# Scalable Visual Attribute Recognition in E-Commerce Products via Automated Synthetic Label Generation

Victor Martinez[1], Virginia Negri[1], Brayan Impata[1] and Sergio Alvarez[1]

[1]*Amazon*

**Abstract**

E-commerce stores rely on product catalog data, which can be enriched by automated mechanisms like visual attribute extraction, for features like search and filtering. Extracting visual attributes from product images in e-commerce is challenging due to the wide diversity in products and the high cost of manual labeling, making traditional methods that rely on human-annotated data often impractical. In this work, we introduce a scalable approach that uses CLIP models to automatically generate synthetic labeled datasets, significantly reducing the need for manual annotation. To ensure dataset quality, we employ an ensemble of CLIP models for zero-shot classification and selectively incorporate text-based models for choosing high-accuracy examples, such as DeBERTa and BART-large pre-trained for natural language inference to perform zero-shot text classification. For downstream classification model training, a CLIP image encoder is used as the backbone, while only a single vector per class is fine-tuned using the automatically labeled data. We evaluate our models through human assessments, guided by automatically generated annotation guidelines based on high-accuracy examples. We evaluated our approach on a large-scale experimental dataset spanning over a thousand product types and over a hundred visual attributes. To ensure high-quality predictions, we computed rejection thresholds to achieve at least 90% accuracy per class and found that, on average, 60.8% of predictions exceeded the threshold. Our results demonstrate that synthetic labeling with vision-language models could serve as a practical and effective approach for large-scale visual attribute extraction in e-commerce product catalogs.

**Keywords**

E-commerce, Visual Attribute Extraction, CLIP Models, Synthetic Labeling, Zero-shot Classification, Product Catalog Data



## 1. Introduction

Modern e-commerce stores catalog millions of products across thousands of categories, each requiring detailed attribute annotations to power search, recommendation, and personalization systems. Visual attributes such as color, style, material, pattern, and design elements are fundamental to how customers discover and evaluate products online [1, 2]. For instance, a fashion retailer may need to classify footwear across dozens of attributes including heel height, toe shape, closure type, and material, with each attribute containing multiple possible values. An overview of this visual attribute extraction task is illustrated in Figure 1.

Despite its importance, the scale of modern e-commerce presents unprecedented challenges for visual attribute extraction [3]. Traditional supervised learning methods depend on manually labeled datasets [4, 5]. However, generating a sufficiently comprehensive dataset to cover thousands of product categories and hundreds of attributes would demand millions of human annotations, an effort that remains prohibitively expensive and time-consuming, even when using semi-supervised learning to reduce the annotation workload [6]. Furthermore, e-commerce product images exhibit extreme diversity in lighting conditions, camera angles, backgrounds, and image quality, making models trained on limited manually-curated datasets prone to poor generalization. The challenge is compounded by the rapid introduction of new products and attributes, requiring continuous model updates that traditional annotation pipelines cannot sustain.

To address these challenges, we introduce a scalable framework that leverages vision-language models for automated visual attribute extraction in e-commerce. Our key insight is that pretrained models like

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

vicmg@amazon.com (V. Martinez); vrgngr@amazon.com (V. Negri); biimpata@amazon.com (B. Impata); sbalanya@amazon.com (S. Alvarez)

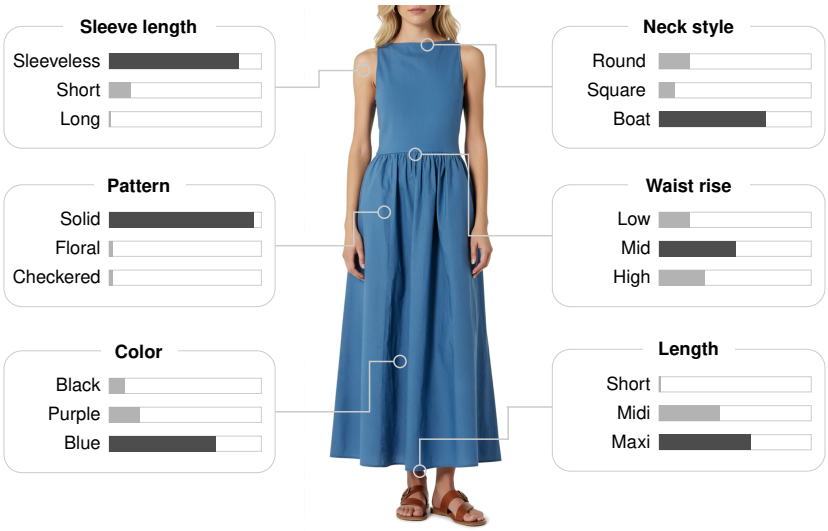


**Figure 1:** Example illustration of the visual attribute recognition task from product images. Given an input image, a model predicts scores (represented as bars) for multiple predefined attributes. Each attribute is associated with a set of candidate classes, and the model assigns confidence scores indicating the likelihood of each class. The class with the highest confidence (represented by the darkest bar) is selected as the final prediction for each attribute.

CLIP [7, 8] can be effectively adapted to generate high-quality training datasets with synthetic labels, eliminating the manual annotation bottleneck while maintaining classification accuracy. However, directly applying general-purpose zero-shot models to e-commerce data presents significant challenges: domain gaps between web-scraped training data and product catalogs, vocabulary mismatches between natural language descriptions and commercial attribute taxonomies, and the lack of reliable methods to assess prediction quality without expensive human validation at scale.

Our approach addresses these limitations through three main components. First, we employ an ensemble-based synthetic labeling strategy that combines multiple CLIP models with complementary text-based models to improve label quality and reliability. Second, we utilize a unified architecture where CLIP image encoders serve as frozen feature extractors while only lightweight, class-specific vectors are fine-tuned, enabling efficient multi-attribute prediction from a single forward pass. Third, we develop an automated annotation guideline generation system that uses high-quality model predictions as class examples to ensure consistent human assessment of model performance across diverse product categories.

We validate our approach through evaluation across diverse product-attribute combinations using human annotators guided by our automated guidelines. To assess scalability, we conduct experiments at volumes comparable to those found in real e-commerce catalogs, demonstrating that the framework can handle hundreds of millions of attribute predictions across diverse product categories with a low inference time per prediction.

The main contributions of this work are: (1) a scalable synthetic data generation pipeline using ensembles of vision-language models that eliminates manual annotation requirements, (2) a unified CLIP-based architecture for efficient multi-attribute classification that balances accuracy and computational efficiency, (3) an automated system for generating annotation guidelines that enables consistent human evaluation at scale, and (4) an experimental validation of scalability through large-scale evaluations across diverse product categories. Our results show that synthetic labeling with pretrained models can serve as a practical alternative to traditional supervised learning using human-labeled datasets for large-scale visual attribute extraction in e-commerce. The scope of this work is limited to the offline experimental evaluation of the proposed approach. Translating these findings into production-ready environments remains a task for future research.

## 2. Related work

The task of visual attribute extraction in e-commerce has received limited attention within the computer vision community, despite its critical role for e-commerce.

Traditional supervised approaches for visual attribute extraction have primarily relied on manually annotated datasets combined with convolutional neural networks to classify product attributes [9, 4, 5]. While these methods established foundational techniques for understanding visual features, they require extensive human annotation efforts that become prohibitively expensive when scaling to thousands of product categories and hundreds of attributes across millions of products. Semi-supervised learning has been explored as a way to alleviate this annotation bottleneck by leveraging large amounts of unlabeled data alongside a smaller set of labeled examples [6, 10]. However, while it reduces the dependency on exhaustive manual labeling, it does not entirely eliminate the need for high-quality annotations and still faces challenges in generalizing across diverse product domains and attribute types.

Zero-shot learning with vision-language models has emerged as a promising direction for attribute classification without manual annotation. CLIP and similar models using contrastive learning to align visual and textual representations [7, 11, 12] enable inference of visual attributes from natural language descriptions. These approaches have demonstrated effectiveness in extracting visual features from product images [13, 14], though their application to large-scale e-commerce taxonomies with domain-specific attribute vocabularies remains underexplored.

Simultaneously, synthetic data generation has emerged as a viable strategy for building and expanding datasets. Researchers have demonstrated the feasibility of constructing large-scale annotated datasets by automatically generating labels using pretrained models [15, 16], eliminating much of the manual overhead typically required. Synthetic labeling addresses the issue of dataset volume and allows for continuous updates in dynamic e-commerce environments, where product catalogs are frequently revised.

In addition, ensemble methods have been applied to improve prediction robustness and quality in various classification tasks. Combining predictions from multiple models [17, 18, 19, 20] can reduce prediction variance and improve reliability. However, ensemble approaches typically require multiple inference passes, which can be computationally expensive for systems processing millions of predictions.

Complementing this, multi-modal fusion approaches [21, 22, 23] combine visual and textual information to improve attribute classification accuracy. By leveraging complementary strengths of both modalities, these methods use visual models to capture image features while text models extract information from product descriptions. The integration of multiple modalities can improve label quality, though consensus-based filtering may reduce dataset volume.

Our work addresses the specific challenges of large scale visual attribute extraction in e-commerce through a unified architecture that processes multiple attributes simultaneously while maintaining computational efficiency. Rather than training separate models per attribute, our approach leverages frozen CLIP image encoders with learnable class vectors to enable multi-attribute extraction in a single forward pass. We focus on ensemble-based synthetic labeling combined with automated quality assurance methods to evaluate scalability and reliability under conditions representative of large-scale e-commerce datasets.

## 3. Methodology

Our approach frames visual attribute extraction as an image classification problem and focuses on generating high-quality synthetic labeled data to overcome the limitations of manual annotation. The complete pipeline consists of four main components: synthetic label generation using CLIP ensembles, unified model training with frozen image encoders, quality assurance through confidence-based rejection thresholds, and human evaluation supported by automatically generated annotation guidelines. An overview of our approach is presented in Figure 2.

Our work focuses exclusively on *visual* attributes—those whose values can be determined from

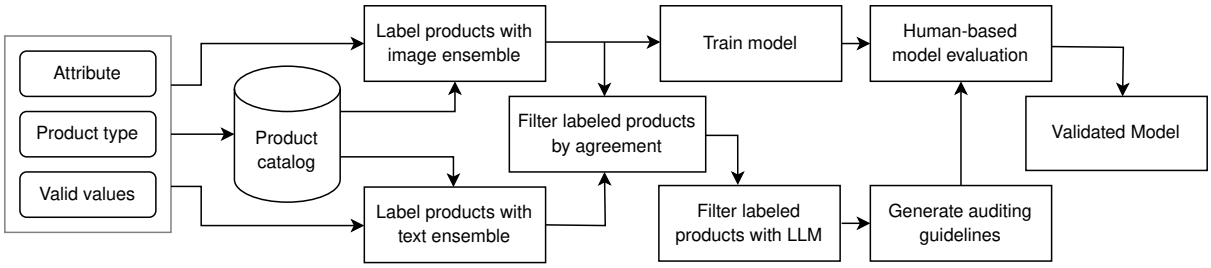


**Figure 2:** Overview of the proposed system for scalable visual attribute extraction. The pipeline consists of synthetic labeled dataset generation using CLIP-based zero-shot classification and text model ensembles, followed by training an image classifier with the generated labels. The final model is validated using human annotations assisted by automatically generated annotation guidelines. The diagram illustrates how the trained model can be applied to analyze and extract attributes from catalog data.

product images alone. Non-visual attributes (e.g., connectivity standards, compatibility specifications, or brand identity without visible logos) are out of scope, as they require textual or structured data sources rather than image-based classification.

We operate at the level of product type-attribute pairs, where both product types and attributes are predefined within a given taxonomy. For each product type-attribute combination, we train a model to classify the attribute using a fixed set of classes given by a predefined taxonomy. For example, for the product type *sunglasses* and attribute *age range*, the taxonomy specifies valid values such as *baby*, *kid*, *teen*, and *adult*. These valid values serve as the target output classes for our visual models.

### 3.1. Synthetic Labeling using CLIP Ensembles

We start by collecting a random sample of unlabeled product images from the overall product catalog to create the training, validation, and test datasets. Our product images typically showcase the product against a plain white background and occupying the majority of the frame. To ensure the data reflects different geographic regions, we sample images evenly from a variety of store locations. For each specific combination of product type and attribute we want the model to learn (such as shoe style or color), we calculate the sample size by multiplying the number of predefined classes for that attribute by 10,000. This helps ensure that the dataset is large enough to support accurate model training.

To overcome the limitations of human annotation, we use a set of CLIP models from OpenCLIP [24] pretrained on different datasets to automatically assign attribute labels via zero-shot classification (the list of models can be found in Table 1). Specifically, we employ a CLIP-based ensemble to generate synthetic labels, which are used for both model training and rejection thresholds computation. Each model in the ensemble performs zero-shot classification using the prompt: "A picture of a *{class}* *{attribute}* *{product_type}*", where *{product_type}* represents the product type, *{attribute}* denotes the target attribute, and *{class}* refers to a candidate class. This process is repeated for each class, with the class exhibiting the highest similarity to the unlabeled image being selected as the model's prediction. To combine predictions across the ensemble, we aggregate the softmax probabilities from all models. For a given class $c \in C$, where $C$ is the set of all valid values defined for the product type-attribute combination, the ensemble probability is computed as:

$$p_{\text{ensemble},c} = \frac{1}{|M|} \sum_{m=1}^{|M|} \frac{\exp(\tau_m \cdot \langle v_m, t_{m,c} \rangle)}{\sum_{c'=1}^{C} \exp(\tau_m \cdot \langle v_m, t_{m,c'} \rangle)},$$

where $v_m$ is the normalized image embedding and $t_{m,c}$ the normalized text embedding for class $c$, both from model $m$. The temperature hyperparameter $\tau_m$ is always set to $100.0$, as indicated in OpenCLIP's documentation, to control the sharpness of the similarity scores when passed through the softmax function. The class with the highest combined ensemble probability is selected as the predicted class. We categorize predictions into two confidence levels: high-confidence predictions,

where a majority of models agree on the predicted class, and low-confidence predictions, where this criterion is not met. In practice, low-confidence predictions are treated as inconclusive labels. These are excluded from training to maintain label quality, but are later considered during validation to mitigate evaluation bias.

### 3.2. Unified Model Architecture and Training

The synthetic dataset is split into training, validation, and test splits with a 70/15/15 distribution. The training split is then used to train an image classification model. We use the image encoder of the *convnext_xxlarge/laion2B-s34B-b82K-augreg-soup* OpenCLIP model [25] as the backbone for this classifier, which has a good trade-off between size and performance, enabling efficient feature extraction while leveraging the strengths of pretrained representations. The image encoder from the CLIP architecture is kept frozen, while the class vectors are initialized using the text encoder embeddings generated from the same prompt used during synthetic label creation. These class vectors are subsequently fine-tuned using the synthetic data. We employ stochastic gradient descent (SGD) with momentum, using a batch size of 32, a learning rate of 0.01, a momentum of 0.9, and a weight decay of $1 \times 10^{-5}$. We use early stopping to halt training when the loss on the validation split does not improve for 5 consecutive epochs. This dual use of CLIP, first for zero-shot label generation and then for model training, creates a cohesive and streamlined pipeline for visual attribute extraction. While using the same model family for both labeling and training could propagate systematic biases, this risk is mitigated by three factors: (i) the ensemble employs six architecturally diverse models trained on different datasets, providing decorrelation across label sources; (ii) the text-based validation models (DeBERTa, BART-large, Claude) used for guideline generation are entirely independent of the CLIP pipeline; and (iii) the human evaluation step explicitly validates accuracy on real data, catching systematic errors that survive synthetic evaluation. Furthermore, a key advantage of our unified architecture is computational efficiency: the frozen image encoder computes visual embeddings once per image, while multiple attributes can be extracted simultaneously through lightweight vector operations with different class vectors. This design significantly reduces inference time compared to both the ensemble used for synthetic labeling, which requires multiple model evaluations, and individually fine-tuned models, which would require separate forward passes for each attribute.

### 3.3. Quality Assurance via Rejection Thresholds

As a quality assurance measure, we apply rejection thresholds to filter out low-confidence predictions that are more likely to be incorrect. These thresholds are calculated independently for each product type–attribute–class combination on the validation split to meet a specified target accuracy, enabling class-specific rejection based on predicted confidence scores. This per-class granularity is important because the same class (e.g., "leather") may exhibit different visual characteristics across product types (e.g., shoes vs. bags), requiring different confidence thresholds. This involves evaluating the model's accuracy while progressively discarding predictions with confidence below different thresholds. In particular, we sort the validation samples for each class by their predicted confidence scores and then calculate the accuracy after removing predictions with confidence below each threshold. The final threshold is the lowest confidence value where accuracy meets or exceeds the predefined target. Unlabeled images, which have low-confidence synthetic labels due to disagreement within the ensemble, are intentionally included to avoid bias. These images are considered inherently difficult to classify and are counted as errors in evaluation metrics, regardless of the model's prediction.

This process yields a trained model and class-specific rejection thresholds designed to ensure at least a predefined accuracy, as estimated from synthetic validation labels. To mitigate potential biases shared between training and evaluation sets that may inflate performance estimates, we include human-based evaluation. For each class, a targeted manual review verifies that the model meets the accuracy target as guardrail mechanism. This approach ensures a reliable model accuracy while remaining cost-effective, as it requires only a small, representative sample of predictions rather than a fully human-labeled

training set.

### 3.4. Human Evaluation with Automated Annotation Guidelines

To obtain a reliable measure of model performance, we conducted a human evaluation process guided by automated annotation guidelines. To support human annotators, we provide annotation guidelines that define valid classes and include visual examples. To reduce manual effort, we developed an automated system that selects high-confidence, synthetically labeled images as reference examples. To enhance label precision, we use a multimodal approach that combines image-based model outputs with text-based classifiers applied to available catalog descriptions of the product. Specifically, we use DeBERTa [26] and BART-large [27] models pre-trained for natural language inference to perform zero-shot text classification. Product text fields are concatenated, and class labels are framed as "a *{class} {attribute} {product_type}*". Only samples where both modalities agree, based on a consensus criterion, are retained. As a final step, the Claude 3.5 Haiku [28] large language model uses a binary-answer prompting strategy to validate label correctness based on the catalog descriptions of the product. Only LLM-confirmed examples are included in the guidelines. This cascaded approach maintains high label quality while minimizing LLM usage. For our task, a maximum of five images per class is sufficient, and this approach consistently yields them at scale in our experiments. An example of automatically generated annotation guidelines is shown in Figure 3.

For human annotation, we rely on trained human annotators. Each reviewed example from the test split is independently assessed by three annotators, with the final label determined by majority vote. If no consensus is reached, the example is left unlabeled and treated as an incorrect model prediction during evaluation, consistent with our conservative approach previously described above.

## 4. Results and Discussion

This section presents a comprehensive evaluation of the proposed framework across three key dimensions: component-wise performance analysis, end-to-end system evaluation, and large-scale evaluation.

### 4.1. Component-wise Performance Analysis

To understand the effectiveness of individual components within our framework, we conducted targeted evaluations of each key element in the pipeline. This analysis examines the performance of CLIP ensemble methods for synthetic labeling, the multimodal approach for generating high-quality annotation guidelines, and the overall system's ability to maintain quality through conservative filtering strategies. These component-level insights provide crucial understanding of the trade-offs and design decisions.

#### 4.1.1. CLIP Ensemble Performance

The foundation of our synthetic labeling approach relies on ensemble-based zero-shot classification. We evaluated the performance of six zero-shot CLIP models from OpenCLIP [24] compared to their ensemble for zero-shot synthetic image labeling, as shown in Table 1. The evaluation spanned 20 manually annotated datasets at the product type-attribute level and over 9,700 total samples, covering combinations such as *dress occasion*, *table shape*, and *shoe heel type*, among others. Labels were curated using the three-auditor method with consensus described above using manually crafted guidelines. Results show that pretrained vision-language models can act as effective proxies for manual labeling without task-specific fine-tuning, but their standalone performance is often insufficient. In contrast, model ensembling substantially boosts accuracy, especially when low-confidence predictions are excluded due to a lack of majority agreement among ensemble members.

The reported accuracies should be interpreted in the context of the inherent complexity and ambiguity of the task, where many instances admit multiple valid interpretations (e.g., predicting multicolor versus the most dominant color). Additionally, the fine-grained nature of the problem, where some classes

| Class | Example images |
|---|---|
| gladiator |  |
| slide |  |
| slingback |  |
| stiletto |  |
| wedge |  |

**Figure 3:** Example of automatically generated annotation guidelines for *sandal style*.

differ based on subtle visual details (e.g., classifying knee-length versus midi for garments with hemlines just below the knee), poses challenges not only for generalist models but also for human annotators.

Several key insights emerge from these results. First, individual model performance varies considerably, with the performances spanning from 63.5% to 70.2%. This variation likely reflects differences in training data, pretraining strategy and architecture, which is precisely what makes these models complementary within an ensemble — each captures different aspects of visual representations. Second, the ensemble approach provides an improvement over the best individual model, demonstrating the effectiveness of model diversity in capturing complementary visual representations. Most importantly, the confidence-based filtering mechanism proves highly effective, boosting ensemble accuracy from 72.4% to 78.4% by discarding 18.4% of predictions where models disagree in our experiments. This trade-off between coverage (i.e., the proportion of data labeled) and accuracy is crucial for maintaining high-quality synthetic labels while preserving sufficient training data volume.

**Table 1**
Performance for individual zero-shot models and their ensemble.

| Models | Accuracy |
|---|---|
| ViT-bigG-14/Laion2B-39B-b160k [24, 29] | 63.5% |
| ConvNeXt-XXLarge/Laion2B-s34B-b82K [25] | 65.3% |
| ViT-bigG-14-CLIPA-336/DC-1B [30, 31] | 67.7% |
| ViT-bigG-14-CLIPA/DC-1B [30, 31] | 67.9% |
| ViT-SO400M-14-SigLIP-384/WebLI [32] | 70.1% |
| ViT-L-16-SigLIP-256/WebLI [32] | 70.2% |
| Ensemble | 72.4% |
| Ensemble (high confidence only) | 78.4% |

#### 4.1.2. Annotation Guideline Generation Quality

For the automated auditing guideline creation process described in Section 3.4, text-based models were employed to synthesize annotations from product textual information, aiming to improve the accuracy of examples included in the guidelines. When combined with image-based predictions, a consensus-based strategy was applied, requiring agreement between both modalities, to curate a final dataset with highly reliable labels, achieving more than 97% correctness as reported in Table 2. We performed an evaluation on a dataset manually labeled by an experienced in-house annotator, including the *length*, *sleeve type*, and *neck style* attributes for 1530 products, for which unstructured text attributes (i.e., item name, bullet points, and description) were retrieved. Although this dual-modality filtering significantly reduced the dataset size, resulting in less than 14% coverage under the most conservative configuration, it substantially improved the correctness of the resulting annotations. Since this subset still provided sufficient volume to sample five examples per class for the guidelines, we adopted this most conservative configuration in our pipeline.

**Table 2**
Trade-off between accuracy and coverage across image-text model combinations.

| Method | Accuracy | Coverage |
|---|---|---|
| Text models (DeBERTa and BART-large) | 67.8% | 40.3% |
| CLIP ensemble and text models | 91.9% | 18.0% |
| CLIP ensemble, text models and Claude 3.5 Haiku | 97.6% | 13.7% |

The multimodal approach reveals a clear progression in the accuracy-coverage trade-off. Text-only models (DeBERTa and BART-large) provide moderate accuracy (67.8%) with reasonable coverage (40.3%), demonstrating that product descriptions contain valuable attribute information. However, the combination of CLIP ensemble with text models dramatically improves accuracy to 91.9% at the cost of reduced coverage (18.0%). The addition of Claude 3.5 Haiku as a final validation step pushes accuracy to 97.6% with only marginal coverage reduction to 13.7%. This cascaded filtering approach proves essential for guideline generation, where high-quality examples are more valuable than quantity. The final 13.7% coverage still provides sufficient examples for creating robust annotation guidelines, as our approach requires only five examples per class, while ensuring near-perfect label correctness for annotators.

### 4.2. End-to-End System Evaluation

To ensure that the performance of the trained visual classifiers aligns to a predefined target accuracy, which we set to 90% in our experiments, we developed a comprehensive model evaluation pipeline incorporating a human-based evaluation process. This pipeline included a manual evaluation step conducted by human annotators, guided by automatically generated annotation guidelines, where underperforming classes were disabled by setting an unreachable rejection threshold. The internal

manual assessment showed that this evaluation protocol consistently yielded an average model accuracy slightly above 90% for the enabled classes, based on 2,225 manually audited examples across 21 product type–attribute combinations. Summary statistics for this evaluation are provided in Table 3. These findings highlight the effectiveness of our framework in maintaining high label quality, even with synthetically generated datasets, while also demonstrating the need for limited human supervision to mitigate potential biases and disable classes accordingly.

**Table 3**
Summary of generated guidelines and human-based evaluation.

| Metric | Value |
|---|---|
| Evaluated product type-attribute combinations | 21 |
| Average guideline number of classes | 6.8 |
| Average guideline example correctness rate | 94.7% |
| Average percentage of enabled classes | 56.4% |
| Average model accuracy for enabled classes | 90.9% |

The human evaluation results reveal several critical insights about our framework's end-to-end performance. The 94.7% correctness rate for automatically generated guideline examples validates our multimodal filtering approach, demonstrating that synthetic labels can effectively support human annotation tasks. However, the selective class enabling (56.4% on average) highlights an important limitation: while our approach produces high-quality predictions for classes where it performs well, it necessarily excludes challenging cases to maintain the 90% accuracy threshold. From our analysis, disabled classes predominantly fall into three categories: (i) *semantically ambiguous classes*, where visual concepts overlap (e.g., distinguishing "platform" from other heel types); (ii) *visually subtle distinctions*, requiring fine-grained discrimination that zero-shot models struggle with (e.g., "knee-length" versus "midi"); and (iii) *taxonomy mismatches*, where products do not entirely fit predefined catalog categories. This conservative strategy proves essential for quality assurance, but it underscores the inherent difficulty of certain visual attribute classifications even with sophisticated synthetic labeling.

The consistent achievement of at least 90% accuracy across enabled classes, validated through human evaluation, demonstrates that our framework successfully bridges the gap between synthetic labeled training data and the target domain. This result is particularly significant given that the models were trained entirely on synthetically labeled data, suggesting that high-quality pseudo-labels can effectively substitute for manual annotation when combined with appropriate quality control mechanisms.

## 4.3. Large-Scale Experimentation

To evaluate the scalability and robustness of the proposed framework, we conducted a large-scale experiment involving millions of products. This experiment was carried out by developing an end-to-end automated pipeline for synthetic dataset generation, CLIP-based model training, and human-in-the-loop evaluation via automated guidelines. The goal was to assess the system's ability to scale across diverse product types and large data volumes, while maintaining efficiency, low inference time, and quality.

In our experiments, the system exhibited strong scalability and broad applicability across a wide range of product categories. In our experimental dataset, which comprised a randomized sample of over a billion attribute slots across multiple geographic regions, the system generated hundreds of millions of predictions covering thousands of unique product types, over a hundred distinct visual attributes, and several thousand unique product type–attribute combinations. This highlights both its scalability and generalizability. Despite processing data at such a large scale, the system maintained low inference time, keeping sub-second 99th percentile prediction times. On average, it generated predictions for 60.8% of entries per product type–attribute model, with the rest filtered out by previously defined rejection thresholds calibrated to ensure a minimum prediction accuracy of 90%. These results underscore the system's practical effectiveness and scalability in large-scale product catalogs.

The synthetic data labeling in our solution provides a highly efficient and scalable alternative to manual annotation. In contrast to human labeling, which often requires multiple auditors and incurs substantial per image costs, synthetic labeling reduces costs by several orders of magnitude. This significant cost reduction enables large scale training that would otherwise be economically unfeasible. For example, projects requiring millions of label annotations can be carried out at a fraction of the traditional cost, making large scale model development both feasible and cost effective.

Our scalability experiments highlight several key advantages of our approach. First, the massive scale of the experiment, generating hundreds of millions of predictions across thousands of unique product type–attribute combinations, demonstrates the framework's ability to handle e-commerce catalog diversity. The 60.8% prediction coverage indicates that our rejection thresholds effectively maintain quality while ensuring reasonable throughput. Second, the sub-100ms inference time per prediction is highly advantageous for potential large scale data processing. Third, the cost efficiency makes large scale application economically viable, enabling comprehensive catalog coverage that would otherwise be prohibitively expensive with manual annotation.

Overall, this large-scale experiment validates the scalability of our proposed framework. It shows that high-throughput, low inference time, and cost-efficient visual attribute extraction is feasible even across highly diverse and evolving e-commerce catalogs. Importantly, the experiment reveals that synthetic labeling not only enables economic scalability but also supports training for long-tail product types, substantially broadening the scope of automated metadata enrichment in complex domains.

### 4.4. Discussion and Limitations

While our results demonstrate the effectiveness of the proposed framework, several limitations and areas for improvement warrant discussion. The most significant limitation is the accuracy-coverage trade-off inherent in our conservative approach. The selective enabling of only 56.4% of classes on average, while ensuring high accuracy, means that nearly half of the attribute classes remain unaddressed. This limitation particularly affects rare or visually ambiguous attributes that are challenging for both synthetic labeling and human annotation.

The reliance on CLIP model quality represents another constraint. Our ensemble approach, while effective, is fundamentally limited by the capabilities of the underlying vision-language models. Attributes requiring fine-grained visual discrimination or cultural context may be poorly captured by models trained on web-scale data. Additionally, the framework's performance may degrade for product types or visual styles significantly different from the CLIP training distributions. While our uniform averaging strategy sets a robust baseline, exploring non-uniform weighting schemes, such as optimizing ensemble weights or leveraging model disagreement patterns, could potentially improve the accuracy-coverage trade-off and reduce rejection rates.

Our framework currently assumes that product text descriptions used for guideline generation are of reasonable quality. In practice, e-commerce text data can be noisy, containing keyword stuffing or inaccurate descriptions. However, the cascaded multi-model consensus approach inherently mitigates this: misleading text is unlikely to produce agreement across both image and text ensembles simultaneously, limiting the propagation of low-quality text signals into the pipeline.

The multimodal filtering approach, while highly effective for guideline generation, introduces coverage limitations that may not be suitable for all applications. The final 13.7% coverage, while sufficient for creating annotation guidelines, may be restrictive for applications requiring broader coverage of training examples.

Despite these limitations, our results demonstrate that the presented framework successfully addresses the fundamental challenge of scalable visual attribute extraction in e-commerce. The combination of ensemble-based synthetic data generation and human-based quality control, based on reviewing a small sample of predictions, creates a practical solution that balances accuracy, coverage, and cost-effectiveness. The positive outcomes of our large-scale experiments suggest broader viability of the approach and can serve as a basis for future research in automated e-commerce catalog enhancement.

## 5. Conclusions

This experimental study presents a practical framework for scalable visual attribute extraction in e-commerce using synthetic labeled data. The proposed framework successfully addresses the cost and scalability challenges of manual annotation by leveraging zero-shot classification and multimodal consensus strategies. Our approach demonstrates cost efficiency while maintaining high accuracy, enabling visual attribute extraction at scale that would be economically challenging with traditional manual annotation approaches.

The technical approach demonstrates effective application of existing methods to the e-commerce domain. Our ensemble-based strategy shows clear improvements over individual CLIP models, with confidence-based filtering proving valuable for maintaining label quality. The multimodal consensus approach, combining vision and language models with large language model validation, produces high-quality automated annotation guidelines. In addition, we demonstrate that models trained on synthetic labels can achieve high accuracy levels across diverse product categories in experimental settings. Finally, our scalability experiments demonstrate the practical viability of our approach across diverse product categories and stores. The system is capable of handling large volumes of predictions while maintaining acceptable inference time and accuracy under experimental testing conditions.

Our framework addresses key practical challenges in e-commerce computer vision: annotation costs, scalability, and quality control. By combining automated training data generation with guideline creation, our approach provides a complete pipeline for visual attribute extraction that can be applied to new product types and attributes with minimal manual intervention.

While this work demonstrates the viability of synthetic labeling for e-commerce, several directions for future research emerge from these findings. First, advancing ensemble techniques and confidence calibration could improve the accuracy-coverage trade-off, potentially enabling higher class coverage while maintaining quality standards. Second, incorporating domain-specific visual understanding through targeted fine-tuning of CLIP models could enhance performance on challenging attributes requiring fine-grained discrimination. Third, developing adaptive filtering mechanisms that adjust confidence thresholds based on attribute difficulty could optimize coverage for different use cases. Fourth, extending the multimodal consensus approach to incorporate additional modalities, such as structured product specifications or user-generated content, could further improve label quality. Finally, the methodology's applicability to other domains beyond e-commerce, such as medical imaging or autonomous driving, warrants investigation.

The successful application of synthetic labeling in this domain provides insights for similar computer vision challenges. As foundation models continue to improve, synthetic data approaches may become increasingly viable for training specialized vision models in domains where manual annotation is expensive or impractical.

## Declaration on Generative AI

During the preparation of this work, the authors used Claude Sonnet 4.5 for grammar and spelling refinement. Authors also utilized generative AI technologies for visual asset creation, specifically to generate representative product imagery that preserves brand anonymity and avoids the use of actual branded materials in this manuscript. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

## References

[1] T. E. Mofokeng, The impact of online shopping attributes on customer satisfaction and loyalty: Moderating effects of e-commerce experience, Cogent Business & Management 8 (2021) 1968206. URL: https://doi.org/10.1080/23311975.2021.1968206. doi:10.1080/23311975.2021.1968206. arXiv:https://doi.org/10.1080/23311975.2021.1968206.

[2] M. Niemir, B. Mrugalska, Product data quality in e-commerce: Key success factors and challenges, Production Management and Process Control 36 (2022) 1–12.
[3] K. Pham, K. Kafle, Z. Lin, Z. Ding, S. Cohen, Q. Tran, A. Shrivastava, Learning to predict visual attributes in the wild, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13013–13023. doi:10.1109/CVPR46437.2021.01282.
[4] P. Gutierrez, P.-A. Sondag, P. Butkovic, M. Lacy, J. Berges, F. Bertrand, A. Knudson, Deep learning for automated tagging of fashion images, in: Proceedings of the European conference on computer vision (ECCV) workshops, 2018.
[5] V. Parekh, K. Shaik, S. Biswas, M. Chelliah, Fine-grained visual attribute extraction from fashion wear, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3973–3977.
[6] M. Lagunas, B. Impata, V. Martinez, V. Fernandez, C. Georgakis, S. Braun, F. Bertrand, Transfer learning for fine-grained classification using semi-supervised learning and visual transformers, 2023. URL: https://arxiv.org/abs/2305.10018. arXiv:2305.10018.
[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR, 2021, pp. 8748–8763.
[8] S. Shen, L. H. Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, K. Keutzer, How much can clip benefit vision-and-language tasks?, arXiv preprint arXiv:2107.06383 (2021).
[9] A. Nodari, I. Gallo, M. Vanetti, Automatic visual attributes extraction from web offers images., in: CompIMAGE, 2012, pp. 127–132.
[10] B. Impata, M. L. Arto, V. M. Gomez, V. F. Arguedas, C. Georgakis, S. Braun, F. Bertrand, Semi-supervised learning and visual transformers for product attribute extraction from e-commerce images (2024). URL: https://www.amazon.science/publications/semi-supervised-learning-and-visual-transformers-for-product-attribute-extraction-from-e-commerce-images.
[11] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, T. Duerig, Scaling up visual and vision-language representation learning with noisy text supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 4904–4916. URL: https://proceedings.mlr.press/v139/jia21b.html.
[12] Q. Qian, J. Hu, Online zero-shot classification with clip, in: European Conference on Computer Vision, Springer, 2024, pp. 462–477.
[13] J. Gong, M. Cheng, H. Shen, P.-Y. Vandenbussche, J. Jenq, H. Eldardiry, Visual zero-shot e-commerce product attribute value extraction, arXiv preprint arXiv:2502.15979 (2025).
[14] M. Hendriksen, M. Bleeker, S. Vakulenko, N. Van Noord, E. Kuiper, M. De Rijke, Extending clip for category-to-image retrieval in e-commerce, in: European Conference on Information Retrieval, Springer, 2022, pp. 289–303.
[15] M. Mendikowski, M. Hartwig, Creating customers that never existed: Synthesis of e-commerce data using CTGAN, in: 18th International Conference on Machine Learning and Data Mining (MLDM-22). New York, US: IBAI Publishing, 2022, pp. 91–105.
[16] M. Gabryel, E. Kocić, M. Kocić, Z. Patora-Wysocka, M. Xiao, M. Pawlak, Accelerating user profiling in e-commerce using conditional GAN networks for synthetic data generation, Journal of Artificial Intelligence and Soft Computing Research 14 (2024) 309–319.
[17] G. Valentini, F. Masulli, Ensembles of learning machines, in: Neural Nets: 13th Italian Workshop on Neural Nets, WIRN VIETRI 2002 Vietri sul Mare, Italy, May 30–June 1, 2002 Revised Papers 13, Springer, 2002, pp. 3–20.
[18] V. Manian, E. Alfaro-Mejía, R. P. Tokars, Hyperspectral image labeling and classification using an ensemble semi-supervised machine learning approach, Sensors 22 (2022) 1623.
[19] J. Lu, Optimizing e-commerce with multi-objective recommendations using ensemble learning, in: 2024 4th International Conference on Computer Systems (ICCS), IEEE, 2024, pp. 167–171.
[20] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, P. N. Suganthan, Ensemble deep learning: A review, Engineering Applications of Artificial Intelligence 115 (2022) 105151.

[21] G. Li, N. Li, Customs classification for cross-border e-commerce based on text-image adaptive convolutional neural network, Electronic Commerce Research 19 (2019) 779–800.
[22] A. De la Comble, A. Dutt, P. Montalvo, A. Salah, Multi-modal attribute extraction for e-commerce, arXiv preprint arXiv:2203.03441 (2022).
[23] O. Amel, S. Stassin, S. A. Mahmoudi, X. Siebert, Multimodal approach for harmonized system code prediction, arXiv preprint arXiv:2406.04349 (2024).
[24] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, L. Schmidt, OpenCLIP, 2021. URL: https://doi.org/10.5281/zenodo.5143773. doi:10.5281/zenodo.5143773.
[25] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
[26] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing, arXiv preprint arXiv:2111.09543 (2021).
[27] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461 (2019).
[28] Anthropic, Claude 3.5 Haiku, Large Language Model, 2026. URL: https://www.anthropic.com/claude, accessed: June 15, 2026.
[29] C. Schuhmann, R. Beaumont, R. Vencu, C. W. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. R. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, J. Jitsev, LAION-5b: An open large-scale dataset for training next generation image-text models, in: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2022. URL: https://openreview.net/forum?id=M3Y74vmsMcY.
[30] X. Li, Z. Wang, C. Xie, An inverse scaling law for CLIP training, in: NeurIPS, 2023.
[31] X. Li, Z. Wang, C. Xie, CLIPA-v2: Scaling CLIP training with 81.1% zero-shot ImageNet accuracy within a $10,000 budget; an extra $4,000 unlocks 81.8% accuracy, arXiv preprint arXiv:2306.15658 (2023).
[32] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, arXiv preprint arXiv:2303.15343 (2023).

# A. Appendix: Dataset Specifications

This appendix provides comprehensive details about the datasets used throughout our study. The information is organized into three main subsections corresponding to different phases of our evaluation methodology.

## A.1. Image Model Evaluation Datasets

Table 4 presents the datasets used to evaluate CLIP models and ensemble performance for synthetic label generation (Section 4.1.1). These datasets span 20 diverse product type-attribute combinations, totaling 9,776 manually annotated examples. The annotation process was carried out by trained annotators. Each image was independently assessed by three annotators, with final labels determined by majority vote to ensure annotation quality. This multi-annotator approach addresses the inherent subjectivity in visual attribute classification while maintaining consistent labeling standards across different product categories.

The dataset composition reflects the diversity of e-commerce product catalogs, covering fashion items (such as tops, dresses, and shoes), home goods (such as tables, pillows, and rugs), and specialized products (such as wall art and electronics). The sample sizes for each category, ranging from 237 to 850 examples, were selected to be proportional to the number of classes, with approximately 50 examples per class.

**Table 4**
Datasets used for image models evaluation.

| Product Type | Attribute | # Examples |
|---|---|---|
| tops and dresses | sleeve type | 349 |
| dress | occasion | 573 |
| fashion products | pattern | 528 |
| handbag | silhouette | 800 |
| dresses and skirts | length | 241 |
| shoes | closure type | 350 |
| wall art | form | 648 |
| shoes | toe style | 237 |
| table | shape | 300 |
| hat | form type | 650 |
| boot | form type | 650 |
| shoes | height map | 650 |
| pillow | shape | 300 |
| shoes | heel type | 600 |
| ring | form type | 450 |
| table | base type | 450 |
| electric fan | design | 500 |
| sofa | type | 300 |
| fashion products | color | 850 |
| rug | shape | 350 |
| **Total** | | 9776 |

## A.2. Text-Image Integration Evaluation Datasets

Table 5 details the datasets used to evaluate text-image integration within the multimodal annotation guideline generation process (Section 4.1.2). These datasets were specifically designed to assess how DeBERTa, BART-large, and the LLM models complement visual information when performing zero-shot text classification on product descriptions, titles, and bullet points. The evaluation focused on three key fashion attributes: dress length, neck style, and sleeve type, representing different levels of classification complexity where textual descriptions provide complementary information to visual analysis.

Unlike the image evaluation datasets, these examples were labeled by a seasoned in-house expert. This approach was chosen due to the specialized nature of text-based attribute extraction and the need for consistent interpretation of product descriptions. The constrained dataset size (1,530 total examples) reflects the intensive manual effort required for expert annotation, but provides high-quality ground truth for evaluating text model performance in the context of guideline generation.

**Table 5**
Datasets used for text-image integration evaluation in the multimodal annotation guideline generation process.

| Product Type | Attribute | # Examples |
|---|---|---|
| dress | length | 316 |
| tops and dresses | neck style | 1064 |
| tops and dresses | sleeve type | 150 |
| **Total** | | 1530 |

## A.3. End-to-end Performance Evaluation Datasets

Table 6 presents the datasets used for the final human-based evaluation phase, where trained models are assessed against the 90% accuracy target using automatically generated annotation guidelines

(Section 4.2). This evaluation encompasses 21 diverse product type-attribute combinations totaling 2,225 manually reviewed examples, representing the most comprehensive validation phase of our methodology.

The annotation process for this phase employed four in-house experts, each with at least one year of experience in visual attribute classification, supplemented by consultation with subject matter experts when needed. This expert-based approach was chosen to ensure high-quality ground truth labels for the final model validation phase. The dataset covers a broad range of attributes including strap types, style classifications, form factors, and mounting types across various product categories from fashion to electronics and home goods. The sample sizes per product type-attribute combination (ranging from 50 to 375 examples) are proportional to the number of classes, with 25 examples per enabled class, while considering the practical constraints of expert annotation.

**Table 6**
Datasets used for end-to-end model performance evaluation guided by automatically generated annotation guidelines.

| Product Type | Attribute | # Examples |
|---|---|---|
| vacuum cleaner | form factor | 125 |
| dishware plate | style | 100 |
| digital camera | mounting type | 375 |
| shirt | style | 125 |
| clock | mounting type | 50 |
| dryer | form factor | 50 |
| keyboards | style | 100 |
| cellular phone case | form factor | 100 |
| dress | strap type | 100 |
| irrigation sprinkler | style | 100 |
| handbag | strap type | 75 |
| suitcase | strap type | 75 |
| electric light | style | 100 |
| backpack | strap type | 100 |
| GPS and navigation systems | mounting type | 75 |
| sandal | style | 100 |
| shoes | strap type | 75 |
| chair | form factor | 100 |
| window shade | mounting type | 50 |
| portable electronic device stand | form factor | 150 |
| portable electronic device stand | mounting type | 100 |
| **Total** | | 2225 |

## B. Appendix: LLM-Based Correctness Validator

This subsection provides technical details about the Large Language Model (LLM) used as the last filter to validate the correctness of the generated training examples. We used Anthropic's Claude 3.5 Haiku model with the following prompt:

> *Correctly answer the following question using only information from the product details. Always try to correctly answer but if there is no answer in the product details say "I don't know". Provide a yes or no answer, no additional text, or answer I don't know when you don't know.*
>
> *Question: Do the product details indicate that the* ***{attribute}*** *of the* ***{product type}*** *is* ***{class}****? Options: yes, no, I don't know*
>
> *Product details:*

*{product details}*

## C. Appendix: Example Predictions

This appendix provides additional qualitative context through a set of visual examples. Figure 4 presents a range of model predictions across different product categories and attributes. The examples illustrate the diversity of visual inputs and attribute types encountered during example labeling, offering insight into the variety of scenarios in which predictions were made.


**Figure 4:** Examples of correct and incorrect classifications across different product categories and attributes. Green labels indicate correct predictions and red labels indicate misclassifications (ground-truth in parentheses).