# COMET: Compatibility-Oriented Multi-modal Embedding Transformer for Visual Recommendations

Dween Rabius Sanny
Amazon
Bengaluru, India
drsanny@amazon.com

Prateek Sircar
Amazon
Gurugram, India
sircarp@amazon.com

Deepak Gupta
Amazon
Gurugram, India
dgupt@amazon.com

## Abstract

Recommending visually compatible products in fashion and interior design is a significant challenge, as compatibility rules are nuanced, context-dependent, and reliant on fine-grained details that traditional models fail to capture. Existing methods often struggle with heterogeneous compatibility rules (e.g., sofa-table vs. sofa-curtain) and an over-reliance on global visual features, missing critical textual cues like style or material. To address these limitations, we introduce COMET (Compatibility-Oriented Multi-modal Embedding Transformer), a scalable, vision-language framework for visual recommendations. COMET replaces rigid, category-specific subspaces with an attribute-conditioned cross-attention mechanism, reframing compatibility as a conditional retrieval problem. By formulating a textual compatibility prompt that encodes relational context and structured attributes, COMET dynamically conditions which visual features are attended to, producing a joint, context-aware representation. This multi-modal approach allows the model to leverage descriptive text (e.g., "mid-century modern" or "oak finish") to understand visually ambiguous stylistic nuances. The model is trained using a triplet loss with hard negative sample strategy to effectively distinguish between compatible and incompatible item pairs. COMET was evaluated on established benchmarks for both fashion (Polyvore) and furniture (Bonn Furniture), in addition to our in-house datasets.

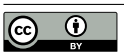
## CCS Concepts

• **Information systems → Top-k retrieval in databases**.

## Keywords

visual compatibility, multi-modal learning, fashion recommendation,

**ACM Reference Format:**
Dween Rabius Sanny, Prateek Sircar, and Deepak Gupta. 2026. COMET: Compatibility-Oriented Multi-modal Embedding Transformer for Visual Recommendations. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3805712.3808473




*SIGIR '26, Melbourne, VIC, Australia*

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3808473

## 1 Introduction

In visually-driven e-commerce, a key challenge is moving beyond single-item recommendations to suggest cohesive, stylistically harmonious ensembles. This capability is governed by Fill-in-the-Blank (FITB) and Complementary Item Retrieval (CIR) tasks that are critical for enhancing user experience in domains like fashion and interior design.

Despite significant progress, existing methods encounter fundamental scalability bottlenecks. Early approaches [7, 13] pioneered the use of category-specific subspaces to address non-transitive compatibility rules. However, this architectural design mandates a dedicated subspace for every possible category pairing such as "shirt-pant", "sofa-table". For a dataset containing $c$ categories, the number of required subspaces scales quadratically as $\frac{c(c-1)}{2}$. This combinatorial explosion leads to rigid, non-scalable frameworks that practically unfeasible for diverse, large-scale catalogs in real-world applications.

More recent methods, including powerful Transformer-based models like [10], and [4], have advanced retrieval performance. Yet, these models are often *primarily visual*, meaning they do not fully leverage the conceptual, textual attributes that define true stylistic coherence. For example, they may not easily distinguish between visually similar mid-century modern" and Scandinavian" chairs, as these nuances are locked in textual data. This same reliance on visual features is a constraint for fashion-specific models. While [1] pioneered style compatibility and [9, 12] advanced the field using Graph Neural Networks (GNNs), both approaches remain limited by their inability to deeply incorporate the fine-grained, non-visual descriptors (e.g., specific wood finishes, historical periods) essential for high-fidelity interior design recommendations.

This "multi-modal gap" is fundamentally a data bottleneck. E-commerce catalogs are notoriously noisy, with crucial attributes often missing, unstructured, or inconsistent.

To address these challenges, we propose a novel two-stage framework. First, we resolve the data bottleneck by fine-tuning Qwen 2.5-VL for a Question-Answer Generation (QAG) task, extracting clean, structured attributes from noisy catalog metadata. Second, we introduce COMET, which reframes compatibility as a flexible, query-driven problem replacing the rigid subspaces. By fusing these high-quality textual attributes with visual features via cross-attention, COMET captures nuanced stylistic cues. Trained via a triplet loss, our approach achieves new state-of-the-art performance on both Polyvore and Bonn Furniture benchmarks, confirming that solving the underlying attribute data problem is a critical prerequisite for robust, generalizable compatibility modeling.

## 2 Related Work

Research in visual compatibility has evolved from sequence models treating outfits as ordered sets using LSTMs [3] and Graph Convolutional Networks [2], to learning subspace embeddings that address heterogeneous compatibility rules. Recognizing that compatibility is non-transitive (e.g., shirt-pants rules differ from pants-shoes), [13] proposed category-specific subspaces, later extended by SCE-Net [11] and CSA-Net [7]. While effective, these methods require $\frac{c(c-1)}{2}$ dedicated subspaces for $c$ categories, making them impractical for large-scale catalogs.

More recently, Transformer-based approaches such as Outfit Transformer [10] and HAT [4] advanced retrieval performance but remain primarily visual, missing fine-grained textual attributes (e.g., "mid-century modern" vs. "Scandinavian") essential for true stylistic coherence. This limitation extends to furniture-specific models [1, 9], which similarly lack deep multi-modal integration.

COMET addresses both gaps: we resolve the noisy catalog *data bottleneck* via a generative VLM pipeline for clean attribute extraction, and replace the rigid subspace architectures of [13] and [7] with a flexible, query-driven cross-attention framework that fuses high-quality textual attributes with visual features.

## 3 Data Collection

To overcome the costly manual annotation bottleneck, we developed a unified, weakly supervised pipeline using item co-occurrence as a compatibility proxy (Figure 1b). We harvest lifestyle images (e.g., outfit photos, interior scenes) and apply GroundingDINO [8] to extract co-occurring items, which are mapped to our catalog via SkiLL [15], a fine-grained visual similarity model. This pipeline yielded a 2M-pair fashion dataset and an expert-validated (97% compatible) furniture dataset, providing a scalable foundation for training without manual annotation.

## 4 Proposed Method

Our proposed method, COMET, learns a nuanced, multi-modal representation of product compatibility through two stages (Figure 1a):
**Multimodal Attribute Extraction:** A generative pipeline using a VLM to extract clean, structured attributes (e.g., style, material) from noisy catalog data.
**COMET Compatibility Model:** A flexible, query-driven architecture replacing rigid category-specific meta-spaces, trained via triplet loss to pull compatible items closer.

### 4.1 Multimodal Attribute Extraction

To address the "multi-modal gap," we curated a large-scale dataset of 8 million samples from our internal catalogs. We formulated attribute extraction as a source-grounded generative task [5], employing two specialized surrogate reasoning models (DeepSeek-R1-Distill-Qwen-7B)—a vision-only and a text-only reasoning system—to independently verify the provenance of each ground-truth attribute. This allows us to identify whether a value is inferable from the image, the unstructured text context, both, or is absent. Each training instance is formatted as a structured XML-style triplet: (1) `<is_present>`, (2) (*text, image, both*, or *absent*), and (3) `attribute:value`. For example, given a sofa image with the context *"A luxurious velvet finish in mid-century style"*, the model generates:

*yes*
*both_text_and_image*
*material: velvet; style: mid-century modern*

This explicit grounding forces the model to "learn to refuse" when information is missing (outputting *no*), effectively mitigating hallucinations caused by visual or linguistic biases.

We fine-tune Qwen 2.5-VL using Low-Rank Adaptation (LoRA), updating only 6% of the total parameters to ensure a low memory footprint. The model is optimized using a causal language modeling objective, where cross-entropy loss is computed exclusively on the structured response tokens. This objective encourages the model to synthesize information only from verified sources, providing the clean signal required for downstream compatibility modeling.

### 4.2 COMET: Compatibility Model

The COMET model (Figure 1a) replaces fixed, pairwise meta-spaces with a flexible, query-driven multi-modal encoder consisting of three components.

*4.2.1 Visual Encoder.* A standard Vision Transformer (ViT) serves as the visual backbone. It processes an input item's image by dividing it into patches, which are linearly projected into patch embeddings. This encoder outputs a sequence of visual features, which serve as the Key (K) and Value (V) for the fusion module.

*4.2.2 Text Encoder and Query Formulation.* The textual component serves as the compatibility query, replacing the $\frac{c(c-1)}{2}$ subspaces. It is formulated from: 1) the **Compatibility Context** (e.g., *"current category sofa, searching for coffee table"*) and 2) the **Multi-modal Attributes** from our extraction pipeline (e.g., *"style: bohemian"*, *"material: oak"*). A Text Encoder produces the query vector (Q), which attends to visual embeddings (K, V) via cross-attention. This allows the query to selectively focus on the most relevant visual features for each compatibility context, producing a single joint embedding vector $e$. We term this *query-driven* because the textual prompt actively conditions which visual features are attended to, functioning as an attribute-conditioned query that dynamically selects relevant visual evidence.

### 4.3 Training Objective

To organize the embedding space, we employ a triplet loss. This loss ensures that compatible items are pulled closer together in a subspace specified by the compatibility context prompt while incompatible items are pushed apart. Given an anchor item $a$ (e.g., a sofa), a compatible positive item $p$ (a matching table), and an incompatible negative item $n$ (a clashing table), their respective joint embeddings $(e_a, e_p, e_n)$ are computed using the COMET.

The triplet loss is then defined as:

$$\mathcal{L}(a, p, n) = \max(0, d(e_a, e_p) - d(e_a, e_n) + m)$$

where $d$ is a distance function (e.g., Euclidean distance) and $m$ is a predefined margin. Positive pairs $(a, p)$ are sourced from ground-truth sets in datasets like Polyvore and Bonn. For hard negative mining, we employ a more structured approach. We first perform an agglomerative clustering of items on a category-wise basis, using embeddings from the SkiLL visual similarity model [15]. This creates a hierarchical tree of visually similar items. For a given positive item, its cluster is identified, and the hard negative

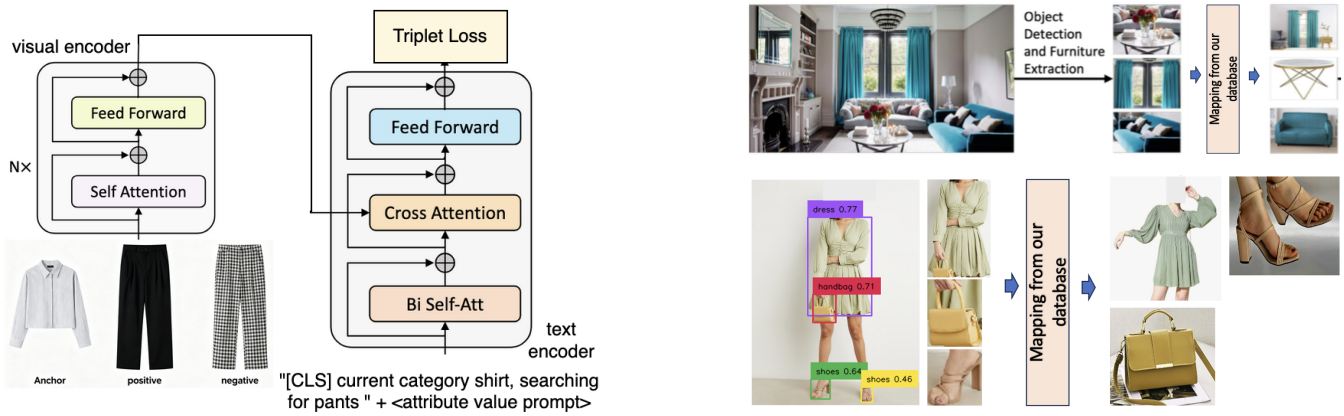



**Figure 1: a) Architecture of our approach, b) Data collection method for furniture and fashion**

**Table 1: Comparison of our model with state-of-the-art methods on the Polyvore FITB ( accuracy) and CIR tasks (recall@top-k).**

| Methods | Polyvore disjoint | | | | Polyvore nondisjoint | | | |
|---|---|---|---|---|---|---|---|---|
| | FITB | R@10 | R@30 | R@50 | FITB | R@10 | R@30 | R@50 |
| Type-Aware | 55.65 | 3.66 | 8.26 | 11.98 | 57.83 | 3.50 | 8.56 | 12.66 |
| SCE-Net Averag | 53.67 | 4.41 | 9.85 | 13.87 | 59.07 | 5.10 | 11.20 | 15.93 |
| CSA-Net | 59.26 | 5.93 | 12.31 | 17.85 | 63.73 | 8.27 | 15.67 | 20.91 |
| Outfit transformer (Vision only) | - | 6.03 | 12.20 | 16.51 | 58.92 | 9.29 | 16.94 | 21.82 |
| Outfit transformer | 59.48 | 6.53 | 12.12 | 16.64 | 67.10 | 9.58 | 17.96 | 21.98 |
| HAT | 57.32 | 5.13 | 10.04 | 15.29 | 64.87 | 7.46 | 15.74 | 20.38 |
| **COMET (Ours)** | **62.24** | **7.25** | **15.04** | **20.58** | **68.42** | **10.14** | **19.38** | **25.86** |

sample $n$ is then selected from an adjacent cluster. This strategy ensures that negatives are visually similar (e.g., same category) but stylistically distinct, forcing the model to learn fine-grained, multi-modal rules.

## 5 Results and Evaluations

We conducted rigorous experiments over various tasks and metrics to evaluate the proposed COMET framework. We compare against several state-of-the-art approaches, including Type-aware [13], SCE-Net [11], CSA-Net [7], Outfit Transformer [10], and HAT [4]. We evaluate on three axes: (1) **Attribute Extraction Quality**, measured by exact-match accuracy against mPLUG [6] and MoE-MoE [14] (Table 2a); (2) **Fill-in-the-Blank (FITB)**, evaluated by accuracy (Tables 1, 2b); and (3) **Complementary Item Retrieval (CIR)**, evaluated by Recall@K (Tables 1, 2c).

### 5.1 Fill-in-the-Blank (FITB)

We evaluated the FITB task on both the Polyvore [3] and Bonn [1] datasets. For the Polyvore datasets, we trained our model on the Polyvore disjoint and nondisjoint training sets and utilized the respective test sets to evaluate the metrics. The task is to select the most compatible candidate item given a partial outfit, evaluated by overall accuracy. The results are shown in Table 1. We observe that for both Polyvore disjoint and non-disjoint sets, our proposed COMET approach shows consistently strong results. COMET achieves an accuracy of 62.24% on the disjoint set and 68.42% on the nondisjoint set, outperforming all multi-modal baselines including HAT and Outfit Transformer.

Additionally, we followed the FITB task methodology for the Bonn Furniture dataset as described by [9]. This task consists of choosing, among a set of four possible choices, the item that best completes a furniture set. The key challenge in this benchmark is that the incorrect answers are randomly selected items from the same category, but a *different style*, from the correct answer. The task is addressed by forming all possible sets between the partial set and the item choices, running the sets through the model, and selecting the item that produces the set with the highest compatibility score.

The results for this task are presented in Table 2b. Our multi-modal COMET framework achieves an accuracy of 82.4%, establishing a new state-of-the-art. This significantly surpasses both the original Siamese Network baseline (75.0%) and the more recent GNN-based models (78.5%), demonstrating the superiority of our multi-modal approach.

### 5.2 Complementary Item Retrieval (CIR)

We trained our model on the Polyvore disjoint and nondisjoint train sets and performed the CIR task on the respective test sets. For this task, we used recall@top-k (abbreviated as R@k) as the metric. For the calculation of R@k, we adopted the same methodology described in CSA-Net to evaluate the Polyvore disjoint and nondisjoint sets. We indexed our dataset by category, treating each category as the target category. For indexing, we used the same methodology described in CSA-Net. The results for this task are

**Table 2: (a) Attribute extraction exact-match relative (w.r.t. Raw Catalog) accuracy on 5,000 annotated samples; (b) FITB Accuracy on Bonn Furniture; (c) Retrieval performance on Softlines and OHL datasets (relative Recall@K w.r.t. Type-aware).**

**(a) Attribute Extraction**

| Method | Style | Mat. |
|---|---|---|
| Raw Catalog | 0 | 0 |
| mPLUG | 0.113 | 0.118 |
| MoEMoE | 0.119 | 0.133 |
| **Qwen 2.5-VL** | **0.141** | **0.163** |

**(b) Bonn FITB Acc.**

| Method | Acc. ↑ |
|---|---|
| Fine-Tuned CNN | 0.571 |
| Siamese Net. | 0.750 |
| GNN I | 0.772 |
| GNN II (GAT) | 0.785 |
| **COMET** | **0.824** |

**(c) In-house Retrieval (rel. R@K w.r.t. Type-aware)**

| | Softlines | | | OHL | | |
|---|---|---|---|---|---|---|
| Method | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| Type-aware | 0 | 0 | 0 | 0 | 0 | 0 |
| CSA-Net | 0.26 | 1.22 | 2.14 | 0.31 | 1.35 | 2.20 |
| Out. Trans. | 0.58 | 1.99 | 3.98 | 0.62 | 2.10 | 4.15 |
| HAT | 0.52 | 1.36 | 3.53 | 0.55 | 1.40 | 3.60 |
| **COMET** | **4.86** | **12.0** | **14.6** | **5.12** | **13.2** | **15.8** |

**Figure 2: Qualitative results for CIR task for Polyvore**

**Table 3: Strategic Ablation Analysis: Isolating Data-Driven vs. Model-Driven Performance Gains on Polyvore (Disjoint/Non-disjoint) and Bonn FITB acc.**

| Method | Poly. (D) | Poly. (ND) | Bonn FITB |
|---|---|---|---|
| *Data-Driven: Modality & Quality* | | | |
| 1. Vision Only | 56.70% | 60.57% | 75.82% |
| 2. Only Text Context | 61.95% | 67.10% | 76.25% |
| 3. Text Context+title | 62.00% | 67.35% | 78.10% |
| *Model-Driven: Arch. & Strategy* | | | |
| 4. w/o Cross-Attn. | 61.88% | 67.50% | 79.15% |
| 5. w/o Neg. Sampling | 60.45% | 65.20% | 77.50% |
| **6. COMET (Full)** | **62.24%** | **68.42%** | **82.40%** |

presented in Table 1. We observe that our model outperforms all recent baselines in almost all R@k metrics for both vision-only and joint vision-text models. On the Polyvore nondisjoint set, COMET achieves an R@10 of 10.14%, an R@30 of 19.38%, and an R@50 of 25.86%. This represents a clear improvement over the strongest multi-modal baselines, Outfit Transformer (9.58% R@10) and HAT (7.46% R@10), validating the superior retrieval capabilities of our query-driven, multi-modal fusion.

To further validate the generalizability of our approach, we conducted a relative R@k analysis, benchmarking all models against the baseline [13]. The results are shown in Table 2. This comparison demonstrates that COMET's architecture generalizes robustly across different e-commerce verticals. While prior multi-modal methods like Outfit Transformer and HAT show modest gains over the baseline (e.g., 3.98 R@10 for Outfit Transformer), our COMET model achieves a massive 14.56 R@10 on Softlines (Fashion). Crucially, this strong performance is replicated on the OHL (Furniture) dataset. Here, COMET achieves a relative R@1 of 5.12 and R@10 of 15.8, which is more than 3.8x the R@10 performance of the next-best baseline (Outfit Transformer at 4.15). This consistent, state-of-the-art performance across both domains validates that our multi-modal, query-driven approach is a more general and effective solution for compatibility modeling. We have launched this experience in our e-commerce plateform and it achieved **+0.06% lift in Same-Day Delivery Ordered Product Sales** and **cart conversion of 7.3%** compared to the current widget's 1.55% in a four week A/B testing phase.

## 6 Ablation Study

We evaluate the impact of multi-modal signals, attribute quality, and training strategies on FITB accuracy across the Polyvore (Disjoint/Non-disjoint) and Bonn Furniture datasets. Our ablation study (Table 3) confirms that COMET's success stems from both Data-Driven Quality and Model-Driven Architecture. Replacing vision-only features with structured VLM attributes yields +6.58% on Bonn Furniture (rows 1 vs.6), while our extracted attributes outperform raw titles (row 3 vs. 6), confirming the value of clean attribute extraction. On the architecture side, removing cross-attention (row 4) or cluster-based hard negative mining (row 5) causes sharp drops across all domains (e.g., −3.22% on Polyvore Non-disjoint), validating that COMET's fusion and training strategies are essential for maximizing compatibility performance.

## 7 Conclusion

We introduced COMET, a two-stage framework that resolves the persistent data and modeling bottlenecks in visual compatibility. By leveraging Qwen 2.5-VL for clean attribute extraction and a query-driven cross-attention architecture with cluster-based hard negative mining, COMET replaces rigid, pair-specific subspaces of prior work. Evaluations on Polyvore, Bonn, and in-house datasets demonstrate SOTA performance in both FITB and CIR, with real-world deployment achieving **+0.06% OPS lift** and **7.3% cart conversion**. This validates that resolving the attribute "multi-modal gap" is a critical prerequisite for robust compatibility systems.

## References

[1] Divyansh Aggarwal, Elchin Valiyev, Fadime Sener, and Angela Yao. 2018. Learning style compatibility for furniture. In *German Conference on Pattern Recognition*. Springer, 552–566.

[2] Guillem Cucurull, Perouz Taslakian, and David Vazquez. 2019. Context-aware visual compatibility prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10373–10382.

[3] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. 2017. Learning fashion compatibility with bidirectional LSTMs. In *Proceedings of the 25th ACM International Conference on Multimedia*. 1078–1086.

[4] Myong Chol Jung, Julien Monteil, Philip Schulz, and Volodymyr Vaskovych. 2025. Personalised outfit recommendation via history-aware transformers. In *Proceedings of the 18th ACM International Conference on Web Search and Data Mining*. ACM, 633–641.

[5] Anant Khandelwal, Happy Mittal, Shreyas Kulkarni, and Deepak Gupta. 2023. Large Scale Generative Multimodal Attribute Extraction for E-commerce Attributes. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, Sunayana Sitaram, Beata Beigman Klebanov, and Jason D Williams (Eds.). Association for Computational Linguistics, Toronto, Canada, 305–312. doi:10.18653/v1/2023.acl-industry.29

[6] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. 2022. mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 7241–7259.

[7] Yen-Liang Lin, Son Tran, and Larry S. Davis. 2020. Fashion outfit complementary item retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3311–3319.

[8] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. *arXiv preprint arXiv:2303.05499* (2023).

[9] Luisa F Polania and Shashank Gupte. 2020. Learning furniture compatibility with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 1022–1023.

[10] Rohan Sarkar, Navaneeth Bodla, Mariya Vasileva, Yen-Liang Lin, Anurag Beniwal, Alan Lu, and Gerard Medioni. 2022. OutfitTransformer: Outfit representations for fashion recommendation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2263–2267.

[11] Reuben Tan, Mariya I Vasileva, Kate Saenko, and Bryan A Plummer. 2019. Learning similarity conditions without explicit supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10373–10382.

[12] Sambeet Tiady, Arihant Jain, Dween Rabius Sanny, Khushi Gupta, Srinivas Virinchi, Swapnil Gupta, Anoop Saladi, and Deepak Gupta. 2024. MERLIN: Multimodal & Multilingual Embedding for Recommendations at Large-scale via Item Associations. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (Boise, ID, USA) *(CIKM '24)*. Association for Computing Machinery, New York, NY, USA, 4914–4921. doi:10.1145/3627673.3680106

[13] Mariya I. Vasileva, Bryan A. Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. 2018. Learning type-aware embeddings for fashion compatibility. In *Proceedings of the European Conference on Computer Vision*. 390–405.

[14] Vinay Kumar Verma, Shreyas Sunil Kulkarni, Happy Mittal, and Deepak Gupta. 2025. MoEMoE: Question Guided Dense and Scalable Sparse Mixture-of-Expert for Multi-source Multi-modal Answering. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 62–69.

[15] Vinay K Verma, Dween Rabius Sanny, Shreyas Sunil Kulkarni, Prateek Sircar, Abhishek Singh, and Deepak Gupta. 2023. SkiLL: Skipping Color and Label Landscape: Self Supervised Design Representations for Products in E-commerce. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3502–3506.

## Author Biography

**Dween Rabius Sanny** is an Applied Scientist at Central Machine Learning, Amazon, Bengaluru, India. His research focuses on multimodal representation learning, visual compatibility modeling, and scalable recommendation systems for e-commerce. He has published at top-tier venues including IEEE/CVF CVPR and ACM SIGIR, with work spanning vision-language models, self-supervised learning, and industrial-scale ML deployment.