ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Label-Efficient and Adaptable Image Selection for Large-Scale E-Commerce Catalogs

Maytal Messing and Guy Shani

eBay - R&D Center

Abstract. E-commerce catalogs depend on high-quality, representative product images for user trust and engagement. Candidate images often include distracting backgrounds, irrelevant elements, or non-compliant branding or certification marks. We present a label-efficient, adaptable image selection pipeline combining: (i) prompt-guided zero-shot image cropping, which reduces non-product or branded elements and improves downstream quality predictions, and (ii) weakly supervised embedding-based outlier filtering that removes images inconsistent with the target product. Evaluated on high-impact categories, the approach adapts rapidly to new products, image sources, and catalog requirements while requiring no category-specific labels or retraining, providing a practical, scalable solution for industrial catalog curation.

1 Introduction

E-commerce C2C platforms such as eBay host vast inventories of heterogeneous, user-generated listings. To support product-based search and structured browsing, eBay maintains a curated product catalog in which each product (e.g., iPhone 16 Pro) is represented by a canonical image. This representative image shapes buyer trust, click-through behavior, and conversion [1, 2]. High-quality images should be sharp, well-framed, and free of clutter, while relevant images must depict the correct product model rather than accessories, packaging, or visually similar variants.

Producing such images at industrial scale is challenging. Manufacturer-provided photos are often copyright-restricted, so catalog images must be sourced from seller-uploaded pictures, partner inventories, and third-party authentication services. These sources exhibit substantial variability in lighting, framing, reflections, and background clutter, and at times include visually similar but incorrect product models. Moreover, a general marketplace spans a broad range of categories (unlike a single-domain platform such as fashion), each with its own visual conventions and evolving presentation guidelines. New image sources (e.g., from partners or acquired catalogs) continually introduce additional variability, making adaptability crucial. With millions of products and constantly changing content, manual curation is infeasible.

Existing e-commerce pipelines mitigate some of these issues via quality scoring, clustering, or category-level filtering [3]. However, many approaches assume relatively clean supplier images or rely on supervised training and per-category annotation. They struggle with fine-grained distinctions between closely related product models and rarely address anomaly detection under noisy, weakly

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964.
Available from http://www.i6doc.com/en/.

labeled conditions. Recent work on vision–language and foundation models offers new opportunities for data-efficient automation. Zero-shot object detectors [4, 5, 6] enable prompt-based localization without per-category training, and Vision Transformer [7] models provide strong generalizable representations for quality and relevance scoring.

Building on these advances, eBay developed the Image Selection Pipeline (iSel), a production-scale system that continuously ingests candidate images from multiple sources and selects a single catalog image per product. This paper focuses on two complementary modules that improve efficiency and robustness of this pipeline without requiring category-specific labels or retraining:

1. a zero-shot, prompt-guided object detector that automatically crops the main product region, and

2. an embedding-based outlier detector that filters mismatched or inconsistent images via cluster-level anomaly detection.

Together, these modules form a data-efficient and resilient foundation for large-scale catalog curation. They improve both overall image quality (e.g. better framing and clarity) and relevance to the intended product, while maintaining low operational overhead. This illustrates how foundation models can be operationalized in a noisy, real-world environment for industrial image selection at scale.

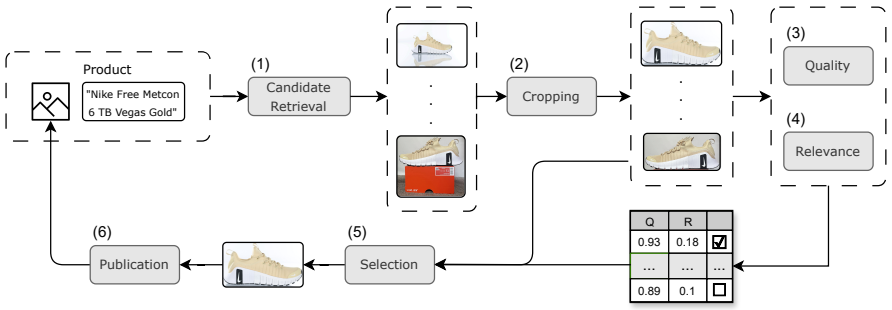

Fig. 1: High-level structure of the Image Selection Pipeline (iSel), which processes candidate images through modular filters and scorers before final ranking.

2 Approach

2.1 Pipeline Overview

The Image Selection Pipeline (iSel) retrieves candidate images from seller listings, partners, and authentication centers, and processes them through modular components before selecting a single representative image per product (Figure 1). We focus on two label-efficient modules designed for adaptability and long-term

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

maintainability: (i) zero-shot object detection for image cropping, used as a quality-improving preprocessing step, and (ii) embedding-based outlier removal for relevance. Both modules require only lightweight, category-specific configuration, and can be adapted quickly as catalog imagery standards, product types, or image sources evolve.

2.2 Zero-Shot Cropping

Candidate images often include reflections, slabs (protective cases for graded trading cards), accessories, or distracting backgrounds. We localize the main product region using Grounding DINO [6], a prompt-guided, zero-shot vision–language detector that requires no category-level training or annotated boxes. Given an image I and a set of textual prompt terms $\mathcal{W} = \{w_1, \ldots, w_m\}$, the model generates candidate boxes $\{b_i\}$ and assigns each box a similarity score for each prompt term. Boxes whose highest similarity score across all prompts exceeds a configurable `box_threshold` are retained. Predicted labels for each box are extracted from the prompt terms with similarity above a `text_threshold`.

We use the product-type name (e.g., "sneaker", "card", "cellphone") as the base prompt. When errors are observed in periodic evaluations, we manually refine prompts by (i) broadening descriptors ("shoe" instead of "sneaker" to better capture novel or atypical product shapes such as the Yeezy Foam Runner), and (ii) adding domain-specific distractor terms (e.g., "reflection", "plastic case") to help the model better contextualize its detections and distinguish the product from common confounders. We may retain the single highest-scoring box or combine multiple boxes (e.g., to preserve both shoes or front and back views), depending on catalog guidelines.

Initial Detection	Improved Detection	Final Crop
		

Table 1: Prompt refinement improves object localization in challenging cases.

For instance, reflective turntable images from a newly integrated authentication partner caused the model to confuse reflections for actual shoes. Updating the prompt to "shoe. reflection." eliminated false detections. This required no retraining or labeled boxes, only prompt and threshold adjustments, enabling fast, scalable adaptation to new image sources and emerging photographic styles.

2.3 Embedding-Based Outlier Detection

To filter mismatched models, accessories, or packaging, we learn a product-level embedding space using a Vision Transformer [7]. Training uses weak positive

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

links derived from an existing model that associates listings with catalog products, and hard negatives sampled from visually similar models, typically within the same brand or product line. To encourage separation, we also include easy negatives from unrelated brands or categories. This results in tight per-product clusters without requiring manual annotation.

At inference, only images that pass the quality model are embedded. For each product, we compute a centroid and discard outliers using a Modified Z-Score threshold, supplemented by lightweight brand and intra-cluster similarity checks.

3 Evaluation and Results

3.1 Evaluation Setup

Each quarter we sample ∼1000 listings per key category (**Sneakers**, **STC**, **Cellphones**) using impression-stratified sampling. Annotators map each listing to its catalog product and assess the selected image for visual quality, relevance, and whether the pipeline selected an image for the product.

To evaluate a new module, we perform pre–post analysis on the same listings: the pipeline is rerun with the updated component, all other modules remain fixed, and the newly selected images are re-annotated. For cropping we report quality precision and product coverage (i.e., the fraction of products with a selected image). for relevance filtering we report relevance of the selected image and corresponding coverage.

3.2 Impact of Cropping on Quality and Coverage

Cropping removes distracting regions and improves framing, enabling more candidate images to meet quality requirements. Figure 2 shows precision–coverage performance before and after cropping in **Sneakers** and **STC**.

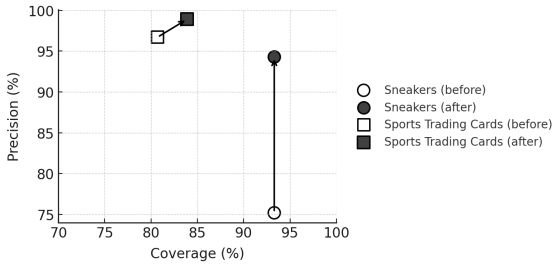

Fig. 2: Precision-coverage before and after cropping deployment.

In Sneakers, cropping increases precision from 75.2% to 94.3%, reflecting its ability to correct common framing and background issues while maintaining high coverage (93.3%). In STC, where baseline visual quality is already high, precision improves from 96.7% to 98.9%, and coverage increases from 80.7% to

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

83.9%. The latter is particularly important in this long-tail category, where many products have only a small number of seller-uploaded images.

These gains illustrate that prompt-guided, zero-shot cropping allows a higher fraction of surfaced images to meet quality standards without sacrificing, and in some cases improving, catalog coverage.

3.3 Relevance Filtering Performance

We compare two families of relevance methods: (1) LLM-based prompting and (2) embedding-based outlier detection. LLM-based approaches either judge whether an image matches a product name (LLM-N) or compare a listing image to a reference image (LLM-I). While they offer strong semantic reasoning, they incur latency and cost, and are less reliable for newly released products that may fall beyond the model's training data.

Embedding-based methods treat each product's candidate set as a visual cluster and identify outliers based on distance to the centroid. We evaluate four variants: basic outlier detection (OD), OD with brand matching constraint (OD+B), OD with noisy-cluster removal (OD+N), and the full pipeline (OD+B+N).

Table 2 reports relevance and coverage, split by product age. BASIC denotes no filtering beyond the upstream catalog mapping.

Method	**Relevance (%)**			**Coverage (%)**		
	All	New	Old	All	New	Old
BASIC	75.0	77.5	73.3	100	100	100
LLM-N	89.3	88.0	90.2	82.4	82.9	82.0
LLM-I	90.7	88.2	92.8	83.5	91.9	77.6
OD	90.2	96.2	85.9	90.4	93.7	88.2
OD+B	93.7	99.0	89.6	87.1	92.8	83.2
OD+N	92.7	99.0	88.2	86.0	88.3	84.5
OD+B+N	**95.6**	**100**	92.1	82.7	88.3	78.9

Table 2: Relevance and coverage for different filtering methods. OD+B+N combines brand constraints, noisy-cluster filtering, and embedding-based outlier detection.

The full embedding-based pipeline (OD+B+N) achieves the highest overall relevance (95.6%), perfect relevance for newer products (100%), and coverage comparable to LLM methods. This is crucial because newly released products tend to dominate user impressions. Compared to LLM-based filtering, the embedding approach scales better in production, avoids per-query prompt costs, and can be monitored using standard cluster-level diagnostics.

4 Discussion and Conclusion

We presented a production-scale catalog image selection pipeline that automatically identifies high-quality, relevant catalog images from heterogeneous sources. By combining zero-shot object detection for content-aware cropping

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

with embedding-based outlier filtering, the system adapts rapidly to new categories, image sources, and evolving catalog standards with minimal human intervention. This design provides strong operational scalability and maintainability, while delivering improved visual quality, coverage, and relevance across multiple high-impact categories.

Several limitations remain. Zero-shot detection performance depends on prompt quality and may require occasional updates as unique product shapes or imaging styles emerge. Future work includes leveraging LLMs to automate prompt generation and distractor discovery.

Overall, the proposed pipeline provides a robust and data-efficient framework for large-scale catalog image selection and can be extended as business needs evolve. Our experience demonstrates that foundation-model capabilities can be effectively integrated into industrial systems to achieve measurable improvements in e-commerce image quality and user experience, while keeping labeling and maintenance costs under control.

References

[1] Wei Di, Neel Sundaresan, Robinson Piramuthu, and Anurag Bhardwaj. Is a picture really worth a thousand words? -on the role of images in e-commerce. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 633–642, 2014.

[2] Stephen Zakrewsky, Kamelia Aryafar, and Ali Shokoufandeh. Item popularity prediction in e-commerce using image quality feature vectors. *CoRR*, abs/1605.03663, 2016.

[3] Abon Chaudhuri, Paolo Messina, Samrat Kokkula, Aditya Subramanian, Abhinandan Krishnan, Shreyansh Gandhi, Alessandro Magnani, and Venkatesh Kandaswamy. A smart system for selection of optimal product images in e-commerce. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1728–1736. IEEE, 2018.

[4] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European conference on computer vision*, pages 728–755. Springer, 2022.

[5] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024.

[6] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.