https://doi.org/10.2352/EI.2026.38.12.GENAI-177

Shelfie: A vision-language system for structured product understanding in real-world consumer environments

Ankur Purwar, Alexander Hollingworth, Laveena Satwani, Eu Jack Tan, Nandita Mishra and Pranav Mishra

Abstract

Personalized consumer experiences increasingly depend on understanding actual product usage in everyday settings. We present Shelfie, a consumer-centric vision-language system that extracts structured product metadata from user-submitted images of lifestyle care products arranged in real-world contexts – such as consumer shelves, countertops, and vanity spaces. Unlike conventional systems developed for controlled retail environments and dependent on barcode scanning, Shelfie is intentionally designed to operate effectively in cluttered, unconstrained home settings and is fully barcode-independent. Shelfie integrates object detection, instance segmentation and large language model (LLM)-based reasoning to infer rich metadata for each visible product. This includes product brand, name, category, form, package type, key ingredients, benefits, and size. The Shelfie system is trained and validated on a diverse user-sourced dataset covering personal care to home lifestyle products, demonstrating strong generalization in producing high-accuracy highly structured output across packaging styles and product categories. Shelfie establishes a vision-language foundation for real-world consumer-facing product understanding and discovery systems. It can enable downstream applications such as community-driven recommendation systems, ingredient sensitivity tracking, and in-depth consumer behavior analysis all while keeping consumer habits, needs and convenience at the center. By bridging visual input with structured metadata output, Shelfie can enable more informed, personalized decisions through peer-driven insights.

Introduction

Consumers tend to seek advice on products from retail staff, but the recommendations will generally be made based on point-of-sale data and barcode-driven analytics. This method, however, has not been adequate in an environment characterized by the abundance theory, where consumers are overwhelmed by numerous options as a result of the huge and constantly growing number of lifestyle care products, including home cleaning agents, air fresheners, laundry detergents, and personal hygiene items. The fast-paced and dynamic nature of this market makes it challenging to deliver intelligent advice on the sole basis of conventional retail tools. The continuous stream of product launches along with reformulated products and ever-changing buying habits of the consumer must be taken into account while providing recommendations. Traditional systems are unable to keep up with these factors and they also do not have insights into the user's utilization in their homes, resulting in a disconnect between the advice and the individual's taste. Moreover, the consumer must trust the worker to provide accurate recommendations based on their understanding of the consumer and their needs. This leaves room for misunderstandings and misinterpretations which often result in faulty recommendations. This lack of insight into in-home presence creates a blind spot for personalized retail strategies like product recommendations and analytics. For example, knowing what a user prefers in their skincare routine will help deliver tailored recommendations for them along with managing their allergies and sensitivities. Extracting structured product information from user-captured photos faces several technical challenges. These challenges include a cluttered environment, similar appearing products, variable lighting and partially visible products. These challenges are akin to those explored in scene text recognition systems [1], where models contend with distorted, occluded, and low-quality text. The integration of contextual and visual cues, as demonstrated in [1], motivates the development of systems that go beyond barcode-driven analysis. These challenges include a cluttered environment, similar appearing products, variable lighting conditions and more likely than not a partially visible product. Conventional barcode-based systems cannot handle these challenges as they are created with a controlled and clearly visible environment in mind. Our proposed solution Shelfie is an intelligence system that extracts rich structured data from user photos of lifestyle care products. Shelfie can identify and interpret each visible product of a given image and extract fields such as brand, product name, category, form, container type, ingredients, benefits, and size from it. Shelfie follows a unique approach making it independent of barcode lookup and can read and process product information in any language by utilizing LLM capabilities to reason over visual and textual cues present in the image. Shelfie enables identification of products in environments where traditional methods fail. It helps pave the way to unlock a wide range of consumer and brand facing applications like personalized recommendations and product usage analytics. By converting unstructured visual information to structured, machine-readable data, Shelfie offers a novel approach to retail understanding in everyday settings. Our specific contributions include: A pipeline that segments and identifies individual products from user shelf and countertop photos using a high-resolution object detection framework. A large language model-based architecture that directly infers structured metadata from products in any language without relying on barcodes. Extensive experiments demonstrating Shelfie's ability to extract fine-grained, structured product information from user submitted images, outperforming vision-only baselines in fieldlevel accuracy across multiple product types.

Related Work

Latest findings in multimodal learning have significantly shown improvement in product-level understanding for ecommerce and retail. The two key areas which can be related to Shelfie are cross-modal product retrieval and attribute-value extraction. Zhan et al. [4] presented Product1M, a massive weakly

supervised, instance-level product retrieval dataset using cross-modal contrastive learning. Their framework, CAPTURE, a hybrid-stream transformer, aligns fine-grained visual and textual representations from more than one million image-text pairs. CAPTURE shows strong retrieval using self-supervised learning on noisy e-commerce data with scalable instance-level matching without needing human-verified bounding boxes or attribute annotations. In counterpart to this, Zhu et al. [5] introduced a multimodal model for joint attribute prediction and value extraction from product listings in e-commerce. Their framework integrates structured understanding of products, predicting current attributes and value extraction from raw text. Through gated fusion-based incorporation of visual features, the model anchors attribute values to corresponding visual regions, enhancing robustness on noisy, incomplete, or semi-structured listings, particularly in categories with overlapping vocabularies of attributes. In contrast to these works, which depend on catalog-style inputs or structured listings, Shelfie addresses the novel domain of unstructured, user-generated product photos taken in everyday home environments. While Product1M and Zhu et al.'s methods excel in large-scale e-commerce contexts, they presuppose the availability of clean captions or structured attribute-value pairs. Shelfie, however, infers structured product metadata solely from in-the-wild images, without relying on OCR, barcodes, our textual supervision, thus pioneering a new direction for multimodal retail systems deployed in uncontrolled settings. Prior efforts in scene text recognition and document understanding provide important insights for Shelfie's vision-language design. Veit et al. [1] introduced an end-to-end system for scene text recognition that jointly models visual and linguistic context, improving robustness to distortions and occlusions. Xu et al. [3] further showed that real-world document understanding requires a multimodal approach that incorporates layout and spatial features beyond basic OCR pipelines. These ideas underpin Shelfie's multi-modal design, especially in cluttered, multilingual packaging scenarios. Vivino [2] is a well-established consumer application that demonstrates how user-submitted images can be transformed into structured metadata and personalized recommendations at scale. By allowing users to scan wine labels, Vivino aggregates contextual information such as expert and consumer ratings, tasting notes, food pairings, region, grape type, and winemaking styles. This information is enriched by a large and active community, enabling informed decision-making through a combination of collective intelligence and personalization. Vivino serves as a compelling example of how unstructured, user-generated visual data can be converted into meaningful, actionable insights within a narrowly defined and visually consistent product domain. The success of Vivino illustrates the value of guided discovery, community feedback, and preference-aware recommendation systems in consumer decision-making. Shelfie draws conceptual inspiration from this paradigm but addresses a fundamentally different and more complex problem space. Unlike wine bottles-which belong to a relatively homogeneous category with standardized labeling-Shelfie operates across diverse lifestyle and personal care categories, including home, fabric, and personal care products. These products are often captured in unconstrained environments with cluttered backgrounds, partial occlusions, non-standard packaging, and significant visual variability. Shelfie extends Vivino's underlying idea to a broader and noisier real-world setting by leveraging multimodal reasoning through large language models and prompt engineering. Shelfie focuses on deducing structured metadata directly from in-the-wild images, without relying on well-defined labels or category constraints. In this sense, Vivino represents an inspiration for what structured intelligence can enable at the consumer level, while Shelfie explores how similar outcomes can be achieved in far less structured and more visually complex domains.

Shelfie

Shelfie is a vision-language-based pipeline engineered to extract structured product metadata from user-submitted images of real-world environments. Shelfie does well at processing casual, cluttered photos taken in personal settings like bathroom shelves, kitchen counters, and laundry areas, where a traditional model often fails. The system integrates traditional computer vision components with a powerful LLM to perform end-to-end reasoning and generate structured outputs. The overall Shelfie pipeline can be visualized in Figure 1. The figure shows how a user generated image undergoes the process of product detection and segmentation to finally retrieve metadata for each individual product identified in the image. The pipeline visualization re-iterates Shelfie's capability to work in cluttered natural settings of a user's home. Here are the key technology building blocks for Shelfie pipeline -

Stock Image Classifier

Shelfie employs a dedicated classifier to detect and reject stock images, differentiating between authentic user-taken photos and high-quality catalogue product shots. If a stock image is detected, it is ignored. This step helps ensure that only real-world user content is processed by subsequent components.

Training Implementation - A binary image classification model was trained using Inception v3 to differentiate between stock and non-stock images on a custom dataset. Inception v3 was selected due to its strong representational ability, efficient use of parameters, and proven performance on large-scale visual recognition tasks. The model was fine-tuned on a custom dataset to adapt pretrained features to domain-specific characteristics such as composition, lighting, and background consistency typically seen in stock imagery. This approach enables robust discrimination between curated stock images and real-world images while maintaining good generalization performance.

Product Detector (Yolo V5)

Upon receiving an input image, Shelfie initiates preprocessing with product detection using a high-resolution object detection model. The detected bounding boxes are extracted as candidate product crops. These crops are then passed on for quality assessment. YOLOv5 was chosen for its strong balance between detection accuracy, inference speed, and deployment flexibility, making it well-suited for real-time and large-scale retail environments. Its architecture enables robust performance on small, densely packed objects (common in shelf imagery), while maintaining low latency for high-throughput pipelines.

Training Implementation - Product detection is implemented using YOLOv5 to identify individual products within shelf images. YOLOv5 was chosen for its consistent detection accuracy across diverse product categories, packaging styles, and shelf layouts, making it suitable for our use-case. The model effectively handles dense product arrangement and visual clutter, ensuring

correct localisation of products.

Image Quality Classifier

The product quality classifier filters out visually inadequate crops, such as those that are blurry or contain partial products. This ensures that only clearly visible and well-formed product crops proceed. This rigorous filtering minimizes false positives and ensures that only real-world user content is processed by subsequent components.

Training Implementation - A binary image quality classifier was trained using Inception v3 to distinguish between good and bad quality product images. Images exhibiting issues such as blur, poor focus, low resolution, occlusion, or lighting artifacts were labeled as bad quality, while clear and well-composed images were labeled as good quality.

Product Segmentation

Once high-quality crops are approved, Shelfie segments them to identify distinct product instances, particularly in images where multiple products are tightly clustered.

Training Implementation - Product segmentation is performed using the segmentation capabilities of YOLOv8 to precisely separate products from their surrounding background. This approach was chosen to achieve more accurate product isolation compared to bounding-box only methods, especially in scenarios where products are closely placed or partially overlapping. By generating pixel-level masks, the segmentation step enables cleaner product representations, which improves the reliability of downstream task of attribute extraction. This added precision helps reduce background noise, minimize false interpretations, and ensures more consistent results across different shelf layouts and packaging styles.

Product Object Rotation

Many product identification tasks rely on the recognition of text and graphical attributes, such as brand names, logos, and product descriptions. However, these attributes are often presented in varying orientations, making it difficult for the LLM to accurately extract the relevant information. Rotating the segmented product to a canonical orientation (e.g., upright) improves the accuracy of text recognition (OCR) and the extraction of graphical attributes..

GPT Vision API

The core of the Shelfie system lies in its integration with a large multimodal language model. Rather than relying on OCR to extract and interpret text, Shelfie directly feeds the cropped product image to the LLM. This approach is similar in philosophy to real-world document understanding [3], where multi-modal reasoning over layout and visual structure replaces rigid OCR-based pipelines. Shelfie leverages this capability to support multilingual and interpret packaging structures more robustly.

Attribute Extraction

In order to transform unstructured shelf and countertop images into rich, machine-readable data, Shelfie performs a structured attribute extraction step. For each detected product, the system identifies and extracts a standard set of attributes that together capture the product's identity, form, usage characteristics,

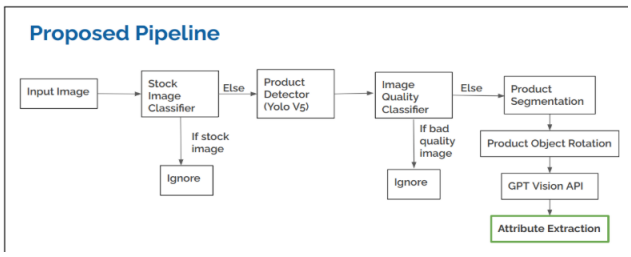


Figure 1. *Shelfie Pipeline*

and packaging details. Shelfie's attribute extraction covers the following fields:

Brand

The manufacturer or corporate name under which the product is marketed. Capturing the brand is essential for grouping products, analyzing brand loyalty, and aligning recommendations with user preferences.

Product Line

The definitive name of the product as labeled (excluding brand or line names), such as "Daily Moisturizer" or "Lavender Fabric Softener." This attribute provides the core identifier for search and retrieval tasks.

Product Name

The definitive name of the product as labeled (excluding brand or line names), such as "Daily Moisturizer" or "Lavender Fabric Softener." This attribute provides the core identifier for search and retrieval tasks.

Form

The physical form of the product (e.g., lotion, spray, powder, liquid). Knowing the form supports usage analytics (e.g., application method) and prioritizes packaging-appropriate recommendation logic.

Container Type

The style of packaging (e.g., bottle, tube, pouch, aerosol can, sachet). This attribute informs inventory management, logistics planning, and user convenience features such as refill reminders.

Container Primary Color

The dominant color of the container's exterior. Primary color can serve as a robust visual cue in low-text environments and aids in visual search and recognition when text is unreadable.

Benefit or Feature

Key selling points highlighted on the label (e.g., "moisturizing," "stain-fighting," "eco-friendly"). Extracting these succinct descriptors supports personalized recommendations based on user needs and product positioning.

Product Claims

Specific assertions or certifications such as "dermatologist tested," "paraben-free," or "USDA organic." Claims are critical for health- and value-oriented filtering, allergy avoidance, and

compliance monitoring.

Named Ingredients

A list of principal active or highlighted ingredients (e.g., "salicylic acid," "lavender oil," "sodium laureth sulfate"). Ingredient extraction enables allergy and sensitivity checks, as well as feature-based recommendations (e.g., "fragrance-free").

Size

The net quantity of the product (e.g., "250 mL," "16 oz," "500 g"). Size information is necessary for inventory tracking, consumption rate analytics, and cost-per-unit comparisons. By systematically extracting these attributes, Shelfie creates a comprehensive, structured profile for each product instance. This level of detail not only surpasses the capabilities of barcode-only systems but also empowers advanced analytics and personalized user experiences in dynamic, real-world environments.

Prompt Engineering for Shelfie

Shelfie's success can be attributed to a series of carefully engineered prompts which have proven to produce the most optimised output. The accuracy and consistency of the structured outputs is highly dependent on these prompts. Unlike normal image captioning tasks, Shelfie demands for a fine-grained level of extraction across a diverse set of products. Through repetitive experimentation, it was found that prompt structure, scope and inclusion of example-based conditioning significantly increased the quality of the output.

Attribute Scope and Prompt Splitting

Initially, a single comprehensive prompt was used to extract all the fields together, an approach that led to inconsistent results especially in the "product line" and "benefit or feature" fields. It was observed that larger prompts resulted in degraded outputs due to information overload. To avoid this, a multi prompt strategy was adopted, which involved querying error-prone fields in isolation or using dedicated prompts in order to improve the accuracy.

Category-wise Prompting

It was observed that a standard prompt across all categories of products was not effective due to the varying attribute meanings across categories like, "Smoothing" is a hair care benefit, while "Acne" is a skincare concern and on the other hand "10X tough stain removal" is a product claim for a cleaning product. The need to move from a universal prompt to a more specific prompt for each category was clear which also facilitated domain specific attributes.

Prompt Examples and Negative Conditioning

Incorporating example-based prompting further improved the results' accuracy as the model was taught to pick up subtle differences through both positive and negative examples. For example, the benefit or feature field was able to identify "Cleaning" for "Classic Clean Shampoo" and "Micro-sculpting" as a feature which in turn helped to reduce the false inclusions under "skin concern" and distinguish the marketing language from functional attributes.

Attribute Definition and Refinement

Defining what exactly to look for when extracting information and categorizing it into a particular attribute, improved the model's ability to distinguish between small nuances. For example, by clearly defining that the package type will be pump only if the product has the pump dispenser at the top as nozzle the model was able to distinguish between a pump and bottle, which was proving to be challenging earlier. The precise definition for each attribute helped to eliminate the ambiguity of the output.

Attribute Mapping

Some attributes exhibit high diversity yet convey the same underlying concept, so we map these varied outputs onto a unified category. For example, all Form values, "Hairspray," "Volumising Mousse," "Serum," and the like, are consolidated under Hair Styling, while diverse Benefit terms, "Volumizing," "Volume enhancement," "Maximum bounce," etc.; are normalized to VOLUME. This approach yields controlled, consistent outputs that streamline automated evaluation, product comparison, and downstream analytics. In summary, Shelfie's attribute extraction quality is not solely a function of the underlying model's capacity but also a direct result of meticulously engineered prompt strategies. By carefully tuning the granularity, domain context, example framing, and attribute definitions within our prompts, we were able to significantly boost the reliability of structured output from the vision-language model. These learnings underscore the critical importance of prompt engineering as a first-class design component in LLM-powered vision systems.

From Full Images to Object Detection: A Necessary First Step

Initially, one might consider directly feeding full images into a multimodal LLM for product identification. However, this approach suffers from several critical limitations. First, processing entire high-resolution images is computationally expensive, placing a significant burden on the LLM. Second, irrelevant background information can dilute the signal and introduce noise, hindering the LLM's ability to focus on the product of interest. Third, the LLM's attention mechanisms, while powerful, may struggle to effectively isolate the product from the surrounding clutter. Object detection provides a crucial first step in addressing these limitations. By employing object detection algorithms (e.g., YOLO, Faster R-CNN), we can identify and localize the product within the image, generating a bounding box that encapsulates the region of interest. This significantly reduces the computational cost by limiting the area that the LLM needs to process. Testing on the extraction of product ingredient lists highlights several key points related to how best to process images with LLMs. Multimodal LLMs (and even specialized OCR/text extraction pipelines) leverage both context and local content. Cropping to the full product outperforms retaining the whole image because multimodal LLMs rely on contextual cues beyond the immediate ROI. While background clutter degrades performance (as seen in the whole image case), the full product crop preserves crucial packaging, labeling, and spatial context, which aids the LLM in accurately identifying and extracting attributes like ingredient lists. This contextual information helps resolve ambiguities, mitigates OCR errors, and leverages spatial/semantic relationships between different elements on the packaging. As an additional observation,

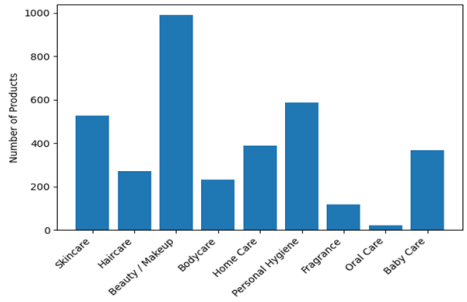


Figure 2. *Distribution of products across different categories.*

cropping exclusively to the ingredient list removes this context, making it harder for the model to disambiguate poorly imaged or partially occluded text and reducing extraction accuracy. Thus, retaining the whole product (minus background) maximizes relevant signal for LLM-based attribute extraction. These phenomena play out across all the product identification attributes we are interested in.

Results of Ingredient List Extraction by Image Type. Values marked * are averaged across 10 repeat runs (model used gpt4o).

Metric	Full Image	Product Crop
Ingredient List Length	24	24
Total Detected* (num)	24	24
TP* (original list match)	21	24
FP* (original list not in our list)	3	0
FN* (original present, ours not)	4	0
Detection Rate*	0.96	1.00
Precision*	0.875	1.00
Recall*	0.84	1.00
F1 Score*	0.857	1.00

Experiments and Evaluation

Dataset

Shelfie was evaluated on a proprietary dataset consisting of approximately 1000 user-submitted images captured in real-world arrangements of their personal environments. These contained images of a wide variety from countertops to vanity tops, cabinets and even included images of products in their showers. The images represented an uncontrolled environment that is found in most users' homes. They lack any system of sorting or arrangement and are varied in terms of their orientations, lighting conditions and image quality. Each product instance in the dataset was manually annotated with structured metadata. Figure 2 represents the distribution of categories across different products in the dataset.

Experimental Setup

Experiments were conducted to assess Shelfie's effectiveness in terms of its accuracy against the ground truth. The evaluation was performed on the field-level accuracy of each extracted attribute against the manually created structured attributes. All images underwent the complete processing through the pipeline and the final structured output was compared attribute-wise to gain insights.

Evaluation Metrics

The following metrics were used for evaluation:

- Attribute Accuracy: The percentage of correctly predicted values for each product field.
- Overall Trends: To get a holistic understanding of the effectiveness across all common fields.

Due to the fine-grained nature of the task and the diverse products, we focus on per-attribute breakdowns.

Results

Shelfie exhibited high accuracy across nearly all attributes for both skincare and haircare products. The results prove the system's strong ability to interpret both visual and textual cues in uncontrolled environments. Table 2 presents a comparative analysis of accuracy levels across various attributes. It highlights the performance distinctions among different GPT models, Gemini's latest iteration, and Qwen. Although attributes such as Brand and Product Name are extracted with consistently high accuracy across all model types, others show marked degradation. For example, the Product Line attribute achieves substantially lower scores with both Gemini and Qwen, a gap driven primarily by semantic ambiguity: line names often appear adjacent to-or even visually integrated with-the brand logo or product name (e.g. "UltraClean"), forcing the model to discriminate which text span denotes the sub-brand versus the core brand identity. Likewise, Container Type extraction is hindered by overlapping visual cues: many packaging formats (cylindrical bottles, spray cans, tubs) share similar silhouettes-round bodies, caps, and labels-yet differ only in subtle three-dimensional or material characteristics. Without explicit depth or tactile information, models trained on generic caption data lack the granularity required to reliably distinguish these form factors, resulting in reduced extraction accuracy. Shelfie robustness is judged by its performance in real-world environments, such as under varying lighting conditions, occlusion, and image quality. The pictures that are usually uploaded by users tend to be uncontrolled, thereby offering a few serious challenges. Lighting is fundamental; it is expected that non-uniform light, shadows, or glare will reduce visual cues and increase the difficulty of segmentation and classification. While reliable in normal light, the extreme conditions, such as dim corners or harsh reflections, can lower confidence or cause misclassifications. Occlusion is commonplace, which is mostly mitigated by good crop/bad crop segregation. Heavily blocked or poorly visible items are filtered as "bad crops," allowing only well-represented items toward metadata extraction that removes noise. Other problems include odd viewpoints, blur, or tiny product appearances. These factors will affect the quality of segmentation and classification. Though robust against ordinary conditions, extremes result in incomplete results. Shelfie's stability relies on smart cropping and steady normal light performance for consistent metadata despite environmental shifts. Shelfie supports a number of different quantities of products per image with

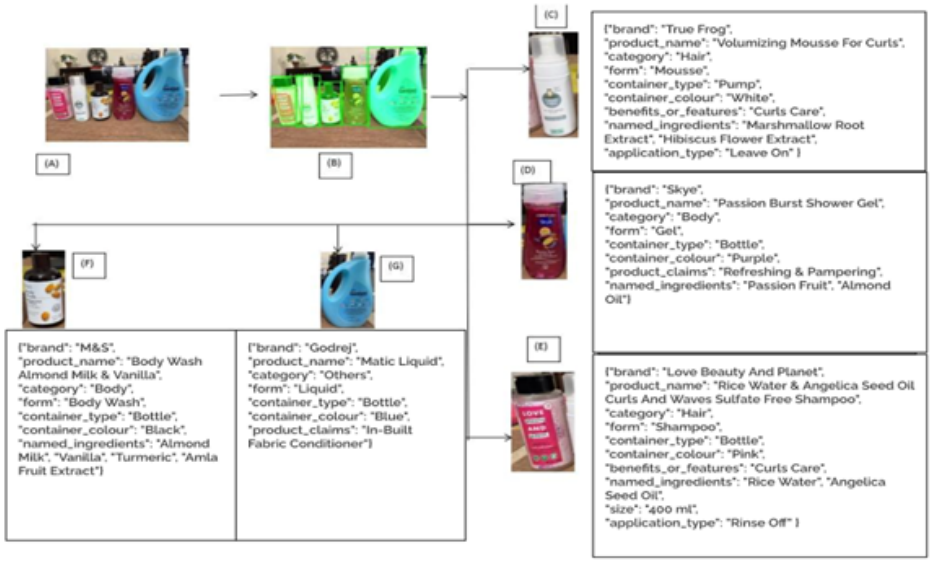


Figure 3. *(A) User Input (Shelfie Image) (B) Transformed images showing good crops (in green bounded boxes) and bad crops (in red bounded boxes) (C) Product Crop-1 (D) Product Crop-2 (E) Product Crop-3 (F) Product Crop-4 (G) Product Crop-5*

Results of attribute extraction by model type.

Attribute	Accuracy (%) (GPT-4-turbo)	Accuracy (%) (GPT-4o)	Accuracy (%) (GPT-o1)	Accuracy (%) (GPT-o3)	Accuracy (%) (Gemini-2.5-pro-preview)	Accuracy (%) (Qwen2.5-VL-3B-Instruct)
Brand	98.40	96.80	97.60	97.60	100	88.80
Product Line	96.00	89.60	88.00	89.60	70.51	64.80
Product Name	96.80	97.60	96.80	96.00	92.30	92.80
Form	94.40	92.80	92.00	93.60	85.89	58.40
Container Type	90.40	92.00	91.20	88.80	41.02	48.80
Container Primary Color	85.60	85.60	85.60	87.20	80.76	72.00
Benefit or Feature	82.40	72.00	87.20	83.20	84.61	66.40
Product Claims	76.00	60.80	62.40	60.80	73.07	47.20
Named Ingredients	83.20	79.20	91.20	83.20	75.64	73.60
Size	96.00	87.20	92.00	92.80	96.15	77.60

very little performance degradation. Additional products mean more crops, and that introduces concerns about the added processing overhead. Scalability is made possible by the crop-and-segmentation architecture. Each detected product is segmented as a discrete crop, and its recognition accuracy is unaffected by other products' quantity. System performance is consistently maintained if the input has a single item or a tight shelf. Scalability is also affected by product separation. Well-separated products function as expected. Tightly packed or overlapping products may yield incomplete/occluded crops that Good/Bad crop segregation filters out to avoid unreliable metadata. Processing time increases only incrementally with product count due to parallelization. Accuracy is constant since every product crop uses the same classification pipeline. Shelfie scales well, maintaining robustness and reliability despite increasing input image complexity.

Error Analysis

A qualitative review was performed in order to gain insights as to why Shelfie failed to match ground truth annotations. The following causes were found:

- Partial occlusion of ingredients or brand names which leads to an incomplete extraction of data. Ambiguous packaging design due to which subtle form or type factors like cream vs lotion were not distinguishable.
- Promotional labels being misinterpreted as true facts and being categorized as benefits or features.
- Non-standard packaging or multilingual text causing confusion in attribute alignment.
- Multiple container types for a single product like some products both “Spray” and “Can” had to be dealt with as special cases where multiple types would be correct.
- Missing form information, where only brand and product name were displayed resulting in the prompt to either infer or completely leave out the form attribute.

Future Work

Shelfie has shown great promise in extracting structured product information from user-generated images and thus has opened up important opportunities to further consumerize the system. One such opportunity involves further personalization by relating Shelfie's outputs to individual factors such as skin type, allergen identification, local climate, and direct consumer feedback. Adding features, like monitoring how an individual's preferences change over time, would allow Shelfie to provide dynamic recommendations that are truly personalized. Another area of opportunity is the integration of Shelfie on mobile devices, enabling users to get instant insights and product recommendations by scanning their shelves. Enabling this with on-device inference would safeguard privacy for users while providing secure, offline functionality, which is crucial in the self-care category where data is sensitive. Lastly, extending Shelfie's capabilities with multimodal reasoning, where text the user inputs (e.g., issues such as ”acne” or ”dry skin” or “vanilla fragrance allergy”) is matched with visual product information, has the potential to unlock a new level of personalized insight. By doing so, Shelfie can become an entire self-care guide: one that evolves with the prolific and rapidly shifting world of lifestyle care while holding each user's individual path, routines, and requirements close.

Conclusion

In this paper, we introduced Shelfie, an end-to-end vision language pipeline designed to extract structured metadata from user-submitted images of lifestyle care products. Shelfie helps to bridge the gap between raw, real-world visual data and structured information by utilizing the GPT-4V-powered multimodal language model which provides state-of-the-art computer vision capabilities to our system. This fusion allows the system to analyze, interpret and extract data from user-generated images which are often informally captured, visually complex, and linguistically varied in the natural environments of their homes. We introduced an intelligent filtering mechanism to reject stock images and poor-quality crops and also demonstrated how prompt based engineering can be an effective replacement for traditional methods. Shelfie was evaluated in uncontrolled, real-world environments using images submitted by users from diverse backgrounds. Its ability to maintain consistent and accurate results across this wide variability confirms the robustness and generalizability of the pipeline. The system's capability presents an opportunity for more consumer-oriented uses. These could range from customization of product recommendations based on other users' usage habits, product filtering based on the ingredient sensitivity and product recommendations based on usage analytics. By focusing on actual product use instead of point-of-sale marketing information, Shelfie positions itself as a platform that is consumer-focused and enables people to make better and more personal choices in their own self-care paths.

References

[1] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, A Structured Output Learning Approach for End-to-End Text Recognition, Proc. CVPR, pp. 1778–1786 (2016).

[2] Vivino, Vivino Wine App, Apple App Store, Version 6.33.0, Accessed 20 June 2025.

[3] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, Layout1M: Pre-training of Text and Layout for Document Image Understanding, Proc. ACM SIGKDD, pp. 1192–1200 (2020).

[4] X. Zhan et al., Product1M: Towards Weakly Supervised Instance-Level Product Retrieval via Cross-Modal Pretraining, arXiv:2403.07615 (2024).

[5] Y. Zhu et al., Multimodal Joint Attribute Prediction and Value Extraction for E-Commerce Product, arXiv:2009.07162 (2020).

Author Biography

Ankur Purwar is a Senior Director and Research Fellow at Procter & Gamble R&D based in Singapore. His research interests are at the intersection of data science, computer vision, and consumer product innovation. He received a PhD in Applied Mathematics from Indian Institute of Technology Kanpur, India.

Alex Hollingworth received his BSc in Chemistry from the University of Wales Swansea in 2004. He is a Data Scientist at Procter & Gamble in Singapore. After joining P&G R&D in 2005, he led global product design programs across Fabric Care, Home Care, and Beauty Care. In 2019, he transitioned to Data Science and modelling, leveraging domain expertise to apply Computer Vision and Generative AI to consumer research and product development.

Laveena Satwani received her education from the Indian Institute of Information Technology, Design and Manufacturing, Jabalpur. She is currently a Principal Engineer specializing in Computer Vision and AI at Big Vision. Her work focuses on developing advanced AI and computer vision solutions, applying machine learning techniques to real-world vision problems. She has extensive experience in building and deploying intelligent systems for industry-scale applications.

Eu Jack Tan received his higher education from the National University of Singapore. He is currently a Category CIO at Procter & Gamble, where he leads digital transformation initiatives using AI and cloud technologies. With over a decade of experience, his work focuses on building scalable data and AI platforms across R&D, digital marketing, e-commerce, and enterprise services, driving data-driven innovation and business growth in global markets.

Nandita Mishra is an undergraduate student at the Vellore Institute of Technology. She has worked as an intern at Big Vision, gaining experience in applied technology projects, and is currently an intern at Schneider Electric.With a strong interest in

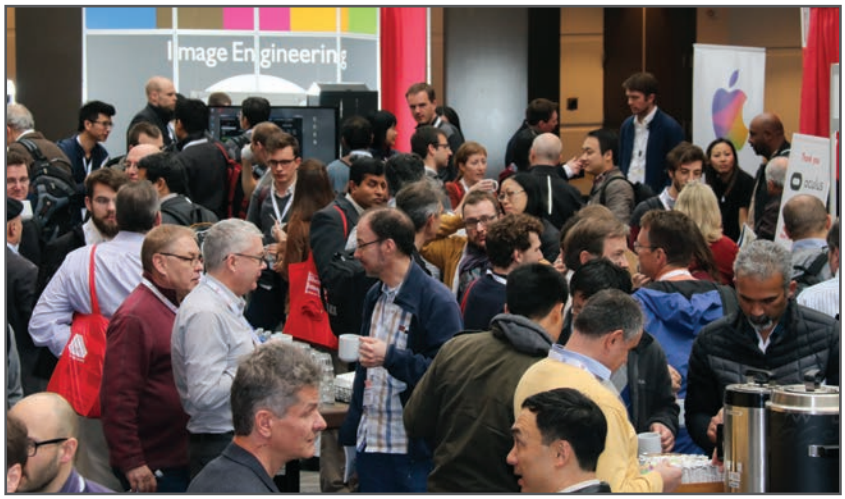
applied machine learning, she has experience working on computer vision projects and is now focused on AI and automation initiatives.

Pranav Mishra received his education from the Indian Institute of Technology, Bombay. He is currently the Vice President of Technology at Big Vision. With over 17 years of experience in computer vision, machine learning, and AI, his work focuses on building state-of-the-art solutions for video analytics, object detection, classification, and semantic segmentation. He has led teams developing scalable vision systems widely adopted across industry and startup ecosystems.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org