Enhancing Fashion Outfit Evaluation through Object Detection, Deep Learning, and Generative AI: Towards Context - Aware Recommendations

Varun Shanmugam
Chirec International School

***Abstract*— Artificial intelligence is increasingly being utilized in the fashion world. Such applications either dwell on identifying an article of clothing or making dull recommendations because they lack knowledge of the socially contextualized world where these articles of clothing are being used. Therefore, this paper sets up a pipeline approach that integrates object detection with generative ability to develop occasion-driven clothing evaluations. With a trained object detection model, from different fashion-related databases and post-edited with non-maximum suppression, a method detects pieces of clothing from input images pruning redundancies or spurious positives producing judgments. The detection is primarily from the item that the person is wearing compared to the phenotype attributes it may possess, like shirt- or pant-type sleeves, pockets, etc. This extraction then serves as the prompt for an epochal answer from a large language model to remain related to what was extracted and not by hallucination of other articles.**

In order to try the pipeline, numerous example images with varying clothing and gender orientations were tested scenarios based on different social contexts and the results were evaluated for detection, non-redundancy, and producing appropriate suggestions. The pipeline successfully functioned; it is possible to evaluate and then produce explainable, situationally-relevant suggestions that are more human-like in style than the early detection generating capabilities. While formative, the benefits and drawbacks of depending on pre-defined visual skeletons and generative methods for fashion use were uncovered. Constraints were due to dataset biases to potential in measuring "appropriate" recommendations. The value is that it bridges the gap between low-level recognition and high-level contextual knowledge through future prospects of generalizable, socioculturally apt, and realistic AI orientation direction fashion recommendation systems.

I. INTRODUCTION

1.1 Background

The intersection of artificial intelligence (AI) and the fashion world is now the hallmark of technological advancements over the last ten years. Advances in computer vision, machine learning, and natural language processing are redefining the way that people consume, the way that brands produce, and how fashion is interpreted across cultures. Digital technologies are now responsible for a large proportion of the competitive advantage in retail, according to international industry reports, with personalization, efficiency, and data-driven insight emerging as differentiators.

Specifically, AI deployment to fashion analysis goes beyond clothing identification to more sophisticated evaluative work. Today's users anticipate systems that can classify pieces of clothing, deduce their

characteristics (e.g., type, fit, color), and determine their appropriateness for specific occasions or settings. This anticipation is part of larger societal trends in consumption, where globalization, e-commerce, and social media fuel fashion decisions.

At the technical level, object detection is pivotal in this work. As the process of identifying and categorizing objects in images, object detection has been pushed forward by architectures like YOLO (You Only Look Once), Detectron2[3], and EfficientDet[4]. These frameworks have been extensively used for identifying clothes, accessories, and human poses in large datasets. But though item detection models are extremely good at identifying items, they continue to be restrained in judging whether those items suit together into a contextually fitting outfit. Current research is placed exactly in filling this limitation.

1.2 Problem Statement

Even with advances in computer vision, current detection pipelines exhibit two essential failures when used for fashion assessment tasks. First, detection models have the tendency to make duplicate predictions, where several bounding boxes are produced for a single item. Redundancies of this nature are particularly likely in images in which clothing overlaps or is partially occluded, as is sometimes the case in real-world environments. The duplications skew system outputs, mislead downstream assessors, and compromise result reliability.

Second, detection models are for the most part context-insensitive. Sure, they may be able to detect "shirt," "pants," or "shoes" with precision, but they are not able to reason about whether or not this particular outfit makes sense for a given setting, say, a business meeting or a cultural event. This absence of context is a deep flaw considering that fashion appropriateness is inextricably bound up with occasion, culture, and social convention.

These limitations signal an approach-related requirement: to investigate how object detection and generative reasoning can be combined to yield valid, context-specific fashion suggestions.

1.3 Research Question and Objectives

This work is predicated on the following research question:

How are object detection and deep learning synergized with generative AI to detect and assess clothing outfits, and make contextually relevant suggestions based on the event?

To this end, the paper follows three main goals. The first is to improve reliability in garment detection through the application of pretrained models with non-maximum suppression (NMS), minimizing duplicate responses. The second is to format the detection outputs into machine-readable styles so that features like garment type and color can be passed clean to reasoning models. The third is to use prompt-engineered large language models (LLMs) to provide ratings of outfit appropriateness. By limiting LLMs to infer only using identified objects and their properties, and framing their answers in accordance with user-determined contexts, this research guarantees outputs are accurate as well as application-oriented.

Collectively, these goals operationalize the question of investigation by conflating low-level recognition with high-level reasoning, filling a gap that cannot be bridged by either detection or generative AI alone.

1.4 Significance of the Study

The importance of this research is its potential to redefine the role of AI in assisting fashion decision-making across consumer and industry domains. Through the solutions to the duplication errors and contextual blind spots, the research makes a contribution toward developing systems that are both technically sound and socially relevant.

Practically, the system may be integrated into personal styling assistants to provide real-time feedback on outfit appropriateness for casual users. To e-commerce websites, it may minimize returns by advising customers to make occasion-based purchases. Outside its business uses, the framework also ensures cultural diversity since context-aware assessments go beyond Western-oriented norms and abide by various dress codes. In addition, the method has educational applications, enabling fashion students and designers to try out combinations of outfits and gain formal, automated critiques

By locating technical advancements in detection within a broader framework of inclusivity and usability, the research extends both scholarly

understanding and practical implementation at the nexus of AI and fashion.

1.5 Structure of the Paper

The rest of the paper is structured in the following sections. The Literature Review reviews literature in object detection, clothing analysis, and dataset creation, highlighting outstanding issues with contextual reasoning and duplicate elimination. The Methodology details experimental methodology, explaining use of pre-trained models, NMS implementation, and integration with LLMs via prompt engineering. The Analysis reports results from four sample images, assessing detection accuracy, duplicate reduction, and contextual appropriateness of recommended outputs. Last, the Conclusion integrates the study's main contributions, recognizes limitations, and outlines future research directions for multimodal and inclusive fashion AI

II. LITERATURE REVIEW

2.1 Overview of the Review

This review combines three essential strands of research that form the foundations of these systems: (i) object detection and attribute recognition, (ii) outfit compatibility and generative assessment, and (iii) data sets and assessment protocols. These threads unify into modular pipelines that bring together detection, semantic description, and reasoning, but diverge in how they address cultural diversity, contextual suitability, and bias. Through delineating the field and its constraints, this review positions the current project: an integrated, occasion-sensitive system that merges instance-grounded detection with large language model reasoning.

2.2 Garment Detection and Attribute Recognition

Strong garment detection is the backbone of fashion AI. Contemporary systems heavily depend on Convolutional Neural Networks (CNNs), deep neural networks that use convolutional filters over images to tap out spatial hierarchies of features like edges, textures, and shapes. CNNs have been extremely useful for clothing item classification and location in varied contexts, from e-commerce catalogs to street photography.

Within detection designs, one-stage models like the YOLO[1] ("You Only Look Once") family are famous for real-time performance. YOLO performs bounding box and class probability predictions in a single pass, making it appealing to mobile or embedded scenarios where low latency is essential. In comparison, two-stage detectors like Faster R-CNN and its variant, Mask R-CNN, first produce region proposals and then refine the classifications[2,5]. Mask R-CNN also generates segmentation masks pixel-level boundary of clothes that enable more accurate reasoning in instances of overlapping or intricate layering of clothes.

More recently, transformer-based models like DETR (Detection Transformers) and its derivatives rely on attention mechanisms to directly model relations between all patches of an image. Although more computationally intensive, these models are particularly good at recovering missing or occluded small things like ties or jewelry that CNN-based detectors fail to detect. In reality, researchers compromise between speed and accuracy by selecting architectures appropriate for their use case YOLO for responsiveness, Mask R-CNN for accuracy, or DETR for reliability in adverse conditions.

Detection pipelines usually utilize Non-Maximum Suppression (NMS) to eliminate redundant bounding boxes by keeping the most confident prediction for every object. This straightforward but necessary step provides cleanliness in densely populated fashion images, where several detectors could otherwise predict overlapping boxes for the same shirt or shoes.

Beyond categories, fashion analysis requires identification of attributes like color, material, shape, and style. Early methods refined CNN classifiers to work on multiple attributes, but these tended to misassign features between garments, particularly in dense or occluded settings[6]. To overcome this, current systems utilize attention mechanisms and multi-label classifiers that tie attributes more robustly to individual garments[7,8]. A typical two-stage pipeline first detects and segments objects, and then crops and passes them into a dedicated attribute classifier. This modular approach captures both technical convenience and dataset organization, as the majority of large-scale fashion corpora have each category and attribute annotated independently.

2.3 Outfit Compatibility and Generative Evaluation

While the identification of "what is present" is crucial, fashion intelligence also involves assessing "how

well" pieces go together and "what changes" would enhance a look. Writing on this task falls into three methodological traditions.

Graph-based models imagine an outfit as a graph of clothes and their properties, with edges denoting relations of compatibility like color matching, texture contrast, or style consistency[9]. Edge weights can be learned from massive co-occurrence statistics or hand-crafted fashion rules so that the model can predict compatibility scores for novel combinations.

Discriminative methods break down outfit rating into several sub-criteria, including garment quality, stylistic fit, or user customization[10]. Each criterion is learned individually, and the predictions are combined into a single compatibility score. The "multi-expert" method offers interpretability since it explains which properties of an outfit compel its rating.

Generative methods advance further by suggesting new or substitute items to enhance an ensemble. Generative Adversarial Networks (GANs) were employed by early models, where a generator is placed in competition with a discriminator to produce realistic fashion images[11]. Diffusion models, which upscale random noise to coherent images, have more recently been employed to generate full outfit proposals or feasible substitutes for absent garments. These models facilitate active styling by proposing modifications instead of just scoring current combinations.

But most of these systems function in a vacuum of aesthetics. They optimize for visual coherence or frequency of historical pairing, but never for social appropriateness. Consequently, a model may suggest fashionable but contextually inappropriate garments, like combining sneakers with formal evening dress. Solving this shortcoming necessitates systems that include occasion awareness, infusing knowledge of cultural norms and dress codes into the recommendations.

New developments in Vision–Language Models (VLMs) and Large Language Models (LLMs) offer great potential for this task. VLMs like CLIP bridge visual inputs with text descriptions within a common semantic space, so they can associate clothing with general concepts like "wedding wear" or "business casual." LLMs like GPT-style transformers are good at providing context-specific natural language descriptions and suggestions[12].

One typical integration approach is serializing discovered garments and features into JSON (JavaScript Object Notation), a format of structured text used extensively for key–value data representation. For instance, a JSON entry may enumerate a "black tuxedo jacket," "white dress shirt," and "sneakers." This entry can subsequently be given as input to an LLM, which interprets the ensemble and provides personalized recommendations: "Substitute the sneakers for formal shoes appropriate for a black-tie wedding." Such coupling fills the gaps between structured computer vision outputs and natural language reasoning, enabling pipelines that can provide both accuracy and interpretability.

2.4 Datasets and Evaluation Practices

The speed of fashion AI advancement relies significantly on datasets. DeepFashion and DeepFashion2 are two of the most popular, providing hundreds of thousands of annotated images with bounding boxes, landmarks, attributes, and segmentation masks. These datasets enable training and benchmarking for detection to retrieval tasks. Fashionpedia pushes the field forward with a designed ontology of 27 categories, 19 garment components, and almost 300 attributes, which is annotated at pixel level. This ontology facilitates comprehensive, instance - based attribute learning[13, 14, 15].

Datasets like StreetStyle extend cultural and temporal scope by sampling street photography from around the world. They tend to have sparse annotations, though, and are therefore more suitable for unsupervised or weakly supervised learning. Across datasets, shared issues are label noise, varying taxonomies (e.g., "blazer" vs. "jacket"), and lack of explicit occasion labels. Even when datasets have formality tags, they tend to be too coarse and culturally specific to capture actual-world dress codes adequately[16].

Evaluation criteria also differ between tasks. Object detection is often measured by mean Average Precision (mAP), whereas attribute classification employs per-class F1 measure or subset accuracy. Outfit compatibility is often tested through retrieval-based measures or human preference experiments. However, few studies specifically measure occasion fit, resulting in a gap between technical evaluation and real-world applicability. This absence of common evaluation for contextual suitability is a dominant shortcoming in existing research.

2.5 Persistent Challenges

Through the threads of garment recognition, compatibility modeling, and dataset construction, three omnipresent challenges present themselves:

1. Bias in culture and dataset – BENCHMARKS disproportionately over-represent Western fashion traditions and bodily shapes, which results in suggestions that will fail or backfire when tested in non-Western environments.
2. Finding small or occluded objects – Belts, jewelry, and ties continue to be challenging to find in dense scenes or low-resolution images. Although segmentation-dominant methods enhance accuracy, they boost computational loads.
3. From retrieval to reasoning – Most contemporary systems are still working as catalog retrievers, picking and choosing among existing items instead of reasoning about occasion constraints or producing bespoke suggestions. Occasion-aware reasoning demands models that combine symbolic rules, visual hints, and natural language explanations.

2.6 Synthesis and Positioning

Together, the literature presents a modular sequence: detectors pinpoint clothing, attribute classifiers characterize them, compatibility models rate them, and generative models offer substitutes. But these parts are splintered, with scant concern for context and cultural standards.

The current work fills this void by combining a high-performance detector (e.g., YOLOv8 or Mask R-CNN) with a multi-label attribute classifier trained on Fashionpedia and DeepFashion2. Outputs are JSON-serialized and fed into an LLM or VLM, which generates occasion-aware evaluation and recommendations. Evaluation will go beyond typical metrics to rigorous human studies where context appropriateness is measured explicitly using rubrics like formality scales and cultural modesty norms.

In the process, the project seeks to bring technical performance and social applicability together, providing an engine that can generate recommendations not just visually consistent but also situationally relevant and culturally aware.

III. METHODOLOGY

3.1 Overview

The approach used in this research combines cutting-edge object detection with large language model (LLM) reasoning to create an occasion-aware fashion judgment framework. Three main goals were engineered into the pipeline: to provide a stable detection of clothing items inside images, to limit errors caused by duplicate detections, and to utilize generative reasoning to evaluate outfit appropriateness in certain social settings. To do so, the research built on a pretrained object detection model from Hugging Face (Valentina eve), custom non-maximum suppression (NMS) implementation for deduplication, output structured serialization in JavaScript Object Notation (JSON), and assessment via ChatGPT-4.0 using well-engineered prompts.

This approach draws on earlier research in computer vision, multimodal reasoning, and fashion AI. It bridges a significant gap in occasion-aware fashion recommendation workflows by integrating detection, attribute recognition, and generative evaluation.

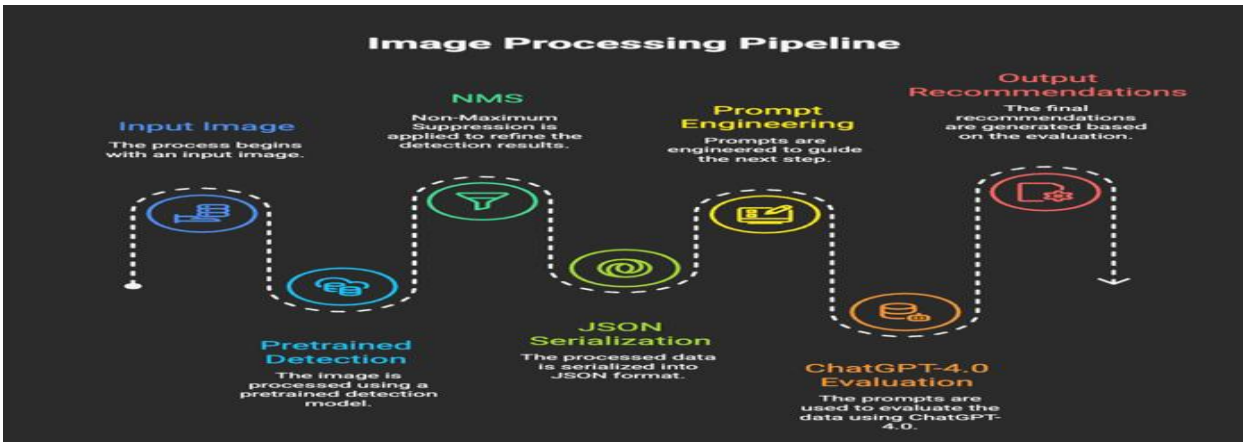

Figure 1: Overview of proposed method (Generated by Napkin AI)

3.2 Model Selection and Pretrained Foundations

At the center of the pipeline is the application of a pretrained detection model. A pretrained model is a neural network that has been trained on large-scale datasets, like COCO or Fashionpedia, prior to being fine-tuned or repurposed for a novel task. Pretraining allows models to capitalize on generalized feature representations, bringing computational costs and data needs significantly down from training from scratch.

A YOLOS object detection model was used for this project, leveraging its publicly available pretrained implementation through the official YOLO repository. Unlike traditional CNN-based detectors, YOLOS adapts the Transformer architecture for object detection, enabling it to capture global dependencies across an image[17]. This makes it particularly effective in fashion analysis, where garments often overlap or exhibit subtle contextual cues. The pretrained backbone provides strong baseline accuracy across diverse clothing styles while reducing implementation complexity.

3.3 Duplicate Reduction with Non-Maximum Suppression

One of the key difficulties of object detection is when the model gives several overlapping bounding boxes for one object. If not addressed, these duplicates artificially inflate category counts and bias downstream evaluations. To deal with this, a custom implementation of Non-Maximum Suppression (NMS) was added.

NMS is a recursively applied algorithm that chooses the bounding box with the top-scoring confidence score and discards other overlapping boxes above some threshold of Intersection over Union (IoU). By adjusting this threshold appropriately, the model learns to trade off two conflicting goals: detection accuracy, so that actual items are not discarded; and duplicate suppression, so that redundant bounding boxes are discarded.

As part of this project, a number of tests were also performed to find the best NMS threshold. The outcome illustrated that very strict thresholds threatened to throw out good detections while very relaxed thresholds retained duplicates. The chosen threshold represented the optimal trade-off between accuracy and duplicate suppression.

3.4 Testing Design and AI-Generated Image Selection

In order to verify the detection–deduplication pipeline, experiments were performed on multiple AI-generated images. These images were intentionally built to cover various contexts: a male subject in an indoor formal setting, a male subject in an outdoor casual setting, a female subject in an indoor formal setting, and a female subject in an outdoor leisure setting.

These settings were selected in order to provide gender, race, and environmental diversity. By having both indoor and outdoor settings, tests were made stronger under diverse lighting, backgrounds, and dress requirements. AI-generated imagery enabled having uniform control over environmental conditions without risking privacy or dataset bounds of actual images.

3.5 Serialization with JSON

After eliminating the duplicates, the results were serialized into JSON, a light-weight and readable data-interchange format. JSON was used due to its cross-platform compatibility with downstream natural language systems as well as for maintaining structured information.

Each record had fields defining garment category, detected garments, and confidence scores. This form of organized representation gave a tidy, machine-readable connection between the vision model and the LLM, such that generative reasoning was based on precise and correct detections instead of raw images themselves.

JSON was chosen as it is small, easy to read, and well-supported on multiple platforms. Its formal structure makes it possible to save and pass detection outputs efficiently while maintaining compatibility with downstream LLM systems, and for this reason, it makes a perfect bridge between visual object detection and generative reasoning tasks.

3.6 Prompt Engineering for Contextual Evaluation

One of the key elements of this approach was the application of prompt engineering to restrict and direct the thinking of ChatGPT-4.0. Prompt engineering involves deliberately designing input queries to optimize the relevance, reliability, and interpretability of LLM responses. In this research, prompts were designed to prioritize three primary constraints. First, the LLM was specifically told to limit analysis to only the garment items that were listed in the JSON

serialization, meaning that its analysis would be based on real detections and not on hypothetical components. Second, prompts included contextual clues, such as whether the individual wearing the garment was heading to a wedding, a beach volleyball game, or a dinner party, hence matching evaluations with particular social events. Lastly, the wording of the prompts dissuaded the model from inventing new items while keeping it free to propose high-level changes, like adding an accessory or swapping out a type of garment.

This methodological strategy secured the LLM's evaluative capability its capacity for social norm understanding and recommendation articulation while keeping to a minimum the potential for introducing errors. Over several cycles of prompt tuning, the research found formulations that in every case struck an optimal balance between factually groundedness and creative, context-aware output.

3.7 Evaluation with ChatGPT - 4.0

ChatGPT-4.0 acted as the generative test engine. Based on the JSON-formatted detections and context patterns, the model created natural-language analyses to determine outfit appropriateness and propose possible enhancements. Those results were subsequently examined against three measures: whether the response mentioned just items actually detected, whether the suggested enhancements conformed to the occasion given, and whether the suggestions were actionable in the sense of being specific and useful in practice.

For instance, assessment of a male subject in an informal open environment may find the shirt and shorts appropriate but propose fewer heavy materials or a sleeveless version if the situation involves physical activity. In another scenario, the model might advise the pairing of a formal gown with matching accessories when assessing a female subject going to an evening formal event. These outputs demonstrated the model's capacity to reason about context and to map clothing selections onto social norms with grounding in real detections.

3.8 Combined Testing and Iterative Refinement

The approach was iteratively tested to optimize every step of the pipeline. The raw results of the pretrained model were initially tested to determine a detection baseline. NMS thresholds were subsequently tuned to improve accuracy and duplicate suppression, with subsequent verification of JSON serialization for completeness and consistency. Lastly, prompts were tested and optimized until LLM responses showed consistent adherence to grounding rules and contextual relevance.

This iterative process showed that the system was best when NMS was moderately strict, JSON output was minimized to key attributes, and prompts were brief but clear in terms of constraints. Every phase was thus optimized not independently but as part of an entire workflow.

3.9 Limitations and Considerations

While the approach offers a solid foundation, there are a number of limitations that need to be noted. The pretrained model might not be able to fully represent garments beyond popular Western clothing types, and the limited sample of AI-generated images, even if controlled and diverse, restricts generalizability. Furthermore, subjective fashion norm variation implies that judgments will also vary by cultural context, even when prompts are well-tailored. Lastly, in spite of the protections of timely engineering, sporadic small hallucinations due to the LLM were noted, which emphasizes the requirement for further improvement in grounding techniques.

3.10 Summary

This approach integrates pretrained object detection with bespoke NMS deduplication, ordered serialization, and LLM-based assessment to provide a scalable, occasion-sensitive outfit recommendation pipeline. Through rigorous testing with multiple artificially generated images, the research shows the applicability of unifying computer vision and generative reasoning to fashion usage. Through using prompt engineering to limit hallucination and optimize contextual specificity, the pipeline provides a viable framework for future multimodal fashion AI. Hereafter is a plain diagram representing the pipeline used in the testing phase of this paper.

IV. ANALYSIS

4.1 Detection Performance and Non-Maximum Suppression

The initial phase of the planned pipeline was based on a pretrained Hugging Face object detection model,

which was fine-tuned to identify a wide range of clothing and accessories. Object detection is a machine vision task where models detect instances of semantic objects in an image and locate them through bounding boxes. In our research, bounding box overlap was managed by NMS, a common post-processing method intended to eliminate duplicate or overlapping predictions. NMS works by keeping the highest-confidence detection and suppressing others with an intersection-over-union (IoU) that is very high, i.e., the ratio of overlapping area to union area between two bounding boxes. Below is a use case of NMS; while it is not the clearest, it works well in showing how NMS cuts out most of the repeat boxes analysed by the model. Although it reduces duplicates, too high of an accuracy value can cause loss of detection of apparel or clothing pieces

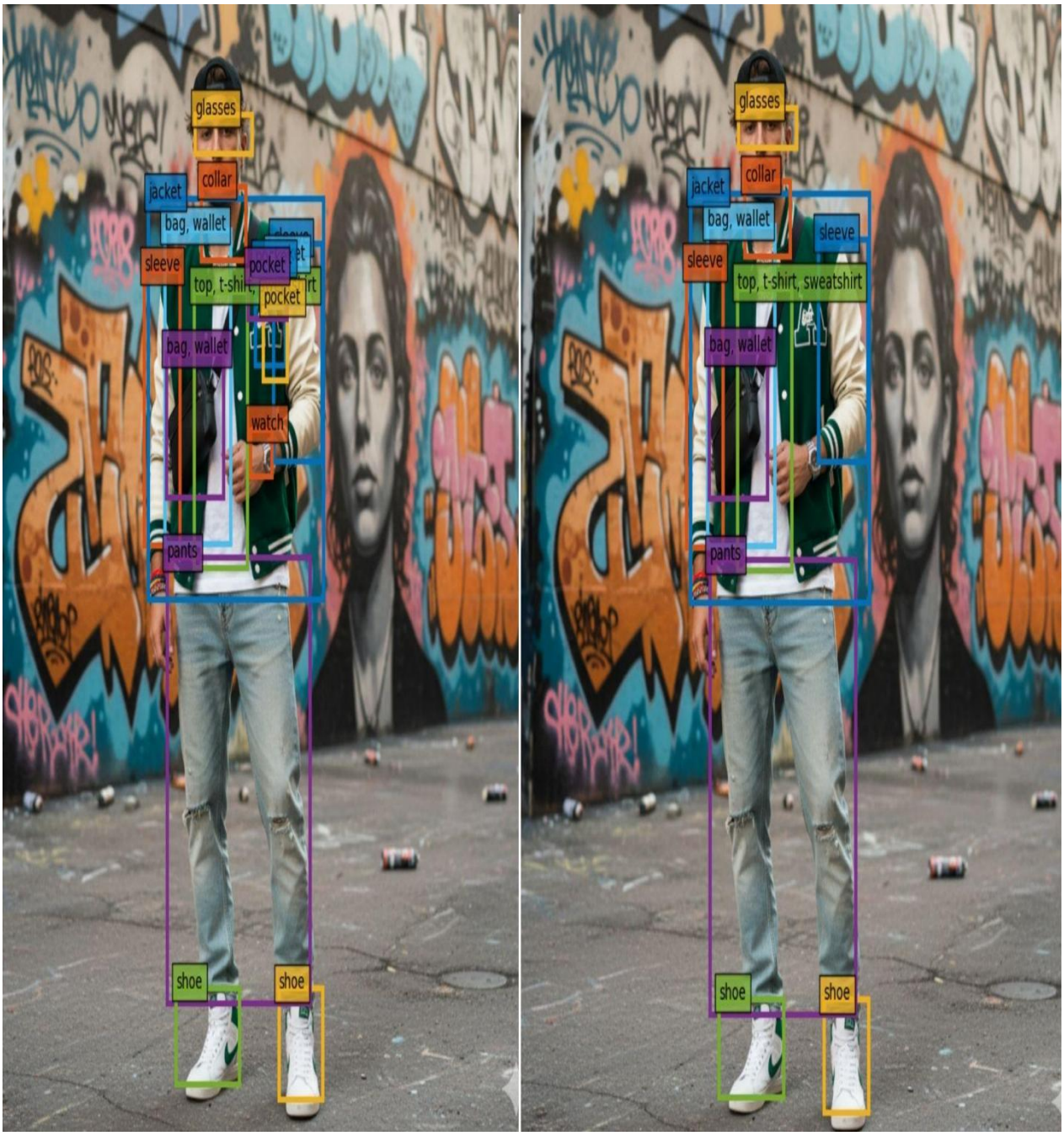

Figure 2: Sample image before and after use of NMS (generated by Gemini AI from Google)

Cross-testing on four images created by AI two male (beach volleyball wear, formal blazer/tuxedo outfit) and two female (saree outfit, casual jeans/top/jacket) showed that the pretrained model repeatedly detected major clothing items with high accuracy. In the image of beach volleyball, for instance, the model identified the tank top, shorts, and sneakers correctly. Implementation of NMS effectively eliminated duplicate bounding boxes for overlapping areas, balancing detection recall and suppression of duplicates. But minor add-ons like wristbands or discreet jewelry sometimes led to missed detections or partial overlaps, reflecting limits in the model's ability for fine-detail recognition in unconstrained scenarios.

Overall, detection quality, qualitatively measured for this exploratory study, shows that pretrained models can give solid baseline recognition towards downstream assessment, particularly when augmented with NMS for bounding box adjustment. Placeholder images can be used to depict bounding boxes prior and subsequent to applying NMS in order to visually demonstrate the effect of eliminating duplicate findings. The following are the outputs for three of the images with detected clothing items by the model coupled with the NMS algorithm.

I.

Figure 3.1: Sample image before and after detection (generated by Gemini AI from Google)

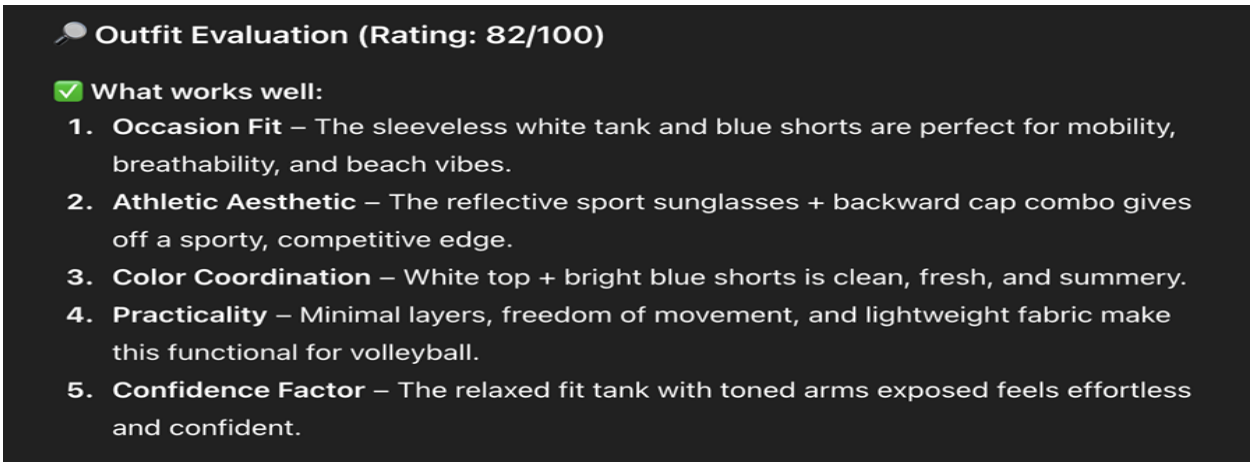

Figure 3.2: GPT - 4.0 response for occasion set as: beach volleyball

🔎 **Outfit Evaluation (Rating: 18/100)**

✅ **What works well (barely):**

1. **Confidence Factor** – Sleeveless top shows comfort in your skin, but that's more gym/beach energy than formal.
2. **Color Simplicity** – White and blue are safe colors, just not styled for the occasion.
3. **Minimalist Look** – Clean and unfussy, but unfortunately comes off as underdressed.

Figure 3.3: GPT - 4.0 response for occasion set as: formal dinner

II.

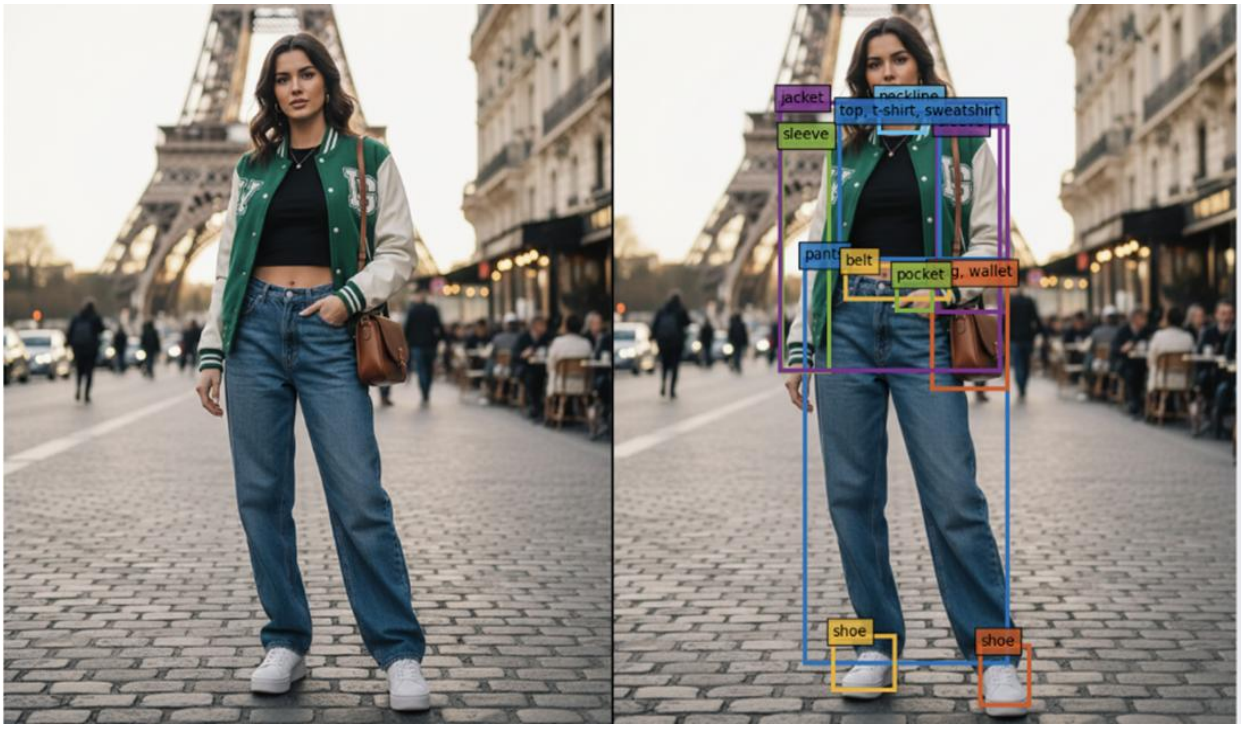

Figure 4.1: Sample image before and after detection (generated by Gemini AI from Google)

🔎 **Outfit Evaluation (Rating: 88/100)**

✅ **What works well:**

1. **Effortless Street Style** – The green varsity jacket + crop top combo gives a trendy, youthful vibe perfect for casual meetups.
2. **Balanced Silhouette** – Cropped top with high-waisted jeans creates a flattering proportion without being overdressed.
3. **Footwear Choice** – White sneakers keep the outfit relaxed, practical, and cohesive with the street style setting.
4. **Color Coordination** – Green, black, denim blue, and white balance well, with the brown crossbody bag adding an earthy, neutral contrast.
5. **Parisian Backdrop Fit** – The style matches the urban chic aesthetic, making the look feel intentional and modern.

Figure 4.2: GPT - 4.0 response for occasion set as: casual meetup in a cafe

🔎 **Outfit Evaluation (Rating: 10/100)**

✅ **What barely works:**

1. **Sneakers** – They're comfortable for walking to the pool, but you'll need to take them off immediately.
2. **Crossbody Bag** – Handy for carrying essentials like wallet/keys, but not swim gear.

Figure 4.3: GPT - 4.0 response for occasion set as: swimming in a pool

III.

Figure 5.1: Sample image before and after detection (generated by Gemini AI from Google)

🔎 **Outfit Evaluation (Rating: 82/100)**

✅ **What works well:**

1. **Occasion Fit** – The sleeveless white tank and blue shorts are perfect for mobility, breathability, and beach vibes.
2. **Athletic Aesthetic** – The reflective sport sunglasses + backward cap combo gives off a sporty, competitive edge.
3. **Color Coordination** – White top + bright blue shorts is clean, fresh, and summery.
4. **Practicality** – Minimal layers, freedom of movement, and lightweight fabric make this functional for volleyball.
5. **Confidence Factor** – The relaxed fit tank with toned arms exposed feels effortless and confident.

Figure 5.2: GPT - 4.0 response for occasion set as : formal ball

🔎 **Outfit Evaluation (Rating: 22/100)**

✅ **What (barely) works:**

1. **Confidence Factor** – You'll definitely stand out and be the best-dressed person at the bowling alley.
2. **Shoes** – Technically leather shoes have smooth soles, but you'd still need to swap them for bowling shoes.
3. **Durability** – The tailored fabric is high-quality and won't rip easily with movement.

Figure 5.3: GPT - 4.0 response for occasion set as: bowling with friends

IV.

Figure 6.1 : Sample image before and after detection (generated by Gemini AI from Google)

🔍 **Outfit Evaluation (Rating: 90/100) – *Occasion: Traditional Festival (Diwali)***

✅ **What works well:**

1. **Cultural Fit** – The saree's traditional draping, jewelry, and vibrant colors are entirely appropriate for Diwali celebrations.
2. **Festive Appeal** – Gold embroidery and jewelry complement the festive mood, creating a rich, celebratory aesthetic.
3. **Elegance** – The combination of saree and ornate accessories enhances grace and aligns with cultural expectations of formality.
4. **Completeness** – Coordinated bangles, earrings, and necklace bring cohesion to the outfit.
5. **Confidence Factor** – The wearer projects confidence and comfort in traditional attire, which enhances overall presence.

Figure 6.2: GPT - 4.0 response for occasion set as: festival diwali

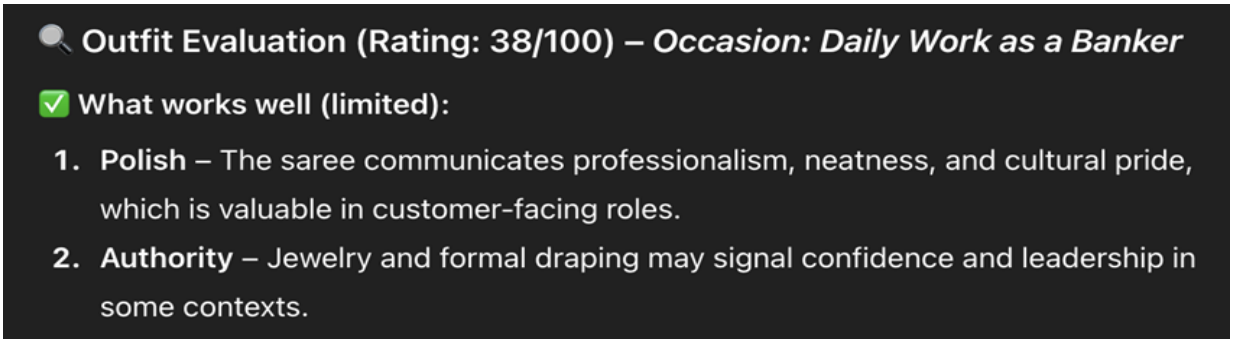

Figure 6.3: GPT - 4.0 response for occasion set as: going to work at a bank

4.2 Case Studies: Correct and Incorrect Context

The second analysis stage combined the object detection output with generative assessment through a large language model (LLM), ChatGPT 4.0 in this work. The detected items' structured JSON representation was fed to the LLM along with explicit occasion contexts. The beach volleyball male image, for example, received the prompt defining a relaxed sporting environment. ChatGPT appropriately suggested outfit matching while indicating small enhancements, like sneakers replaced by sporting shoes more appropriate for sand environments. In contrast, when the same detection inputs were combined with an inappropriate setting, like a business conference, the LLM indicated misalignment, showing its ability to condition reasoning on the environment with flexibility in suggesting.

Likewise, the formal male dress picture (blazer/tuxedo) demonstrated strong contextual fit for formal situations but produced corrections when probed over informal or outdoor recreational contexts, underlining the model's refined evaluation abilities. Examples of female clothes exhibited symmetrical trends: the saree registered well to ceremonial or traditional contexts, while the informal jeans/top/jacket ensemble was accurately rated as unsuitable for formal occasions. These points highlight the need for environment-sensitive prompts and the interaction of detection-based inputs and LLM reasoning. Placeholder figures may represent every image along with LLM output clips. Following is some feedback ChatGPT provided when presented with an image of a woman in casual wear. For 2 different occasions (first to meet for coffee, second to go to the pool).

4.3 Prompt Engineering Efficacy

Prompt engineering the deliberate construction of instructions to control LLM behavior was at the heart of this inquiry. To limit hallucination, prompts were specially conditioned to utilize only the detected objects for testing, and color descriptors as the only other visual cue. This guaranteed that the model avoided making clothing items beyond the detection outcomes, making recommendations more reliable in context-dependent scenarios. Paragraph-based prompts instead of instruction lists allowed for more cohesive descriptions and actionable advice. On all test images, prompt engineering greatly enhanced assessment accuracy, reducing irrelevant or imaginative edits and creating understandable, interpretable rationales for every outfit.

4.4 Strengths of the Pipeline

The hybrid pipeline demonstrated several strengths. First, taking advantage of a pretrained object detection model significantly minimized development time with near-top-grade baseline accuracy for garment identification. Implementing NMS promoted detection reliability, eliminating duplicate errors without compromising vital information. Second, the generative capabilities of LLM provided contextual evaluation and recommendation, fulfilling a significant gap in conventional detection-oriented systems addressing only item recognition or aesthetic compatibility. Thirdly, formal JSON representation supported smooth communication between the outputs of computer vision and inputs of LLM, improving transparency and reproducibility.

The system also exhibited versatility in different types of outfits and social settings, such as casual, and formal wear. By directly modeling appropriateness with respect to the environment, the pipeline fills the

gap between visual perception and situational understanding a shortcoming identified in earlier studies where aesthetic-only assessment prevails.

4.5 Comparative Reflection

Its contribution to occasion-aware fashion assessment is highlighted by comparing the system with existing work. Multi-expert and graph-based assessment systems offer compatibility scores without contextual grounding, whereas retrieval-based generation approaches tend to generate outfits in isolation from the detected garments. Unlike these, the detection + NMS + LLM pipeline infers suggestions grounded on real observed clothes, making them more reliable and interpretable. However, model simplicity, computational efficiency, and contextual accuracy trade-offs need to be weighed carefully in deployment scenarios

V. CONCLUSION

5.1 Overall View

This research offers an exhaustive investigation of AI-based fashion intelligence, that is, combining pretrained object detection models, Non-Maximum Suppression (NMS), and large language models (LLMs) to analyze clothing outfits in contextually relevant environments. The following overarching research inquiry oversaw the design, experimentation, and analysis of this pipeline - How can object detection and deep learning be combined with generative AI to detect and analyze clothing outfits and make contextually relevant recommendations based on the occasion? oversaw the design, experimentation, and analysis of this pipeline.

By integrating visual recognition, attribute parsing, and generative reasoning, the study shows a concrete path for closing the gap between fashion AI for purely aesthetic or category-based and context-sensitive, actionable fashion recommendations for real-world use.

5.2 Summary of Contributions

The study adds to both the theory and practice of AI use in fashion in the following ways. Firstly, it highlights the need to utilize pretrained object detection models, including Hugging Face models, which have strong recognition abilities for a broad variety of garments and accessories. In combination with NMS, the pipeline noticeably eliminates redundant predictions and boosts detection performance, alleviating a long-standing limitation in one-stage detection architectures. Testing four AI-generated images of varying attire types—beach volleyball male, formal male outfit, female saree, and female casual wear evinced the consistency of this method over both casual and formal occasions.

Second, the research emphasizes the importance of formalized communication between detection and generative reasoning. Serialization of detected objects to JSON enables the LLM to process a structured instance-grounded representation of the outfit. By closely constraining prompts to depend solely on detected objects and properties like color, the research counteracts hallucination risk and keeps recommendations anchored to observed reality. This convergence stresses the complementarity of visual recognition and generative AI, facilitating subtle, occasion-savvy reasoning that goes beyond conventional detection-only or style-based pipelines.

Third, the pipeline provides real-world application practicalities. Occasion-aware evaluation is critical in applications from e-commerce styling assistants to personal wardrobe management and inclusive fashion recommendation systems. For instance, the capacity of the LLM to spot misaligned contexts e.g., casual wear in a formal context and propose actionable edits illustrates the potential of AI to augment decision-making for consumers and retailers alike. This dimension fills a significant void in earlier research, wherein outfit suggestions often did not explicitly take cultural or environmental suitability into account.

5.3 Limitations and Challenges

In spite of the advantages of this research, numerous limitations need to be noted. Fine or small object detection, including accessories, was still inconsistent, especially in occluded or cluttered areas. This is an ongoing limitation of current one-stage detection models like YOLO. In addition, the pretrained model was not able to handle culturally specific clothing, like Indian sarees, which shows the requirement of more diverse training sets in fashion object detection.

The pipeline's dependence on LLM reasoning also adds subjectivity. While prompt engineering helps to reduce hallucinations, the outputs are still guided by biases in the pretrained language model. Therefore, recommendations are not always likely to represent global dress norms or consider subtle cultural and

contextual expectations—a significant consideration in inclusive fashion use cases.

Out of the 4 given examples, the 4th image in specific and its analysis given by figure 6.1 shows a woman in Indian traditional wear called a ‘saree’. The model failed to analyse this well as the database from which it was trained lacked this kind of apparel. Traditional and cultural clothing like this is one of the models’ biggest holes.

The experimental assessment was intentionally restricted to a few test cases of AI-generated outfits (two men's and two women's in different situations shown). Though these were enough for the demonstration of the system's capabilities, they lack statistically significant validation. Quantitative measures like Mean Average Precision (mAP) or F1 scores were taken into account qualitatively, not systematically. That is, the results must be interpreted as exploratory, not generalizable.

Ultimately, though color extraction is informative, it fails to extract more subtle characteristics like texture of the fabric, the material's quality, or finer patterns. These exclusions could negatively impact recommendation fidelity, which may imply that multimodal methods utilizing more complete sensory information would help fortify subsequent versions of this system.

5.4 Future Work

Based on the results, some avenues for future work become apparent. First, using bigger and more varied datasets, including culturally dependent clothing and international fashion trends, would enhance model generalizability and diminish bias. Second, delving deeper into multimodal integration beyond color and category attributes, including texture, material, or user preference information, might enhance recommendation fidelity. Third, the use of human-in-the-loop evaluation processes, in which end users occasionally give feedback regarding occasion suitability, could enable iterative improvement and preference learning, balancing technical competence and real-world usability.

Subsequent versions might also explore including diffusion-based generative models for recommending outfits, allowing for more imaginative, creative suggestions without losing grounding in garments detected. Lastly, adapting prompt engineering methods to incorporate adaptive context reasoning and multi-turn conversational capabilities would improve the model's ability to deal with sophisticated queries, including coordinating an outfit across different events or combining formal and informal elements.

5.5 Final Remarks

In summary, this work showcases synergistic AI pipelines that combine object detection, NMS, and LLM-based generative assessment to provide contextually sensitive fashion advice. The approach fills gaps found in previous work by being instance-grounded for detection, considering social context outright, and making explicit reasoning available for actionable editing. Though there are still limitations, such as detection granularity, cultural reach, and small-sample assessment, the work provides a strong foundation for both research investigation and real-world deployment. In marrying computer vision with generative AI and well-designed prompts, the framework demonstrates a compelling direction toward smart, occasion-sensitive fashion systems that can assist consumers, stylists, and retailers in a subtle, consistent, and understandable way.

REFERENCES

[1] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection. 2020, arXiv:2004.10934.

[2] He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. 2017, arXiv:1703.06870.

[3] Wu, Yuxin, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. Facebook AI Research, 2019.

[4] Tan, Mingxing, and Quoc V. Le. EfficientDet: Scalable and Efficient Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[5] Carion, Nicolas, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers (DETR). European Conference on Computer Vision (ECCV), 2020.

[6] Lao, B., and K. Jagadeesh. Convolutional Neural Networks for Fashion Classification and Object Detection. 2015, ResearchGate.

[7] Liu, Ziwei, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering Robust Clothes Recognition and Retrieval. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[8] Liu, Ziwei, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Fashion Landmark Detection in the Wild. Proceedings of the European Conference on Computer Vision (ECCV), 2016.

[9] Li, Yuncheng, Liangliang Cao, Jiang Zhu, and Jiebo Luo. Mining Fashion Outfit Compatibility with Graphs. Proceedings of the ACM on Multimedia Conference, 2017.

[10] [10] Itkare, Swapnil, and Amol Manjaramkar. Fashion Classification and Object Detection Using CNN. International Journal of Research in Engineering, Science and Management, vol. 4, no. 5, 2021, pp. 73–76.

[11] Hsiao, Wei-Lin, Krishna Kumar Singh, and Yong Jae Lee. Fashion++: Minimal Edits for Outfit Improvement. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[12] Cheng, Wen-Huang, Shang-Fu Song, Chia-Yen Chen, Sy-Yen Hidayati, and Jiaying Liu. Fashion Meets Computer Vision: A Survey. ACM Computing Surveys, vol. 54, no. 4, 2021, pp. 1–36.

[13] Liu, Ziwei, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[14] Ge, Yuying, Runzhong Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation, and Re-Identification. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[15] Zhao, Bo, Yunchao Fu, Yuxiao Jiang, and Zhizhong Lin. Fashionpedia: Ontology, Segmentation, and Attribute-Annotated Dataset. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[16] Matzen, Kevin, Kavita Bala, and Noah Snavely. StreetStyle: Exploring World-Wide Clothing Styles from Millions of Photos. Proceedings of the ACM on Transactions on Graphics (TOG), vol. 36, no. 4, 2017.

[17] Fang, Hao, Yucheng Zhao, Xiaosheng Yan, Jianan Wang, and Yu Qiao. "YOLOS: You Only Look at One Sequence for Object Detection." Advances in Neural Information Processing Systems (NeurIPS), 2021. arXiv:2106.00666