Pekka: MLLM-based Product Representation Learning with User Engagement Signals for E-Commerce

Rui Yan[1], Kevin Huang[1], Miao Li[1], Zhenyuan Liu[1], Mina Ghashami[1], Michael Schlichtkrull[1], Miguel Arduengo[1], Wenliang Gao[1], Bella Shi[1], Michael Du[1], Fan Xia[1], Zhaoheng Zheng[1], I-Ta Lee[1], Minghai Chen[1], Yash Desai[1], Rana Alkhoury Maroun[1], Yugandhar Nanda[1], Weijian Han[1], Dongxin Liu[1], Santiago Villa Fernandez[1], Jack Wang[1], Guang Yang[1], Amit Jaspal[1], Saurabh Gupta[1], Moji Solgi[1], Ben Schulte[1], Deepak Chandra[1] and Damien Lefortier[1]

[1]*Meta Platforms, Inc.*

Abstract

Product embeddings underpin modern e-commerce systems, enabling search, recommendation and product discovery across vast and rapidly changing catalogs. Effective representations should capture signals from multimodal product content as well as user preferences expressed through shopping interactions. Existing approaches, however, largely optimize content similarity and often fail to model the behavioral relationships that drive real-world product relevance. Here we present Pekka, a generalizable product embedding framework that adapts pretrained multimodal large language models to e-commerce representation learning. Pekka combines the cross-modal understanding and broad world knowledge of multimodal foundation models with large-scale supervision from user engagement signals, yielding embeddings that bridge semantic understanding and shopping intent. To learn from noisy browsing logs, we develop a scalable two-stage curation pipeline that transforms raw session interactions into high-quality co-clicked product pairs for contrastive learning. Trained on these signals, Pekka substantially improves behavioral retrieval while transferring effectively across domains and tasks, including search ranking and cross-modal retrieval on unseen platforms. Across internal benchmarks, Pekka improves Recall@5 by 30% relative to a strong CLIP-based production baseline. On public benchmarks spanning behavioral retrieval, cross-modal retrieval, text-only search ranking and multimodal search ranking, Pekka achieves 42.3 nDCG@10 versus 24.5 for CLIP on ESCI, and attains the highest content retrieval quality on M5Product (89.9 R@1, surpassing both SigLIP2 and E5-V). These results demonstrate that Pekka's approach of grounding multimodal representations in user engagement signals yields scalable product embeddings that unify semantic understanding with behavioral relevance.

Keywords

Product Embeddings, Multimodal LLMs, E-commerce Search, Representation Learning

1. Introduction

Product embeddings have become a foundational component of modern e-commerce systems, representing products as dense vectors that power retrieval, recommendation and discovery [1, 2]. As online marketplaces expand in scale and modality, the quality of these representations increasingly shapes retrieval relevance, user experience and downstream model performance. Effective embeddings should capture not only multimodal product content, including images, titles and descriptions, but also the patterns of user preference revealed through shopping behavior. Products that differ visually may nonetheless satisfy the same shopping need, whereas embeddings optimized only for surface similarity can yield redundant rankings and limited discovery.

Early product representation methods relied primarily on collaborative signals derived from user interactions. Approaches such as Item2Vec and graph-based methods including PinSage and LightGCN learned item similarity from co-view, co-click, co-purchase and user–item interaction graphs [3, 4, 5, 6]. These methods effectively capture behavioral relevance, but typically operate over item identities or interaction structure rather than the rich multimodal signals present in modern product listings. As a

ECOM'26: SIGIR Workshop on eCommerce, Jul 24, 2026, Melbourne, Australia

ruiy@meta.com (R. Yan); kevh@meta.com (K. Huang)

result, they often struggle with newly listed products, long-tail inventory and fine-grained semantic understanding [7, 8].

Large-scale vision–language pretraining introduced a complementary paradigm. Models such as CLIP [9], SigLIP [10] and SigLIP2 [11] learn aligned image–text embeddings from web-scale paired data, typically through dual encoders trained with contrastive objectives. These models transfer well to retrieval and classification tasks, and are particularly valuable for cold-start products because representations can be derived directly from content. However, content-based embeddings remain incomplete for commercial applications. They often encode what a product looks like or says, but not what it means to a shopper. Products with similar content may serve different needs, whereas products that satisfy the same shopping intent may differ substantially in appearance or category—a child's strawberry-themed slipper and a dinosaur-themed slipper serve the same need, yet share few visual or textual attributes. Content-only models therefore tend to over-emphasize near-duplicate similarity, producing homogeneous rankings while under-representing diversity and behavioral relevance [12, 13].

Recent multimodal large language models (MLLMs) provide a promising foundation for richer product representation learning. Models such as LLaVA [14], Qwen3-VL [15] and Gemma 3 [16] demonstrate strong visual reasoning, language understanding and deep cross-modal integration. Unlike dual-encoder vision–language models, MLLMs fuse images and text through joint self-attention over concatenated visual and textual tokens, enabling richer interactions between modalities. Emerging work has shown that these models can also be repurposed as embedding encoders. Methods such as E5-V [17], VLM2Vec [18] and GME [19] adapt pretrained MLLMs into retrieval-oriented representation models, often outperforming traditional dual-encoder systems on multimodal benchmarks. However, these methods are trained primarily with semantic supervision, such as image–text alignment or query–document relevance, and make limited use of the behavioral feedback available on e-commerce platforms.

Here we present Pekka, a product embedding framework that bridges this gap by adapting pretrained MLLMs with co-click contrastive learning. Pekka jointly encodes product images and text through the MLLM backbone, leveraging deep cross-modal interaction to capture rich multimodal semantics, and aligns these representations with user intent using curated co-clicked product pairs. Through fine-tuning, Pekka produces compact 256-dimensional embeddings that capture both product semantics and shopping intent, improving behavioral retrieval while transferring effectively to search ranking and cross-modal retrieval tasks. Our key contributions are as follows:

1. **Behavior-grounded MLLM product embeddings.** We introduce Pekka as a generalizable framework to train multimodal large language models for product embeddings using large-scale user engagement signals. By leveraging joint attention between image and text tokens, Pekka captures richer semantics than dual-encoder approaches while grounding representations in observed shopping behavior.
2. **Scalable co-click supervision.** We develop a data curation pipeline that converts noisy browsing logs into high-quality co-clicked product pairs for contrastive learning. This supervision enables embeddings that better reflect intent, diversity and substitutability than content-only models.
3. **Cross-domain transfer of behavioral signals.** Through comprehensive evaluation on four public benchmarks spanning behavioral retrieval, cross-modal retrieval, and search ranking, as well as internal attribute classification tasks, we show that co-click supervision from a single platform transfers effectively to unseen datasets, product domains, and task types—outperforming both dual-encoder and MLLM-based baselines trained on semantic supervision alone.

2. Related Work

2.1. Product Representation Learning

Early product representation learning approaches often modeled individual modalities in isolation. On the visual side, metric-learning methods such as lifted structured loss [20], Proxy-NCA [21], and multi-

similarity loss [22] learned discriminative image embeddings by enforcing distance constraints between product images. More recently, self-supervised vision models such as DINOv2 [23] and DINOv3 [24] produce strong general-purpose visual features via self-distillation with no labeled data. These methods are effective for visually distinctive categories such as apparel or furniture, but struggle when key differences are expressed through textual attributes such as size, compatibility or material. On the text side, Sentence-BERT [25] enabled efficient dense retrieval over titles and descriptions through siamese transformer architectures, yet text-only representations miss visual signals such as style, color and design patterns that often shape user preference.

The emergence of large-scale vision–language pretraining introduced a multimodal alternative. Models such as CLIP [9], SigLIP [10] and SigLIP2 [11] learn aligned image–text representations from web-scale paired data using contrastive objectives. These embeddings transfer strongly to retrieval and classification tasks, and naturally support cold-start products because representations can be derived directly from content. Domain-adapted variants such as FashionCLIP [26] further showed that vertical-specific fine-tuning can substantially improve product retrieval. Large-scale e-commerce efforts including M5Product [27] and CommerceMM [28] extended this paradigm through multimodal contrastive learning over millions of products. However, most such approaches rely on dual encoders that process images and text largely independently before similarity matching, limiting deeper cross-modal reasoning and behavioral grounding.

2.2. Multimodal LLMs for Representation Learning

Recent multimodal large language models (MLLMs) provide a richer foundation for representation learning by jointly processing visual and textual tokens through shared transformer layers [14, 15, 16]. Unlike dual-encoder vision–language models, these systems allow repeated interaction between modalities through joint self-attention before embedding extraction, enabling stronger reasoning over compositional semantics, context and fine-grained attribute interactions. Bridging architectures such as BLIP-2 [29] further demonstrated that large language models can be efficiently adapted for vision–language understanding without end-to-end retraining.

A growing body of work has explored repurposing MLLMs as embedding encoders. E5-V [17] showed that pretrained MLLMs can generate universal multimodal embeddings through prompting alone. VLM2Vec [18] demonstrated that contrastive fine-tuning on multimodal retrieval corpora yields substantial gains over zero-shot prompting. UniIR [30] proposed unified multimodal retrievers for heterogeneous query–document settings, while GME [19] introduced generative multimodal embeddings for retrieval. Related ideas have emerged in recommendation [31, 32] and e-commerce product understanding [33]. However, most existing MLLM-based embedding methods remain trained with semantic supervision, making limited use of the behavioral feedback available in commerce systems.

2.3. User Engagement Signals in Product Embeddings

Behavioral supervision has a long history in recommender systems because user interactions often reveal relevance more directly than content similarity alone [34]. Early neural approaches such as Item2Vec [4] adapted Word2Vec [35] to learn item embeddings from co-purchase or session sequences by treating neighboring items as context, while Meta-Prod2Vec [2] incorporated product side information. Graph-based systems such as PinSage [5] scaled collaborative embeddings to web-scale catalogs using user–item graphs, and Airbnb [1] showed that session-level listing embeddings can drive real-time search personalization at industrial scale. Sequential recommendation models, including GRU4Rec [36], SASRec [37] and BERT4Rec [38], learned representations implicitly through next-item prediction, while LightGCN [6] simplified graph propagation for collaborative filtering. At the system level, YouTube [39] demonstrated that deep two-tower models trained on implicit watch signals can serve billions of users.

These methods capture behavioral relevance effectively, but typically operate over item identities, interaction graphs or sparse side features rather than rich multimodal product content. As a result, they

often struggle with newly listed products, long-tail inventory and fine-grained semantic understanding [7, 8].

The choice of behavioral signal is itself important. Purchase events provide strong conversion intent but are relatively sparse, whereas session-level browsing interactions are far more abundant and reflect a broader range of shopping behaviors, including comparison, substitution and exploratory discovery. Products clicked together within a session may therefore encode both complementarity and substitutability signals that purchase-only supervision can miss. Pekka builds on this observation by using curated co-click pairs as large-scale contrastive supervision, combining behavioral relevance with multimodal semantic understanding in a unified embedding space.

3. Method

3.1. Problem Setup

Let $\mathcal{P} = \{p_1, \ldots, p_M\}$ denote a product catalog, where each product $p = (I, T)$ contains visual content I (image) and metadata T. We observe browsing sessions $\mathcal{S} = \{s_1, \ldots, s_K\}$, where each session is an ordered sequence of clicked products. For two products p_i and p_j, we define their co-click count as the number of sessions in which both appear:

$$c(p_i, p_j) = \left|\{s \in \mathcal{S} : p_i \in s,\ p_j \in s\}\right|. \tag{1}$$

Frequent co-clicks often reflect shared intent, substitutability, or complementarity. Our goal is to learn an embedding function $f_\theta(I, T) \in \mathbb{R}^d$ satisfying two criteria: (i) behavioral alignment, so that products frequently co-clicked have high similarity; and (ii) semantic generalization, so that the embedding space remains useful for unseen products and downstream tasks such as search ranking and classification. These objectives are not always aligned: user behavior may connect visually different products, whereas visually similar products may satisfy different needs. Pekka addresses this through three components: (1) an MLLM encoder that jointly models image and text; (2) a scalable pipeline that mines high-quality co-click pairs from browsing logs; and (3) contrastive optimization with in-batch negatives, offline-mined hard negatives, and frequency-weighted supervision.

3.2. MLLM Encoder

Pekka converts a pretrained MLLM into a product embedding model. Unlike dual-encoder systems that process image and text independently, Pekka jointly encodes both modalities through a shared transformer, enabling rich cross-modal interaction before embedding extraction. The encoder is defined as

$$f_\theta = h_\phi \circ g_\psi \circ \left[\pi_\omega(\mathrm{VisionEnc}(I)); \mathrm{Tokenize}(T)\right], \tag{2}$$

where $\mathrm{VisionEnc}$ is a pretrained vision encoder, π_ω projects visual features into the language-model hidden space, g_ψ is the MLLM backbone, and h_ϕ is a lightweight embedding head. This formulation applies to any autoregressive MLLM with a vision encoder frontend. In our experiments, we evaluate Qwen3-VL [15], Llama 3.2-11B [40], and Gemma 3 [16] backbones. Visual tokens are concatenated with tokenized metadata fields such as title, description, category, and attributes. Missing fields are omitted rather than padded, allowing a single model to handle heterogeneous catalogs with varying metadata quality.

To obtain a fixed-size representation from an autoregressive LLM, we append a learnable `[CLS]` token to the multimodal sequence and use its final hidden state as the product representation. Because this token attends to all preceding image and text tokens under the causal attention mask, it learns to aggregate the signals most informative for product similarity. The `[CLS]` hidden state $\mathbf{z}$ is projected into a compact retrieval vector:

$$\mathbf{e} = W_h \mathbf{z} + \mathbf{b}, \qquad \mathbf{e} \in \mathbb{R}^d. \tag{3}$$

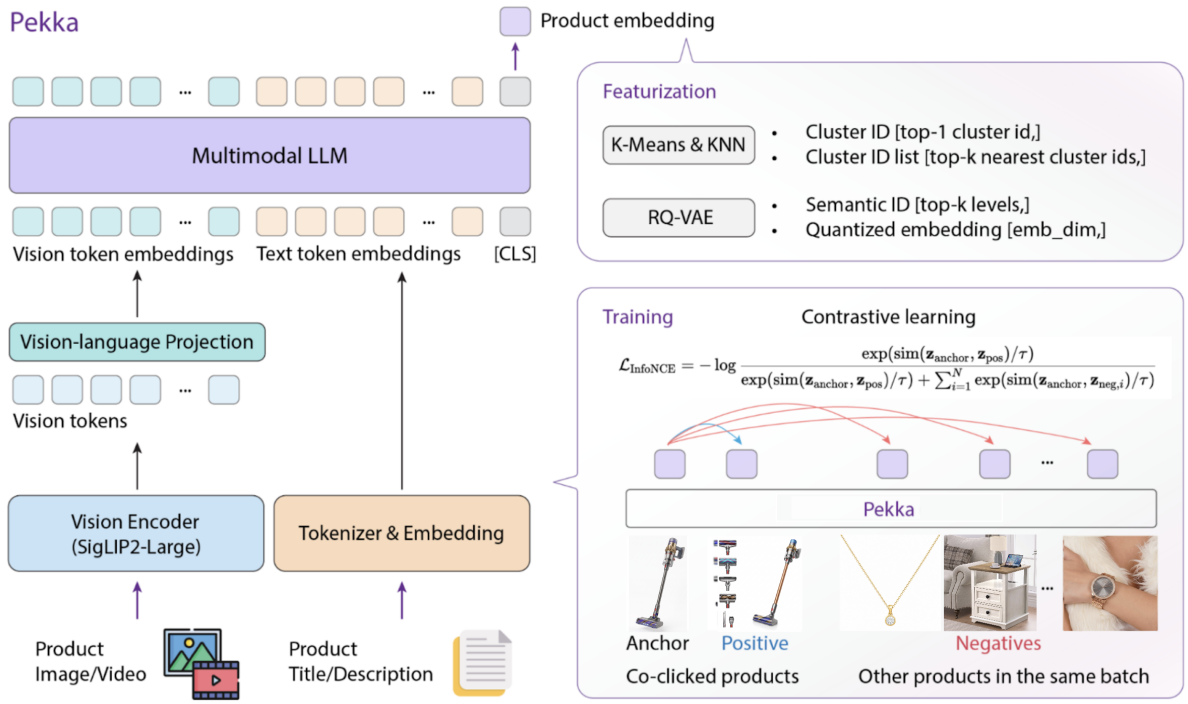

Figure 1: Overview of the Pekka framework. **Left:** A vision encoder and tokenizer process product images and text metadata into a shared token sequence. A multimodal LLM with LoRA adapters processes the concatenated tokens, and a learnable [CLS] token aggregates the sequence into a compact embedding. **Top right:** Downstream featurization derives discrete cluster IDs (K-Means) and hierarchical semantic IDs (RQ-VAE) from frozen embeddings. **Bottom right:** Training optimizes an InfoNCE contrastive loss over co-clicked product pairs with in-batch and hard negatives.

We use $d = 256$, balancing retrieval quality with storage and latency constraints. Embeddings are L2-normalized before similarity computation, so $\mathrm{sim}(\cdot, \cdot)$ denotes cosine similarity throughout.

To preserve pretrained multimodal knowledge, we freeze the backbone weights and adapt only LoRA [41] modules on all linear projections. For each frozen weight matrix $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, LoRA introduces a low-rank update so the effective weight becomes

$$W = W_0 + BA, \qquad B \in \mathbb{R}^{d_{\text{out}} \times r},\ A \in \mathbb{R}^{r \times d_{\text{in}}}, \tag{4}$$

where $r \ll \min(d_{\text{in}}, d_{\text{out}})$. Only A, B, the [CLS] token and the embedding head h_ϕ are optimized; the backbone g_ψ and vision encoder remain frozen. Catalog completeness varies across products; Pekka naturally supports missing modalities by omitting unavailable inputs while using the same encoder. During training, we apply stochastic modality dropout to improve robustness: for each sample, independent Bernoulli masks $m_I \sim \text{Bernoulli}(1-p)$ and $m_T \sim \text{Bernoulli}(1-p)$ gate the image and text token sequences respectively, with $p = 0.1$. To guarantee a valid input, we condition on $(m_I, m_T) \neq (0, 0)$.

3.3. Co-click Data Curation

Browsing logs provide abundant supervision but also contain accidental clicks, bots, and weak associations. Training directly on raw pairs can pull unrelated products together. We therefore construct curated co-click pairs using a two-stage pipeline. First, we apply frequency filtering, retaining only pairs observed multiple times:

$$\mathcal{D}_{\text{freq}} = \{(p_i, p_j) : c(p_i, p_j) \geq \kappa\}, \tag{5}$$

where $\kappa = 2$ in our experiments. This removes one-off coincidences while preserving stable behavioral patterns. Second, we apply semantic filtering using a pretrained vision-language encoder f_0 (CLIP

ViT-L/14 in our experiments):

$$\mathcal{D}_{\text{train}} = \{(p_i, p_j) \in \mathcal{D}_{\text{freq}} : \text{sim}(f_0(p_i), f_0(p_j)) \geq \delta\}, \tag{6}$$

where $\delta = 0.65$, selected to remove clear outliers (e.g., accidental cross-category co-clicks) while retaining visually diverse but behaviorally meaningful pairs. Using anonymized browsing logs over a fixed time window, the pipeline yields tens of millions of curated training pairs.

3.4. Contrastive Optimization

The training objective encourages co-clicked products to be embedded nearby while pushing unrelated products apart. Given a mini-batch of N co-clicked pairs $\{(p_i, p_i')\}_{i=1}^{N}$, we compute embeddings $\mathbf{e}_i = f_\theta(p_i)$ and $\mathbf{e}_i' = f_\theta(p_i')$ and optimize a symmetric InfoNCE loss [42]:

$$\mathcal{L} = \tfrac{1}{2}\big(\mathcal{L}_{\text{fwd}} + \mathcal{L}_{\text{bwd}}\big), \tag{7}$$

where $\mathcal{L}_{\text{bwd}}$ reverses anchor and target roles, and with $s(\mathbf{a}, \mathbf{b}) = \text{sim}(\mathbf{a}, \mathbf{b})/\tau$,

$$\mathcal{L}_{\text{fwd}} = -\frac{1}{\sum_i w_i} \sum_{i=1}^{N} w_i \log \frac{e^{\,s(\mathbf{e}_i, \mathbf{e}_i')}}{\sum_{j=1}^{N} e^{\,s(\mathbf{e}_i, \mathbf{e}_j')} + e^{\,s(\mathbf{e}_i, \mathbf{e}_i^-)}}. \tag{8}$$

Intuitively, this loss maximizes the similarity of each anchor $\mathbf{e}_i$ with its co-clicked partner $\mathbf{e}_i'$ relative to all other products in the batch. The denominator provides two sources of contrast: in-batch negatives (the sum over j, which includes all other batch samples) and an offline-mined hard negative $\mathbf{e}_i^-$ (described below). The learnable temperature τ controls the sharpness of the similarity distribution. To enlarge the effective negative pool, embeddings are gathered across all devices so each sample is contrasted against the global batch.

Frequency weighting. Not all co-click pairs carry equal signal—products co-clicked many times across independent sessions reflect stronger behavioral association than pairs that co-occur only at the minimum threshold κ. We encode this prior through a log-scaled weight:

$$w_i = \log\big(1 + c(p_i, p_i')\big)\,, \tag{9}$$

so that frequently co-clicked pairs receive proportionally more supervision while the logarithm prevents a small number of very popular pairs from dominating the gradient.

Hard negative mining. Random in-batch negatives are typically drawn from different product categories and are easy for the model to distinguish. To learn fine-grained within-category distinctions—such as telling apart product variants, sizes, or formulations—we augment each anchor with a hard negative mined from the same leaf category. Let $\mathcal{C}(p)$ denote the set of products sharing the same leaf category as p and $\mathcal{P}^+(p)$ its known co-click positives. For each anchor p_i, we select the most confusable non-positive under a fixed pretrained checkpoint θ':

$$p_i^- = \underset{p \in \mathcal{C}(p_i) \setminus \mathcal{P}^+(p_i)}{\arg\max} \text{sim}(f_{\theta'}(p_i),\ f_{\theta'}(p))\,. \tag{10}$$

This product is the nearest neighbor within the category that users do *not* co-click with the anchor. Its embedding $\mathbf{e}_i^- = f_\theta(p_i^-)$ enters the InfoNCE denominator (Equation 8), providing targeted contrast that in-batch negatives alone are unlikely to supply.

3.5. Downstream Feature Extraction

Although Pekka embeddings are primarily used for nearest-neighbor retrieval, they can also be converted into discrete features for integration with existing ranking systems that operate on categorical inputs.

K-Means clustering. Given the full set of frozen product embeddings $\{\mathbf{e}_1, \ldots, \mathbf{e}_M\}$, we partition the embedding space into C clusters by minimizing the within-cluster sum of squares:

$$\min_{\{\boldsymbol{\mu}_c\}_{c=1}^{C}} \sum_{i=1}^{M} \min_{c \in [C]} \|\mathbf{e}_i - \boldsymbol{\mu}_c\|_2^2 . \tag{11}$$

Each product receives a hard cluster assignment $a(p) = \arg\min_c \|\mathbf{e} - \boldsymbol{\mu}_c\|_2$. To capture soft membership, we additionally record the top-k nearest centroid IDs via approximate nearest-neighbor (ANN) search, yielding a k-dimensional categorical feature vector per product.

Hierarchical semantic IDs (RQ-VAE). For hierarchical discrete codes, we train a Residual Quantization VAE [43] on the frozen embeddings. The RQ-VAE encodes each embedding into a latent space and applies L levels of residual vector quantization: at each level, the quantizer selects the nearest codebook entry and passes the residual to the next level, producing a hierarchical *semantic ID* $(q_1, q_2, \ldots, q_L)$ per product. Products sharing longer prefixes in this ID are progressively more similar, capturing structure from coarse category groupings down to fine-grained style distinctions. These derived features—cluster IDs and semantic IDs—provide complementary signals for downstream CTR and CVR ranking models; their production impact is quantified in Section 5.

4. Experiments

4.1. Setup

To evaluate whether behavioral fine-tuning generalizes beyond the training domain, we use four public datasets from three independent e-commerce platforms (Table 1). Together, they cover behavioral retrieval, cross-modal content retrieval, text-only search ranking, and multimodal search ranking.

- **H&M Personalized Fashion** [44] — behavioral retrieval from purchase signals. The dataset contains 31M transactions over 105K fashion articles. Co-purchase pairs are constructed within 7-day customer windows, yielding 20K queries and 37K candidates for nearest-neighbor retrieval over the full candidate set. This benchmark evaluates whether embeddings capture shopping intent reflected in purchase behavior.
- **M5Product** [27] — cross-modal content retrieval. The dataset contains 6M multimodal products from a Chinese e-commerce platform across 6,232 categories, with relevance defined by shared leaf-category membership. Evaluation uses the official test split with 24K queries and a 1.2M-product candidate pool, testing whether behavioral fine-tuning preserves semantic category structure.
- **ESCI** [45] — text-only search ranking. Derived from the Amazon KDD Cup 2022, ESCI provides four-level human relevance labels (Exact, Substitute, Complement, Irrelevant) over 2K queries and 1.2M candidates. Products are ranked by embedding similarity; Pekka encodes queries as text-only inputs by omitting visual tokens. This benchmark evaluates asymmetric query-to-product matching without visual information.
- **SQID** [46] — multimodal search ranking. SQID extends ESCI with product images for 36% of items and a 1.8M-product candidate pool across 2K queries. This benchmark evaluates whether visual product information improves large-scale search ranking.

Metrics. For H&M and M5Product, where relevance is binary, we report Recall@k (R@k), defined as the fraction of queries for which at least one relevant item appears among the top-k retrieved results. For ESCI and SQID, where relevance is graded, we additionally report nDCG@10, which discounts lower-ranked results and assigns higher reward to more relevant products appearing near the top of the ranking.

Table 1
Public benchmark datasets. Each dataset targets a different aspect of product embedding quality, spanning three independent e-commerce platforms.

Dataset	Modalities	Queries	Candidates	Task
H&M	Text+Image	20K	37K	Behavioral retrieval
M5Product	Text+Image	24K	1.2M	Cross-modal retrieval
ESCI	Text	2K	1.2M	Search ranking
SQID	Text+Image	2K	1.8M	Search ranking

Query and product encoding. For search benchmarks, queries and products are encoded with the same model. Queries are represented as text-only inputs by placing the query string in the textual field and omitting visual tokens. Products are encoded from available metadata and, for SQID, images when available. Retrieval is performed by cosine similarity between each query embedding and all candidate product embeddings. The same modality availability is used consistently for all models whenever supported; models that cannot process a modality are evaluated using their supported inputs.

Baselines. We compare against representative models spanning three architecture families. *Vision-only:* DINOv3 ViT-L/16 [24] (1024-d), a self-supervised vision model evaluated only on image-available benchmarks. *Dual-encoder vision–language:* CLIP ViT-L/14 [9] (768-d) and SigLIP2-SO400M [11] (1152-d), which encode image and text separately. *MLLM-based:* E5-V [17] (4096-d), which adapts a multimodal LLM for embedding generation. We also include the base Qwen3-VL-2B (2048-d) without fine-tuning to isolate the contribution of the pretrained backbone. No model, including Pekka, is trained on any benchmark dataset. Importantly, all baselines are evaluated off-the-shelf without e-commerce fine-tuning, whereas Pekka is fine-tuned on co-click data from a separate platform. This evaluation therefore measures cross-domain transfer of behavioral supervision rather than a controlled architectural comparison under matched data. The controlled comparison—in which all models are trained on the same co-click data—is provided in Section 4.4.

Implementation. We use Qwen3-VL-2B as the default backbone based on the quality–efficiency analysis in Section 4.5. The Qwen3-VL vision encoder is a SigLIP2-based ViT; it and the visual projection π_ω are frozen throughout training. Fine-tuning applies LoRA (rank 16, $\alpha = 32$) to all linear layers of the language model. We optimize an InfoNCE objective with in-batch and hard negatives, using AdamW ($\mathrm{lr} = 2 \times 10^{-4}$, 500-step linear warmup, cosine decay) and initialize the learnable temperature at $\tau = 0.05$. Embeddings are projected to 256 dimensions. Training runs for one epoch over the curated co-click training pairs on 8×A100 80GB GPUs (batch size 64 per device; global batch size 512; image resolution 384×384). The same recipe is used for all backbone variants in Section 4.5.

4.2. Public Benchmark Results

Table 2 summarizes performance on four public benchmarks spanning complementary retrieval settings: behavioral retrieval (H&M), cross-modal retrieval (M5Product), text-only search ranking (ESCI), and multimodal search ranking (SQID). Across all tasks, Pekka achieves the strongest overall performance (Fig. 2), despite using compact 256-dimensional embeddings that are 3–16× smaller than those of competing baselines. This consistent advantage across heterogeneous benchmarks indicates that the learned representations generalize effectively across modality, domain, and task objective.

Notably, Pekka is trained solely on co-click data from a separate platform, with no overlap with any evaluation dataset. The observed gains therefore reflect cross-domain transfer rather than in-domain optimization, suggesting that large-scale user interactions encode transferable signals of substitutability, preference, and shopping intent that can be distilled into compact embeddings.

Table 2
Public benchmark results (%). H&M: behavioral co-purchase retrieval; M5Product: cross-modal retrieval; ESCI: text-only search ranking; SQID: multimodal search ranking. No model is trained on benchmark data; Pekka is fine-tuned on separate co-click data. Best in **bold**, second underlined.

	H&M			M5Product			ESCI		SQID	
Method	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	**nDCG@10**	**R@10**	**nDCG@10**	**R@10**
Vision-Only										
DINOv3 ViT-L/16* [24]	4.26	15.54	23.18	7.69	9.95	11.03	–	–	–	–
Vision-Language										
CLIP ViT-L/14 [9]	22.65	44.10	52.25	50.68	66.82	72.54	24.48	47.80	22.61	45.05
SigLIP2-SO400M [11]	26.25	44.12	50.01	89.53	96.10	97.39	13.63	53.20	10.99	43.40
MLLM Embedding										
E5-V [17]	18.60	29.90	35.07	81.18	90.31	92.70	12.02	40.50	23.79	53.10
Qwen3-VL-2B (base) [15]	0.21	36.54	44.69	82.11	94.49	96.63	0.10	0.55	0.03	0.30
Pekka	**28.81**	**49.34**	**56.22**	**89.89**	**96.86**	**97.99**	**42.26**	**77.40**	**41.47**	**76.60**

*Vision-only; cannot process text-only queries (ESCI) or text-dominant queries (SQID).

Behavioral retrieval (H&M). On the benchmark grounded in real purchase behavior, Pekka achieves 28.81% R@1, 49.34% R@5, and 56.22% R@10, exceeding CLIP (22.65% / 44.10% / 52.25%) and E5-V (18.60% / 29.90% / 35.07%). This result is notable because Pekka is trained on co-click rather than co-purchase data, yet transfers effectively to purchase-based retrieval. The finding suggests that browsing and purchasing behaviors share common latent structure, enabling click signals to serve as scalable supervision for downstream commercial relevance. By contrast, the vision-only DINOv3 baseline attains only 4.26% R@1, underscoring the importance of textual metadata for capturing user intent and functional product similarity.

Content retrieval (M5Product). Pekka attains 89.89% R@1 and 97.99% R@10, slightly surpassing SigLIP2 (89.53% / 97.39%) and substantially outperforming E5-V (81.18% / 92.70%). Importantly, Pekka is optimized solely with behavioral co-click pairs and does not use category labels during training, yet achieves the strongest results on a benchmark whose ground truth is defined by category correspondence. This indicates that behavioral supervision does not erode semantic fidelity inherited from the pretrained backbone, but can instead refine it. Such retention of content understanding is particularly valuable for cold-start products that lack historical engagement signals.

Search ranking (ESCI and SQID). Pekka achieves 42.26% nDCG@10 on text-only ESCI and 41.47% on multimodal SQID, approximately 1.7× higher than the strongest baseline on both tasks. These gains suggest that co-click fine-tuning learns relevance signals that align closely with human judgments of search quality. In addition, the learned `[CLS]` representation appears well suited to asymmetric matching between short user queries and longer product descriptions, a setting in which generic pooled embeddings often underperform.

Comparing ESCI and SQID further reveals how different architectures utilize visual information, while noting that SQID contains a larger candidate pool (1.8M versus 1.2M), making direct numerical comparison imperfect. E5-V improves markedly when images are introduced (12.02% to 23.79% nDCG), whereas CLIP degrades slightly (24.48% to 22.61%), suggesting that visual features do not uniformly translate into ranking gains. Pekka remains stable across both settings (42.26% versus 41.47%), indicating robust multimodal fusion under changing retrieval conditions. SigLIP2, despite strong performance on M5Product, records relatively low nDCG on ESCI (13.63%) and SQID (10.99%), consistent with weaker alignment to short-query ranking tasks than models benefiting from broader text supervision.

Role of the pretrained backbone. The base Qwen3-VL-2B model, evaluated without contrastive fine-tuning using a simple mean-pooling extraction protocol, provides a diagnostic view of the pretrained

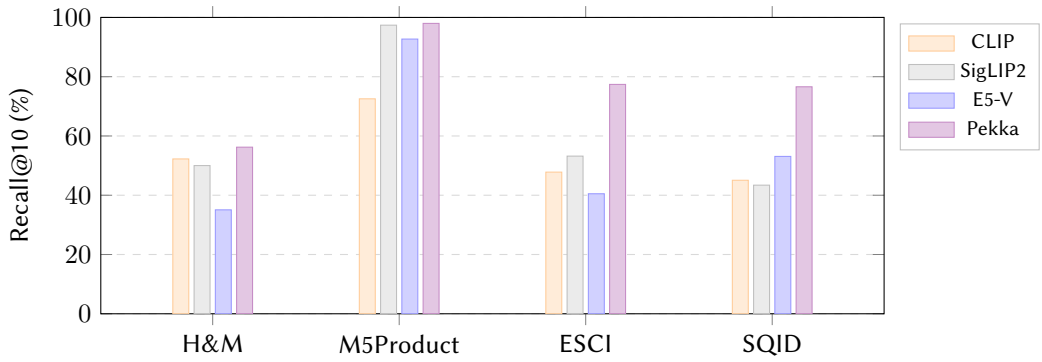

Figure 2: Recall@10 (%) across four public benchmarks (DINOv3 and base Qwen3-VL-2B omitted; see Table 2). Pekka achieves the highest recall on all datasets, with the largest gains on search ranking (∼24 percentage points over the best baseline on ESCI and SQID).

Table 3
Attribute classification accuracy (%) using linear probes on frozen embeddings. Style, color, and category labels are derived from structured catalog metadata. Best in **bold**, second underlined.

Method	Dim	Style	Color	Category	Avg.
DINOv3 ViT-L/16 [24]	1024	65.7	10.7	93.0	56.5
CLIP ViT-L/14 [9]	768	77.3	20.2	93.5	63.7
SigLIP2-SO400M [11]	1152	66.7	18.2	93.4	59.4
E5-V [17]	4096	76.9	21.1	93.6	63.9
Qwen3-VL-2B (base) [15]	2048	71.2	21.7	93.3	62.1
Pekka	256	**79.8**	**23.1**	**94.5**	**65.8**

backbone rather than a fully optimized MLLM embedding baseline. On content retrieval, the base model is already competitive: M5Product R@1 reaches 82.11%, surpassing E5-V and approaching SigLIP2. On search ranking, however, performance collapses (ESCI nDCG = 0.10%, SQID nDCG = 0.03%). Mean pooling under a fixed prompting template likely causes short queries to map to highly similar embeddings, yielding near-random retrieval. Co-click fine-tuning, together with a learned [CLS] aggregation token, resolves this failure mode under our extraction protocol and raises ESCI nDCG from 0.10% to 42.26%.

Together, these results indicate that pretrained multimodal backbones provide strong semantic priors, but require behavioral alignment to become effective retrieval systems. By integrating multimodal understanding with large-scale user interaction signals, Pekka achieves consistent gains across behavioral retrieval, semantic matching, and real-world search ranking tasks.

4.3. Product Attribute Classification

Product embeddings can be used as feature vectors to predict product-level metadata, including categorical product attributes derived from structured catalog annotations. Here we evaluate embedding quality on product attribute prediction tasks derived from an e-commerce catalog, where ground-truth labels are curated from structured product metadata. For each task, a linear probe (logistic regression with 5-fold stratified cross-validation) is trained on frozen embeddings, isolating representation quality from end-to-end classifier capacity. Results are summarized in Table 3.

Pekka achieves the highest accuracy on all three attributes while using only 256-dimensional embeddings, which are 3–16× smaller than those of all baselines. Among baselines, E5-V is strongest overall (63.9% average), while CLIP performs comparably (63.7%) despite lower dimensionality. The base Qwen3-VL-2B reaches 62.1%, showing that pretrained backbones already encode useful attribute information but remain inferior to behaviorally aligned embeddings.

Table 4
Co-click retrieval results (%) on 100K held-out query products. All models are trained on the same internal co-click data. Best in **bold**, second underlined.

Method	Dim	R@5	R@10	R@50
ResNeXt-101	256	37.19	46.67	67.49
Text Transformer	1024	47.76	56.18	74.10
CLIP (fine-tuned)	1024	51.41	60.16	77.75
Pekka	256	**66.79**	**74.38**	**87.39**

The largest gains are observed for style (79.8% vs. 77.3% for CLIP) and color (23.1% vs. 21.7% for base Qwen3-VL-2B), suggesting that user interaction data provides informative supervision for attributes that shape browsing and purchase behavior. Category prediction is strong across nearly all models ($\geq$93%), while color is hard for every encoder. Several factors limit it: adjacent shades (navy vs. black), general-vs-specific labels (purple vs. mauve), multi-colored items collapsed to a single dominant color, and noise from catalog labels or background and packaging colors. Even so, Pekka leads on all three attributes, suggesting that behavioral supervision helps recover subtle properties that content alone misses.

4.4. Co-Click Retrieval

To evaluate behavioral alignment directly, we construct a co-click retrieval benchmark from 100K held-out query products, split by product identity so that no evaluation item appears during training. All models are trained on the same internal e-commerce data at comparable scale.

We compare against several widely used industry product-embedding baselines: *ResNeXt-101*, an image-only CNN trained with ArcFace metric learning (256-d); *Text Transformer*, a text-only encoder with contrastive fine-tuning (1024-d); and *CLIP (fine-tuned)*, a dual-encoder adapted to product data using concatenated image and text embeddings (1024-d). This model is distinct from the generic pretrained CLIP evaluated in Table 2. Results are shown in Table 4.

Pekka achieves the best performance at all retrieval depths, reaching 66.79% R@5, 74.38% R@10, and 87.39% R@50. At R@5, this is a 30% relative improvement over the strongest baseline, CLIP (fine-tuned), while using embeddings that are 4$\times$ smaller. These gains indicate that jointly learned multimodal representations are more effective than larger systems built from separate modality encoders.

The image-only ResNeXt-101 baseline performs substantially worse (37.19% R@5), showing that visual similarity alone is insufficient to capture shopping intent. The Text Transformer improves over image-only retrieval (47.76%), indicating that textual metadata carries stronger behavioral signal, but remains inferior to multimodal models. CLIP (fine-tuned) further benefits from combining image and text, yet still trails Pekka, likely because the two modalities are encoded independently before fusion.

By contrast, Pekka jointly encodes images and text via the MLLM backbone, enabling deeper cross-modal interactions before projection into the embedding space. Combined with co-click supervision, this yields representations that better capture user preference and product substitutability.

Figure 3 shows the relative improvement in category-level Recall@10 of Pekka over the fine-tuned CLIP baseline. Pekka yields positive gains across all 24 product categories. The largest improvements occur in *Housing* (+60%), where text-heavy listings with diverse imagery benefit from co-click supervision that captures preference patterns beyond content similarity; *Clothing & Accessories* (+37%), where visually diverse but functionally substitutable products require joint reasoning over style and occasion; and *Musical Instruments* (+31%), *Electronics* (+28%), and *Jewelry & Watches* (+25%), where retrieval hinges on fine-grained cross-modal distinctions—specifications, materials, and design details—that dual encoders process in isolation.

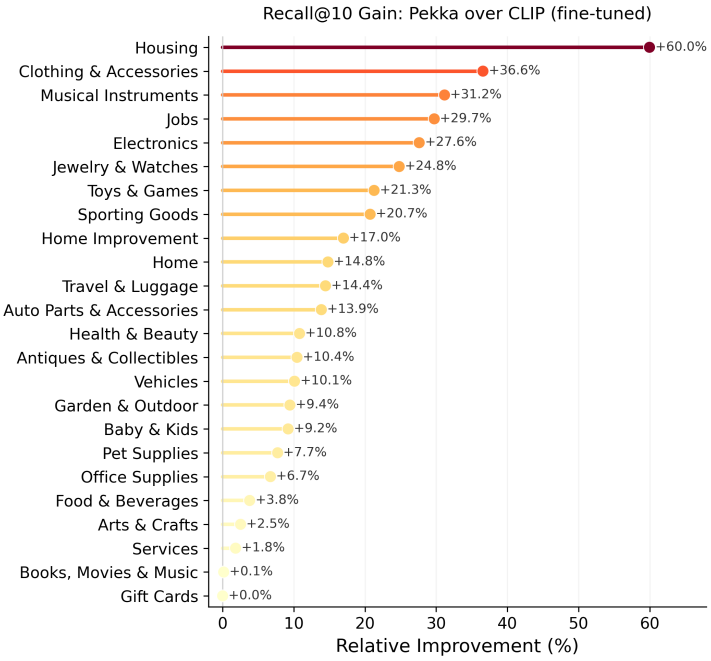

Figure 3: Relative improvement in category-level Recall@10 of Pekka over the fine-tuned CLIP baseline on the co-click retrieval task.

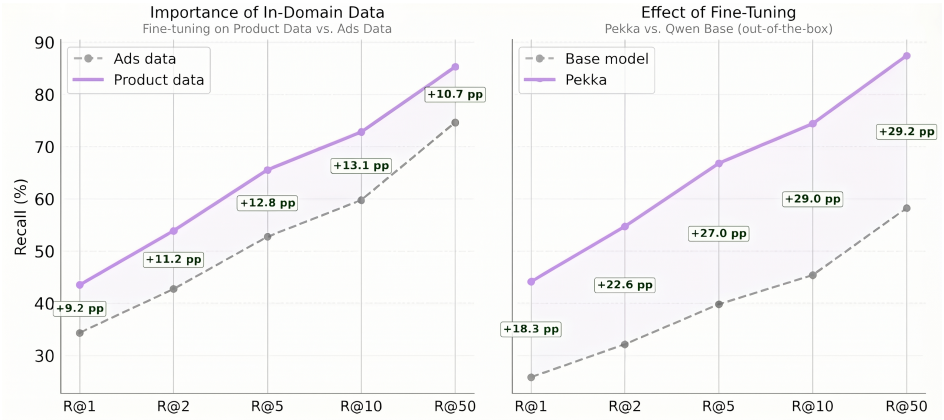

Figure 4: Ablation studies on the co-click retrieval task. **Left:** Product-domain co-click supervision versus out-of-domain behavioral supervision using the same model and training recipe. **Right:** Pekka (fine-tuned, 256-d) versus base Qwen3-VL-2B (no fine-tuning, 2048-d), showing the effect of co-click fine-tuning.

4.5. Ablations and Backbone Study

Ablation studies were conducted on the co-click retrieval task to understand the contributions of in-domain data and fine-tuning, holding all other variables constant (Figure 4).

Effect of supervision data. Product-domain co-click supervision improves retrieval by +9.2 to +10.7 percentage points across recall levels relative to out-of-domain behavioral data under the same architecture and optimization setting (Fig. 4, left). This suggests that retrieval quality benefits not only from scale, but also from alignment between supervision signals and the target product domain.

Effect of fine-tuning. Co-click fine-tuning improves R@10 by +29.0 percentage points over the base Qwen3-VL-2B model (74.38% versus 45.37%) while reducing embedding dimensionality from 2048 to 256 (Fig. 4, right). The gain remains similarly large at deeper retrieval depths (+29.2 points at R@50), indicating that fine-tuning restructures the global embedding space rather than only improving top-ranked matches. These results show that Pekka's performance arises from behavioral adaptation of the pretrained backbone, rather than from the backbone alone.

Table 5
Impact of MLLM backbone on retrieval performance and inference throughput. QPS measured on 8×A100 GPUs. All embeddings are 256-d.

Backbone	R@5	R@10	R@50	QPS
Llama-3.2-11B	66.84	74.22	86.79	416
Gemma-3-4B	69.07	76.33	88.31	312
Qwen3-VL-2B	66.79	74.38	87.39	**840**
Qwen3-VL-4B	67.36	74.91	87.65	504
Qwen3-VL-8B	67.57	75.03	87.68	344
Qwen3-VL-30B-A3B	68.54	75.84	88.15	128

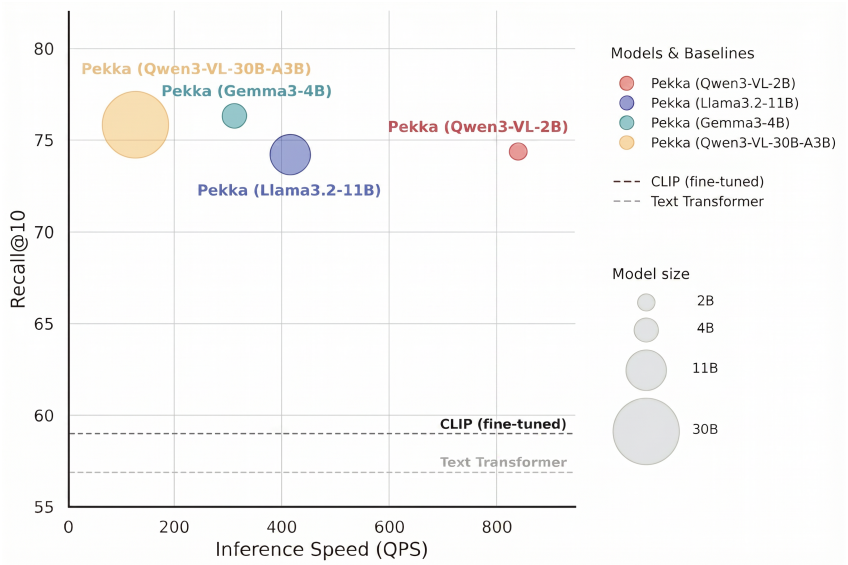

Figure 5: Quality–efficiency Pareto frontier for MLLM backbones on co-click retrieval. Each bubble represents a backbone; size indicates parameter count. Dashed lines mark the CLIP (fine-tuned) and Text Transformer baselines from Table 4. All Pekka variants exceed both baselines; Qwen3-VL-2B offers the best tradeoff at ∼840 QPS.

Table 5 compares retrieval accuracy and inference throughput across MLLM backbones from three model families, all fine-tuned with an identical recipe and projected to 256-dimensional embeddings. Figure 5 shows the resulting quality–efficiency frontier.

Gemma-3-4B achieves the highest retrieval accuracy (69.07% R@5), exceeding Qwen3-VL-2B by +2.3 points, but with 2.7× lower throughput (312 versus 840 QPS). Within the Qwen3-VL family, scaling from 2B to 30B-A3B yields only +1.8 points at R@5 while reducing throughput by 6.6× (840 to 128 QPS), indicating sharply diminishing returns with model scale. Qwen3-VL-30B-A3B approaches Gemma-3-4B in quality, but remains substantially slower, likely owing to mixture-of-experts routing overhead. Qwen3-VL-2B therefore offers the strongest return on inference cost, combining competitive retrieval quality with the highest throughput. We adopt it as the default backbone for deployment, sustaining ∼840 queries per second on 8×A100 GPUs.

Disentangling the source of gains. We separate the effects of fine-tuning, backbone choice, and architecture under matched co-click supervision. *Fine-tuning:* the base-vs-Pekka ablation fixes the backbone and isolates the effect of co-click fine-tuning, yielding a +29 pp gain in R@10. *Backbone:* Table 5 compares three MLLM families fine-tuned with the same recipe. *Architecture:* in Table 4, the fine-tuned dual-encoder CLIP reaches 51.4% R@5, compared with 66.8% for Pekka, suggesting that the remaining gap comes from joint multimodal encoding rather than data or fine-tuning alone.

Query Text	Retrieved Products
Minimalist Fine Jewelry in Timeless Luxury & Investment Pieces: Simple, elegant pieces for daily wear, e.g., gold chain necklace, diamond stud earrings.	
Aesthetic Vases & Greenery in Modern & Curated Home Ambiance: Simple, modern vases with fresh greenery, e.g., ribbed glass or ceramic vases.	
Closet & Drawer Organization Systems in Elevated Daily Essentials & Tech: Bamboo or fabric dividers for accessories.	

Figure 6: Text-to-product retrieval examples from Pekka's unified embedding space. Each row shows a stylistic query and the top-5 retrieved products. The model captures abstract aesthetic concepts ("timeless luxury," "modern ambiance") and functional intent ("closet organization") beyond literal keyword matching, retrieving visually and semantically coherent product sets across categories.

4.6. Qualitative Analysis

Pekka's unified embedding space supports text-to-product retrieval that captures abstract stylistic and functional intent beyond literal keyword matching (Figure 6). For example, the query "Minimalist Fine Jewelry in Timeless Luxury" retrieves rings, necklaces, and earrings spanning different jewelry types but sharing a cohesive aesthetic; "Aesthetic Vases & Greenery in Modern & Curated Home Ambiance" retrieves ceramic vases and planters unified by minimalist design rather than product category; and "Closet & Drawer Organization Systems" retrieves functionally related storage products across different form factors. These results demonstrate that the embedding space learned through co-click fine-tuning captures style, aesthetics, and functional purpose—attributes that are difficult to encode through content similarity alone. We note that this behavioral grounding can also introduce tradeoffs such as color drift and attribute overshadowing, which we discuss in Section 5.

5. Discussion

Pekka shows that adapting MLLM representations with behavioral supervision can yield compact product embeddings that perform strongly across retrieval, ranking, and attribute prediction tasks (Tables 2, 3, 4). Despite using only 256 dimensions, Pekka outperforms substantially larger baselines in our evaluations, offering practical advantages in index size, retrieval latency, and serving cost.

A central finding is that behavioral alignment does not compromise semantic understanding. On M5Product, Pekka matches or exceeds strong pretrained vision–language models on content retrieval despite using no category supervision. This suggests that the pretrained backbone retains sufficient content understanding for new products to receive reasonable embeddings from content alone, though dedicated cold-start evaluation remains an avenue for future work. In production, Pekka embeddings are deployed through the discrete representations described in Section 3.5; integrating them into ranking models yields relative improvements of 0.10% in normalized cross-entropy for second-stage CTR, 0.15% for CVR, and 0.10% for first-stage CTR in internal offline evaluations.

More broadly, co-click supervision reorganizes the embedding space around shopping intent rather than surface similarity. Visually similar products that users treat differently are separated, whereas visually dissimilar products serving the same need are drawn together. Because browsing behavior

reflects both substitutability and complementarity, the resulting geometry is richer than that induced by content-only objectives. This likely underlies the strong transfer to search ranking and attribute prediction observed across benchmarks.

Behavioral supervision also introduces trade-offs. Signals driven by browsing may conflict with strict visual fidelity; for example, users often co-click across colors or styles during exploration, which can weaken sensitivity to fine-grained appearance attributes. Applications requiring exact visual matching may therefore benefit from hybrid objectives that combine behavioral and content-based constraints. In addition, engagement logs may encode exposure or popularity biases: frequently viewed products accumulate more co-clicks regardless of relevance quality. Debiasing strategies and performance analysis stratified by product popularity are important directions for future work.

6. Conclusion

We introduced Pekka, a versatile framework for adapting pretrained multimodal large language models into compact product embedding models through co-click contrastive learning. By combining scalable behavioral data curation with parameter-efficient fine-tuning, Pekka learns product representations that perform strongly across behavioral retrieval, search ranking, cross-modal retrieval, and attribute prediction, while also improving production ranking metrics when used as downstream features. A key finding is that behavioral supervision from a single platform can transfer to unseen datasets, product domains, and task settings, including fashion co-purchase retrieval, category-based cross-modal retrieval, and large-scale search ranking. This suggests that large-scale co-click data captures reusable structure in shopping intent beyond any single marketplace. More broadly, our results indicate that progress in product representation learning may depend not only on larger pretrained multimodal models, but also on how effectively such models are grounded in user behavior. As multimodal foundation models continue to advance, behavioral alignment offers a practical path toward scalable and deployment-ready product retrieval systems.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] M. Grbovic, H. Cheng, Real-time personalization using embeddings for search ranking at Airbnb, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), 2018, pp. 311–320.

[2] F. Vasile, E. Smiber, S. Purushotham, Meta-prod2vec: Product embeddings using side-information for recommendation, in: Proceedings of the 10th ACM Conference on Recommender Systems (RecSys), 2016, pp. 225–232.

[3] M. Grbovic, V. Radosavljevic, N. Djuric, N. Bhamidipati, J. Savla, V. Bhagwan, D. Sharp, E-commerce in your inbox: Product recommendations at scale, in: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2015, pp. 1809–1818.

[4] O. Barkan, N. Koenigstein, Item2Vec: Neural item embedding for collaborative filtering, in: IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), 2016, pp. 1–6.

[5] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, J. Leskovec, Graph convolutional neural networks for web-scale recommender systems, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2018, pp. 974–983.

[6] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, M. Wang, LightGCN: Simplifying and powering graph convolution network for recommendation, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2020, pp. 639–648.

[7] A. I. Schein, A. Popescul, L. H. Ungar, D. M. Pennock, Methods and metrics for cold-start recommendations, in: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2002, pp. 253–260.
[8] M. Volkovs, G. Yu, T. Poutanen, DropoutNet: Addressing cold start in recommender systems, in: Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4957–4966.
[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning (ICML), 2021, pp. 8748–8763.
[10] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2023.
[11] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, et al., SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, arXiv preprint arXiv:2502.14786 (2025).
[12] F. Radlinski, S. Dumais, Redundancy, diversity and interdependent document relevance, in: ACM SIGIR Forum, volume 43, 2009, pp. 46–52.
[13] R. Agrawal, S. Gollapudi, A. Halverson, S. Ieong, Diversifying search results, in: Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (WSDM), 2009, pp. 5–14.
[14] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, Advances in Neural Information Processing Systems (NeurIPS) (2024).
[15] S. Bai, et al., Qwen3-VL technical report, arXiv preprint arXiv:2511.21631 (2025).
[16] Gemma Team, Gemma 3 technical report, arXiv preprint arXiv:2503.19786 (2025).
[17] T. Jiang, M. Ye, S. Huang, S. Luo, Z. Zhang, H. Huang, F. Wei, D. Jiang, E5-V: Universal embeddings with multimodal large language models, arXiv preprint arXiv:2407.12580 (2024).
[18] Z. Jiang, X. Ma, W. Chen, VLM2Vec: Training vision-language models for massive multimodal embedding tasks, arXiv preprint arXiv:2410.05160 (2024).
[19] X. Zhang, Y. Li, W. Wen, M. Xie, Y. Gao, D. Li, M. Zhang, GME: Improving universal multimodal retrieval by multimodal LLMs, arXiv preprint arXiv:2412.16855 (2024).
[20] H. O. Song, Y. Xiang, S. Jegelka, S. Savarese, Deep metric learning via lifted structured feature embedding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4004–4012.
[21] Y. Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, S. Singh, No fuss distance metric learning using proxies, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 360–368.
[22] X. Wang, X. Han, W. Huang, D. Dong, M. R. Scott, Multi-similarity loss with general pair weighting for deep metric learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5022–5030.
[23] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., Dinov2: Learning robust visual features without supervision, Transactions on Machine Learning Research (2024).
[24] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, et al., Dinov3, arXiv preprint arXiv:2508.10104 (2025).
[25] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019, pp. 3982–3992.
[26] P. J. Chia, G. Attanasio, F. Bianchi, S. Terragni, A. R. Manoel, J. Greco, O. L. Scotton, C. Tagliabue, Contrastive language and vision learning of general fashion concepts, Scientific Reports 12 (2022) 18958.
[27] X. Dong, X. Zhan, Y. Wu, Y. Wei, M. C. Kampffmeyer, X. Wei, M. Lu, Y. Wang, X. Liang, M5Product: Self-harmonized contrastive learning for e-commercial multi-modal pretraining, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 21252–21262.
[28] L. Yu, J. Li, C. Deng, B. Liu, Y. Gao, C. Xiong, X. He, H. Hu, Y.-C. Chen, Y.-Y. Wu, Z. Liu, Com-

merceMM: Large-scale commerce multimodal representation learning with omni retrieval, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2022, pp. 4433–4442.

[29] J. Li, D. Li, S. Savarese, S. Hoi, BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, Proceedings of the 40th International Conference on Machine Learning (ICML) (2023).

[30] C. Wei, Y. Chen, H. Chen, H. Hu, G. Zhang, J. Fu, A. Ritter, W. Chen, UniIR: Training and benchmarking universal multimodal information retrievers, in: Proceedings of the European Conference on Computer Vision (ECCV), 2024, pp. 393–412.

[31] C. Zhang, S. Wu, H. Zhang, T. Xie, Y. Gao, Y. Hou, L. Chen, NoteLLM: A retrievable large language model for note recommendation, in: Companion Proceedings of the ACM Web Conference 2024 (WWW), 2024, pp. 170–179.

[32] C. Zhang, S. Wu, H. Zhang, Y. Gao, Y. Hou, L. Chen, NoteLLM-2: Multimodal large representation models for recommendation, arXiv preprint arXiv:2405.16789 (2024).

[33] D. Zhang, C. Fu, Z. Nie, J. Liu, W. Guan, Y. Gao, J. Song, P. Wang, J. Xu, B. Zheng, MOON: Generative MLLM-based multimodal representation learning for e-commerce product understanding, in: Proceedings of the 19th ACM International Conference on Web Search and Data Mining (WSDM), 2026.

[34] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, BPR: Bayesian personalized ranking from implicit feedback, in: Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI), 2009, pp. 452–461.

[35] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems (NeurIPS), 2013, pp. 3111–3119.

[36] B. Hidasi, A. Karatzoglou, L. Baltrunas, D. Tikk, Session-based recommendations with recurrent neural networks, in: International Conference on Learning Representations (ICLR), 2016.

[37] W.-C. Kang, J. McAuley, Self-attentive sequential recommendation, in: IEEE International Conference on Data Mining (ICDM), 2018, pp. 197–206.

[38] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, P. Jiang, BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformers, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM), 2019, pp. 1441–1450.

[39] P. Covington, J. Adams, E. Sargin, Deep neural networks for YouTube recommendations, in: Proceedings of the 10th ACM Conference on Recommender Systems (RecSys), 2016, pp. 191–198.

[40] A. Grattafiori, A. Dubey, A. Jauhri, et al., The Llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[41] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations (ICLR), 2022.

[42] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, in: Advances in Neural Information Processing Systems (NeurIPS), 2018.

[43] D. Lee, C. Kim, S. Kim, M. Cho, W.-S. Han, Autoregressive image generation using residual quantization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[44] H&M Group, H&M personalized fashion recommendations, Kaggle, 2022. https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations.

[45] C. K. Reddy, L. Márquez, F. Valero, N. Sharma, H. Z. Paka, Shopping queries dataset: A large-scale ESCI benchmark for improving product search, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2022.

[46] M. A. Ghossein, C.-W. Chen, J. Tang, Shopping queries image dataset (SQID): An image-enriched ESCI dataset for exploring multimodal learning in product search, arXiv preprint arXiv:2405.15190 (2024).