. et al. 2024. “Decoding the Recommender System: A Comprehensive Guide to Explainable AI in E-commerce.” Expert Systems with Applications (Elsevier).
19. Zhang, L. et al. 2024. “Multimodal Prediction of E-commerce Customer Satisfaction Driven by Big Data.” Applied Soft Computing (Elsevier).
20. Chinnasamy, P. 2025. “Current Trends in Intelligent Recommendation Systems in E-commerce.” IJCESEN 2025