Journal of Design Intelligence
https://www.ffspub.com/index.php/jdi
ISSN 3104-9001 (print)
E-mail: jdi.office@ffspub.com

Article

Research on Fashion Contextual Style Recognition Methods Based on Scene Graph Technology

Yuming Mai, Xin Zhao, Yan Hong†
The College of Textile and Clothing Engineering, Soochou University, Suzhou 215021, China
†*E-mail: yannichonghk@gmail.com*

Received: May 29, 2026 / Revised: June 15, 2026 / Accepted: June 18, 2026 / Published online: June 26, 2026

Abstract: Driven by the intelligent and personalized transformation of the garment industry, fashion contextual style recognition has emerged as a pivotal technology bridging visual understanding and business decision-making. However, traditional methods anchor styles to isolated garment attributes, failing to capture the deep semantics of entity interactions in complex visual scenes. To address these limitations, we propose a fashion contextual style recognition method based on unbiased Scene Graph Generation (SGG) and hierarchical context fusion. We developed a hierarchical context-aware network that deconstructs fashion contexts into three dimensions: entity context, global context, and scene context, integrated through graph embedding. Furthermore, a Causal Intervention mechanism based on Total Direct Effect (TDE) is introduced to mitigate the long-tail bias inherent in unconstrained scenes. Experimental results on the Visual Genome dataset demonstrate that our approach significantly outperforms baseline models in recognition accuracy and robustness. This research extends the definition of fashion style from visual attributes to holistic situational semantics, providing a new perspective for intelligent fashion analysis.

Keywords: Scene Graph Generation; Fashion Contextual Style Recognition; Unbiased Scene Graph; Hierarchical Context; Personalised Recommendations
https://doi.org/10.64509/jdi.13.115

1 Introduction

The fashion industry is undergoing a transition from automated production toward human-centered intelligence and adaptive digital services, where personalized recommendation has become increasingly dependent on contextual visual understanding [1]. Within this transformation, Fashion contextual style recognition has emerged as a key task connecting semantic image understanding with recommendation-oriented decision-making. Yet despite recent advances in computer vision and multimodal learning, current fashion analysis systems still demonstrate limited contextual adaptability. Accurately identifying fashion styles from complex visual environments therefore remains a challenging problem at the intersection of artificial intelligence and fashion intelligence.

Most existing fashion style recognition methods treat style as an intrinsic property of isolated garments. Early studies such as Style2Vec modeled fashion items through distributional semantic learning, regarding clothing items as "words" and outfit combinations as "contexts" to learn latent style embeddings [2]. Although this strategy captures implicit semantics related to color, texture, and pattern, its analytical unit remains discrete clothing items rather than the complete visual situation presented in an image. Design Issue Graph (DIG) further introduced graph-based modeling to capture co-occurrence relationships among garment attributes [3]. However, its reasoning process still centers primarily on clothing attributes themselves, making it difficult to represent the broader contextual atmosphere formed by interactions among people, objects, and environments.

This limitation largely originates from the isolated recognition paradigm widely adopted in visual analysis research, where models tend to emphasize object-level appearance while overlooking semantic interactions among multiple entities and the environmental context shaping human style perception [4, 5]. From a cognitive perspective, Fashion contextual style recognition is not determined solely by garments worn by an individual. Instead, style perception emerges from the joint configuration of people, surrounding objects, physical environments, and social interactions [6]. The same suit

† Corresponding author: Yan Hong
*Academic Editor: Zhaohui Wang

may convey entirely different stylistic meanings in an office scene and on a beach. Such differences are not caused by changes in the garment itself, but by contextual semantics reconstructed through environment, activity, and social identity. Context therefore functions not as auxiliary information, but as an essential bridge connecting fashion semantic analysis and personalized recommendation [7].

The development of Scene Graph Generation (SGG) provides a possible pathway toward structured semantic understanding of fashion contexts. By representing image content as relational triplets composed of entities, attributes, and predicates, SGG transforms visual scenes into explicit graph structures that support higher-level semantic reasoning beyond pixel representations [8, 9]. This capability has demonstrated effectiveness in tasks such as visual question answering and affective scene understanding [10]. Nevertheless, directly applying existing SGG frameworks to Fashion contextual style recognition still faces two major challenges. First, predicate prediction in scene graphs is heavily affected by long-tail distribution bias. Models tend to predict high-frequency but semantically weak predicates such as "on" or "has," while struggling to identify context-sensitive relations such as "walking on beach" or "sitting in office," which are more important for style interpretation [11]. Second, most current graph reasoning architectures lack effective mechanisms for integrating hierarchical contextual semantics. As a result, they cannot jointly model local entity appearance, global environmental atmosphere, and relational interaction logic within a unified contextual representation [10].

To address these limitations, this paper proposes a Fashion contextual style recognition framework based on unbiased Scene Graph Generation and hierarchical context fusion. Rather than treating clothing as isolated recognition targets, the proposed framework incorporates all entities appearing in an image, together with their interaction relationships and global visual atmosphere, into a unified structured contextual representation. Specifically, we first construct a hierarchical context-aware network that decomposes fashion scenes into three complementary dimensions: entity context, global context, and scene context. These representations are then integrated through graph embedding and adaptive feature fusion. Next, to reduce relational bias in scene graph reasoning, a Causal intervention mechanism based on Total Direct Effect (TDE) is introduced to suppress spurious contextual correlations during predicate prediction [11]. In addition, self-attention and MLP-Mixer modules are employed to strengthen phrase-level relational representation learning under complex unconstrained scenes [12].

The major contributions of this paper are summarized as follows:

1) We propose a Fashion contextual style recognition framework based on unbiased Scene Graph Generation and hierarchical context fusion, extending style recognition from isolated garment attributes toward structured contextual semantics jointly constructed by people, objects, and environments.

2) A hierarchical context representation architecture is developed to jointly model entity context, global context, and scene context, enabling collaborative semantic reasoning from local appearance, environmental atmosphere, and relational interactions.

3) A Causal intervention strategy based on Total Direct Effect (TDE) is incorporated into Scene Graph Generation to alleviate long-tail predicate bias and improve relational semantic discrimination under unconstrained scenes [11].

4) Comprehensive experiments conducted on the Visual Genome dataset demonstrate the effectiveness of the proposed framework through comparative evaluations and ablation studies under complex contextual conditions.

The overall technical roadmap of the proposed framework, including problem definition, methodological design, experimental evaluation, and future application directions, is illustrated in Figure 1.

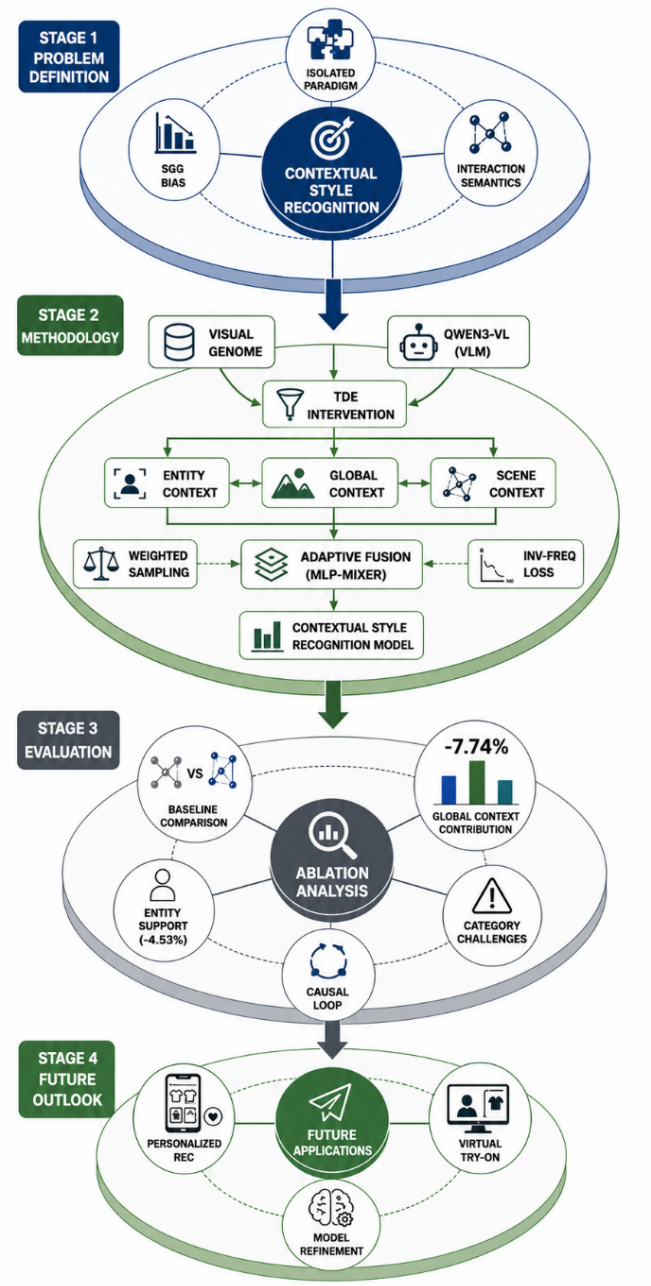

Figure 1: Technology roadmap of the proposed framework. While stage 1 corresponds to chapters 1 and 2, the remaining stages corresponds to each following chapters.

The remainder of this paper is organized as follows: Section 2 reviews related studies on Fashion contextual style recognition, Scene Graph Generation, and hierarchical contextual reasoning. Section 3 introduces the proposed framework and its methodological details. Section 4 presents the experimental settings and evaluation results. Finally, Section 5 concludes the paper and discusses future research directions.

2 Related Work

2.1 Fashion Contextual Style Recognition

Whether fashion style originates from visual garment attributes or from contextual semantics remains central to understanding the nature of style itself [13]. If style were regarded solely as an intrinsic property of clothing items, then the same suit should convey identical aesthetic meaning in both an office and a beach environment. Yet in practical perception, these two situations produce substantially different stylistic impressions. This difference does not arise from changes in the garment itself, but from contextual semantics jointly reconstructed through social identity, physical environment, and human activity. Consequently, approaches that anchor style recognition exclusively to clothing items often fail to capture the full semantic meaning of fashion style.

Most existing fashion image analysis studies still focus primarily on garment classification or attribute recognition, implicitly assuming that style can be directly inferred from visual properties such as color, texture, or silhouette. Early methods such as Style2Vec learned latent style representations through distributional semantic modeling and demonstrated effectiveness in style classification and analogy reasoning tasks [2]. However, the analytical unit of these methods remains discrete clothing items rather than the complete visual context represented in an image. With the development of multimodal recommendation systems, models such as PSGNet further incorporated personal preference prediction and compatibility learning into fashion recommendation frameworks, thereby improving personalized recommendation performance across different scenarios [14]. Nevertheless, the primary objective of these studies remains recommendation optimization rather than direct recognition of overall Fashion contextual style semantics.

DIG introduced graph-based reasoning into fashion analysis by modeling co-occurrence relationships among garment attributes through graph convolutional learning [3]. Although graph structures improved semantic association among clothing attributes, the graph nodes themselves were still restricted to garment-level representations. As a result, interactions between people and surrounding environmental objects could not be effectively modeled. This limitation becomes increasingly evident under unconstrained scenes, where contextual semantics often play a dominant role in style perception.

Building upon these observations, this study defines Fashion contextual style recognition as the task of identifying the socially meaningful overall style category jointly formed by people, garments, environments, and their interaction relationships within an image. Unlike traditional garment-centered recognition paradigms, the proposed task treats the entire image as the analytical unit and emphasizes the collaborative influence of spatial context, environmental atmosphere, and social interaction cues on style perception. The key challenge therefore lies in transforming fragmented pixel-level information into structured semantic representations while enabling effective fusion across multiple contextual levels. This requirement naturally motivates the introduction of Scene Graph Generation and hierarchical contextual reasoning.

2.2 Fashion Contextual Style Recognition Based on Scene Graph Embedding

A major bottleneck in Fashion contextual style recognition lies in converting low-level visual signals into structured semantic representations capable of describing interactions among people, garments, and environments. Scene Graph Generation (SGG) provides a feasible solution to this problem. As a structured visual representation framework, a scene graph models image content as a directed graph composed of entity nodes, attribute labels, and relational edges. Its fundamental representation unit is the semantic triplet <subject–predicate–object>, such as <person–wearing–suit> [8, 9]. Unlike conventional convolutional neural networks that primarily capture local texture or shape information, scene graphs explicitly encode semantic relationships among image elements, thereby supporting higher-level visual reasoning [9].

Within the fashion domain, the value of scene graphs lies in their ability to represent complex interactions among people, garments, accessories, and surrounding environments. In a street-fashion image, for example, the "wearing" relationship between a person and clothing, the "matching-with" relationship between garments and accessories, and the "located-in" relationship between individuals and backgrounds together construct the semantic logic underlying style perception. Previous studies suggest that integrating structured relational semantics with visual appearance can produce more robust semantic embeddings than relying solely on visual features [15] . Scene graph representations therefore provide important semantic support for understanding "what is worn" under "which contextual condition."

However, the raw output of Scene Graph Generation consists of discrete relational triplets, which cannot be directly applied to style classification tasks. Graph embedding techniques are introduced to address this issue by mapping graph nodes and edges into low-dimensional continuous vector spaces while preserving topological and semantic structures. In the context of Fashion contextual style recognition, scene graph embedding transforms detected entities, attributes, and relational predicates into unified semantic representations, thereby constructing a contextual semantic space containing fashion-related interactions. Figure 2 represents the workflow of graph embedding network.

Graph neural networks further enable each entity node to aggregate information from neighboring nodes through message-passing mechanisms, producing embeddings that jointly encode local visual appearance and global relational semantics [16]. These embeddings can subsequently be fed into downstream classifiers to determine the overall Fashion

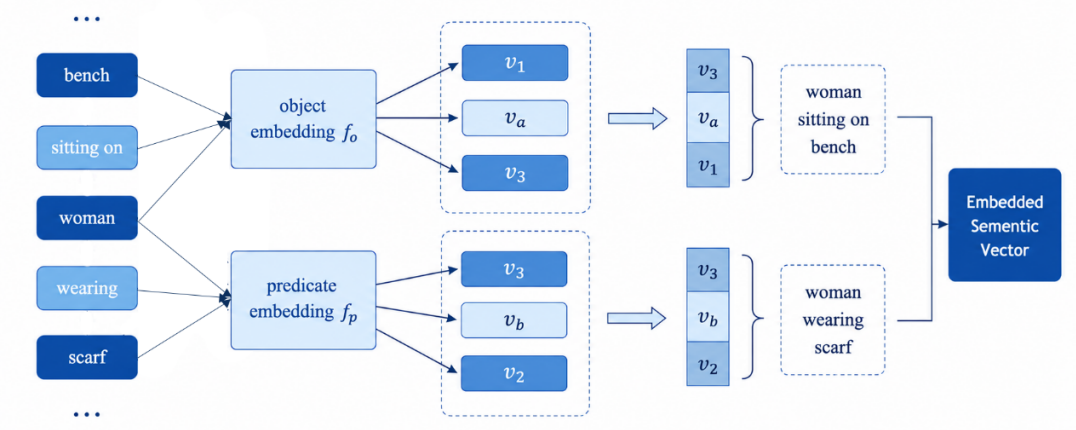

Figure 2: Graph Embedding Network

contextual style category of an image. Nevertheless, directly applying conventional graph embedding frameworks to fashion scenarios still encounters significant limitations. Most existing methods emphasize structural relationship modeling while insufficiently capturing hierarchical contextual semantics, particularly the interaction between local entities, global atmosphere, and scene-level relational logic. This issue motivates the introduction of hierarchical context reasoning frameworks.

2.3 Hierarchical Context Recognition Framework

In visual contextual understanding tasks, extracting complementary semantic information from multiple levels of an image is essential for improving recognition performance. Wu *et al.* proposed a hierarchical context recognition framework that decomposes image semantics into three interconnected levels: entity context, global context, and scene context [10]. In this study, we extend this framework to the task of Fashion contextual style recognition.

Entity context focuses on the fine-grained local representations of salient objects within an image and can be defined as visual features extracted from object regions through detection-based localization techniques [10]. In Fashion contextual style recognition, entity context provides the material basis of style semantics. These entities may include people, garments, accessories, furniture, architectural components, or environmental objects contributing to stylistic perception. For instance, a street-fashion image may simultaneously contain entities such as dresses, handbags, wooden tables, and decorative lighting. Their local visual characteristics — including tailoring structures, material textures, and reflective properties — jointly contribute to the recognition of styles such as urban casual or modern vintage.

Global context captures the macroscopic atmosphere of the entire image and can be defined as environmental representations extracted through global pooling operations over the full scene [10]. Such representations typically encode overall illumination, color distribution, contrast, and spatial composition. Within Fashion contextual style recognition, global context provides the atmospheric foundation of style semantics. A "bright and minimalist" style, for example, is often characterized not by a single object, but by globally consistent visual properties such as high brightness, low saturation, and uniform lighting distribution. Global context and entity context therefore exhibit strong complementarity: the former describes the overall visual impression of a scene, whereas the latter specifies the concrete objects contributing to that impression.

Scene context focuses on interaction logic among entities and can be represented through structured semantic triplets extracted using Scene Graph Generation techniques [10]. In Fashion contextual style recognition, scene context contributes the relational logic underlying stylistic interpretation. The same entity set — such as "person," "sofa," and "coffee cup" — may convey entirely different style semantics depending on their interaction relationships. A scene described by <person–sitting formally on–sofa> and <person–holding–coffee cup> may suggest a business discussion style, whereas <person–leaning on–sofa> and <person–placing casually–coffee cup> may indicate a relaxed leisure style. In other words, style perception is often determined less by the existence of entities themselves than by the way these entities interact within a scene.

Despite its effectiveness, scene-level reasoning still faces a serious challenge caused by long-tail distribution bias in predicate prediction. Existing models tend to over-predict high-frequency generalized predicates such as "on" or "has," while neglecting semantically discriminative relations such as "leaning on" or "holding," which are more important for contextual style interpretation [11] . This issue motivates the introduction of unbiased Scene Graph Generation methods based on Causal intervention and Total Direct Effect (TDE).

3 Proposed Methodology for Fashion Contextual Style Recognition

3.1 Task Definition and Unbiased Scene Graph Generation

This research focuses on the task of fashion contextual style recognition. Given a single RGB image I from a natural scene, the objective is to predict the fashion contextual style category y from a predefined set $\mathscr{Y} = \{1, 2, \ldots, C\}$ (where $C = 10$). The prediction process is formulated as:

$$y = \arg\max_{c \in \mathcal{Y}} p(c \mid I) \tag{1}$$

where $p(c \mid I)$ denotes the probability that image I belongs to category c. Unlike conventional fashion recognition that targets specific clothing items, our approach shifts the focus toward the holistic fashion style emerging from the interaction of people, garments, and environmental entities.

To achieve this, we developed a two-stage recognition framework. The first stage involves extracting structured contextual representations, where object detection and Scene Graph Generation (SGG) are utilized to identify entity sets, relationship candidates, and global visual features. The second stage then integrates these structured inputs through hierarchical feature fusion and classification.

A primary challenge in SGG is the long-tail distribution of relational predicates in training data. In the Visual Genome (VG) dataset, for instance, over 90% of annotations consist of geometric or possessive predicates, while semantically rich relationships essential for style characterization remain scarce [11]. To mitigate this, we adopt an unbiased SGG approach based on a causal inference framework. During the inference phase, predicate scores are calculated using the Total Direct Effect (TDE) [11]:

$$\mathrm{TDE} = R_{X,O}(I) - R_{X,O}(I_{\mathrm{ref}}) \tag{2}$$

where $R_{X,O}(I)$ represents the predicate prediction score based on the original image I, and $R_{X,O}(I_{\mathrm{ref}})$ denotes the score after counterfactual intervention on object features. This subtraction effectively eliminates contextual bias stemming from object co-occurrence frequencies. Consequently, the final predicate scores rely more heavily on genuine visual interactions. In this study, unbiased SGG facilitates the generation of relationship triplets—such as <person–leaning on–sofa> instead of <person–on–sofa>—forming the backbone of scene context.

3.2 Hierarchical Context Feature Extraction

Building upon unbiased SGG, we constructed a hierarchical context feature extraction module to capture complementary semantic representations across three dimensions: entity, global, and scene contexts. As illustrated in Figure 3, the module takes the scene graph generated by the unbiased SGG as input. Through three parallel branches, it produces the entity context h_{ec}, global context h_{gc}, and scene context h_{sc}.

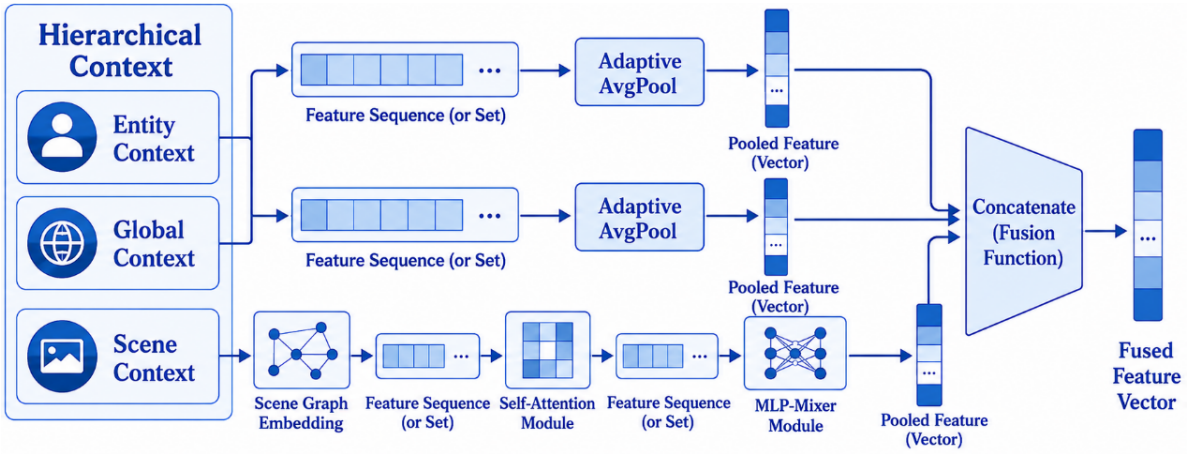

Figure 3: Hierarchical context extraction and fusion network

3.2.1 Entity Context Extraction

The entity context aims to capture fine-grained local features of salient entities, including individuals, clothing items, and environmental objects. We utilized Faster R-CNN with a Feature Pyramid Network (FPN) as the backbone detector, fine-tuned on the Visual Genome dataset. For an input image I, the detector yields N candidate entities, each defined by a bounding box $b_i \in \mathbb{R}^4$, a class label l_i, and a corresponding Region of Interest (RoI) feature map m_i. To construct h_{ec}, we applied adaptive average pooling to the RoI features to obtain fixed-length vectors, which were then mapped to a 256-dimensional space via a fully connected layer. The final entity context matrix $h_{ec} \in \mathbb{R}^{N\times 256}$ was formed by stacking these vectors. To ensure comprehensive coverage of key objects, we retained all entities with a detection confidence above 0.5.

3.2.2 Global Context Extraction

Global context captures the overall visual atmosphere, including macro attributes such as color tone, lighting, contrast, and spatial layout. Specifically, our implementation leverages the multi-scale feature maps P_2, P_3, P_4, P_5 from the FPN, characterized by a constant channel count of 256. After normalizing the spatial resolutions via adaptive average pooling, these maps were concatenated along the channel dimension to form h_{gc}^{raw}. To reduce computational complexity while extracting high-level semantics, we employed a convolutional layer (kernel size 3, stride 2) followed by a global max pooling layer. This yielded a global context vector $h_{gc} \in \mathbb{R}^{512}$, providing the atmospheric foundation for style recognition.

3.2.3 Scene Context Extraction

The scene context constructs structured semantic representations based on the relationship triplets generated by the

unbiased SGG. For the N detected entities, the model predicts predicates for each pair, forming a set of triplets (s, p, o). To maintain computational efficiency, only the top K (where $K = 200$) triplets by confidence score were retained.

To transform these discrete triplets into differentiable representations, we designed a graph embedding network. First, entity and predicate categories were mapped to learnable embeddings, $v_{o_i} \in \mathbb{R}^{32}$ and $v_p \in \mathbb{R}^{32}$ respectively. For each triplet (s, p, o), we concatenated the subject, predicate, and object embeddings into a phrase-level vector:

$$x_{spo} = [v_{o_s}; v_p; v_{o_o}] \in \mathbb{R}^{96} \tag{3}$$

These vectors were stacked to form the matrix $X \in \mathbb{R}^{K\times 96}$. We then applied a Self-Attention layer and an MLP-Mixer module [12] to this matrix. While the attention layer captures dependencies between triplets, the MLP-Mixer blends information across sequence and feature dimensions:

$$X' = \text{SelfAttention}(X), \quad h_{sc} = \text{MLPMixer}(X') \tag{4}$$

The resulting $h_{sc} \in \mathbb{R}^{256}$ aggregates the semantic logic of the entire scene graph, accentuating discriminative cues such as the difference between <person–leaning on–sofa> and <person–sitting on–office chair>.

3.2.4 Coordination of Hierarchical Features

The three components—h_{ec}, h_{gc}, and h_{sc}—collectively characterize the fashion context through local details, holistic atmosphere, and relational logic. This hierarchical approach ensures that the multifaceted nature of fashion style is preserved. The subsequent integration and classification of these features are detailed in Section 3.5.

3.3 Construction of Fashion Contextual Style Data

The realization of fashion contextual style recognition necessitates a dataset enriched with structured situational annotations. We utilize Visual Genome (VG) as our primary data source, as its dense entity and relationship annotations provide the essential substrate for Scene Graph Generation tasks [8]. However, the raw VG dataset contains hundreds of predicate categories—many of which are irrelevant to fashion—and lacks explicit style labels. To address these deficiencies, we performed data preprocessing at two distinct levels.

3.3.1 Predicate Filtering and Fashion Contextual Remapping

Statistical analysis by Tang et al. reveals that over 90% of relationship annotations in Visual Genome are geometric or possessive, leaving semantically rich interactions underrepresented [11]. To refine this data, we implemented a multi-step filtering and remapping protocol.

First, irrelevant predicates were pruned. This involved removing purely geometric terms (e.g., behind, next to, above) and overly generalized possessive terms that provide negligible stylistic information. Second, we retained and categorized 30 fashion-relevant semantic predicates into three distinct subsets: (1) Dressing and Coordination, such as wearing, matching-with, and contrasting-with, which describe the composition between people and garments; (2) Handling and Carrying, including carrying and holding, to capture interactions with accessories; and (3) Spatial and Pose, such as leaning_on, walking_on, and sitting_on, which characterize the relative positioning and posture of entities. Finally, synonymous predicates were standardized into unified forms.This distillation process reduced the average number of triplets per image by approximately 40%. Crucially, however, it significantly increased the information density of the remaining predicates for style discrimination. The predicate filtering and remapping procedure is illustrated in Figure 4.

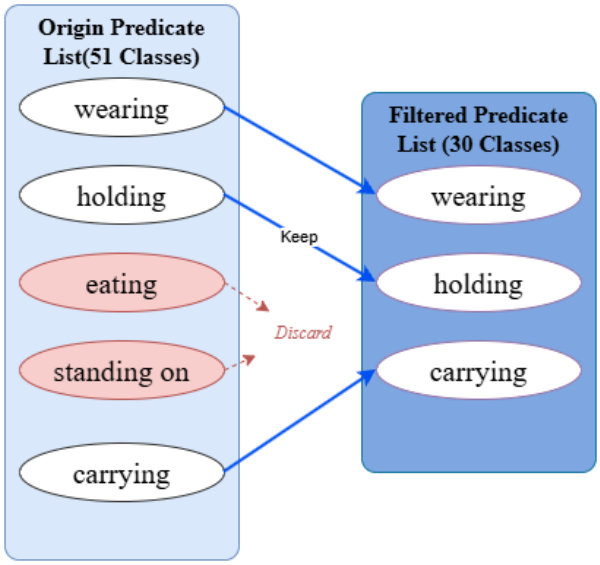

Figure 4: Predicate Set Mapping

3.3.2 Fashion Contextual Style Annotation

Since the original Visual Genome data lacks explicit style metadata, we supplemented it with supervised fashion contextual style signals. The annotation pipeline was engineered to rigorous methodological principles.

Regarding the definition of style categories, as illustrated in Table 1, we deconstructed the fashion contextual style into two complementary dimensions: Environmental Atmosphere and Scene Graph Nodes. Drawing upon the five-class clothing style taxonomy introduced by Kiapour *et al.* [7], we adapted the framework to accommodate the diverse unconstrained scenarios of the VG dataset, the holistic visual attributes of the images, and the frequency distribution of scene graph elements. Following a comprehensive evaluation by a panel of three annotators with expertise in fashion design, we established 10 distinct fashion contextual style categories that balance aesthetic rationale with empirical data support. The environmental atmosphere, scene graph nodes, and naming rationale for each category are detailed in Table 1. This bi-dimensional classification paradigm provides the foundational theoretical scaffolding for the hierarchical context-aware network proposed in this study.

Two strict constraints were enforced during the annotation design phase:

Table 1: Fashion contextual style classification and rationale

Style Tag	Environmental Atmosphere	Scene Graph Nodes	Naming Rationale
Candid/Humanistic	Candid streetscapes and real-life everyday moments	Human figures, interactive gestures, and street elements	Narrative spontaneity
Kinetic/Dynamic	Athletic settings, rapid motion, and dynamic lighting	Sports equipment, stadiums, and directional motion lines	Physical tension and energy
Maximalist/Eclectic	Ornate ornamentation, contrasting colors, and layered styling	Mixed textile patterns, dense accessories, and asymmetry	Visual overload and aesthetic remix
Minimalist/Modern	Stripped-back spaces, low-saturation tones, and geometric lines	Solid backgrounds, unadorned garments, and simple geometry	Pure form and spatial deconstruction
Urban/Modern	City skylines, steel structures, and glass facades	Architectural contours, street signs, and transit infrastructure	Modern urban rhythms and geometry
Luminous/Airy	Translucent lighting, high-key exposures, and open atmospheres	Windows, expansive skies, and soft reflections	Weightlessness through light and shadow
Industrial/Brutalist	Exposed raw materials, cool color temperatures, and rough textures	Concrete surfaces, metal fixtures, and industrial shadows	Stark, unadorned industrial aesthetics
Vintage/Nostalgic	Warm filters, antique decor, and nostalgic color palettes	Weathered furniture, dark wood, and retro appliances	Symbolic nostalgia
Intimate/Cozy	Domestic interior environments and soft material textures	Sofas, rugs, books, and warm domestic textiles	Safe, private, and relaxing narratives
Pastoral/Organic	Natural vegetation, wide-open vistas, and earthy color palettes	Foliage, floral elements, organic soil, and natural landscapes	Harmony between humans and nature

1) Holistic Evaluation: Annotations target the entire image rather than isolated garments. For instance, an image depicting a subject in utilitarian workwear posed against a raw concrete wall is classified under Industrial/Brutalist, even if the clothing itself does not strictly adhere to industrial aesthetics, because the environmental texture dictates the overall situational style.
2) Exclusion of Ambiguous Classes: We deliberately omitted an "Other/Unknown" category. This constraint prevents the model from defaulting to a vague catchment class for borderline samples, thereby forcing the network to learn the highly discriminative features of each style and sharpening the classification boundaries.

Under these guidelines, we leveraged the Qwen3-VL vision-language model for initial style annotation on the Visual Genome dataset. During inference, the model synthesizes image visual content with existing VG structural metadata, including object bounding boxes, category tags, and relationship triplets, enabling a joint semantic understanding of both appearance and scene structure. The annotation process is further guided by a set of 1,000 human-annotated examples used for few-shot calibration, ensuring that the model predictions are consistently aligned with domain-specific stylistic interpretations.

This annotation pipeline operates under a fixed inference configuration, which ensures consistency of generated labels across the entire dataset. As a result, Qwen3-VL is able to produce dense style annotations for the full Visual Genome corpus in a deterministic and structured manner without requiring repeated stochastic sampling.

The final sample distribution across the 10 style categories is summarized in Table 2. The dataset exhibits a pronounced long-tail characteristic, where the majority class (Candid/Humanistic) accounts for approximately 25.9% of the corpus, whereas the rarest class (Maximalist/Eclectic) represents only 1.5%. To mitigate this imbalance and prevent gradient starvation in minority categories, we introduce a dual-level balancing strategy, which is detailed in Section 3.4.

Table 2: Empirical distribution of fashion style labels

Style Tag	Instance Count	Percentage (%)
Minimalist/Modern	2,676	2.48%
Maximalist/Eclectic	1,715	1.59%
Luminous/Airy	4,628	4.28%
Vintage/Nostalgic	10,294	9.53%
Industrial/Brutalist	1,828	1.69%
Pastoral/Organic	16,101	14.90%
Intimate/Cozy	15,404	14.25%
Urban/Modern	11,090	10.26%
Kinetic/Dynamic	16,296	15.08%
Candid/Humanistic	28,038	25.94%

3.3.3 Dataset Splitting

Following annotation, we employed a stratified sampling strategy to partition approximately 100,000 images into training, validation, and test sets. This ensures that the distribution of style categories remains consistent across all subsets. Furthermore, we applied a relational validity filter, retaining only those samples containing at least one filtered fashion-contextual predicate. The final intersection resulted in a dataset comprising 75,653 images for training, 4,115 for validation, and 22,680 for testing.

3.4 Optimization Strategy for Class Imbalance

The 10 defined fashion style labels exhibit a pronounced long-tail distribution. We define a globally consistent imbalance handling strategy operating at both sampling and optimization levels. To mitigate the resultant bias, we implemented a unified weight-based oversampling strategy during the data loading phase. For each category c, given its sample count n_c, the total samples N_{total}, and a predefined balance threshold T (set to 0.05 in all experiments), the sampling weight $W_{\text{sample},c}$ is calculated as follows:

$$W_{\text{sample},c} = \begin{cases} \min\left(W_{\max}, \dfrac{T \cdot N_{\text{total}}}{n_c}\right), & \text{if } \dfrac{n_c}{N_{\text{total}}} < T, \\ 1.0, & \text{otherwise.} \end{cases} \tag{5}$$

where $W_{\max} = 100$ serves as a global weight ceiling shared across all categories and all experimental settings, without any category-specific tuning. These weights are passed to a weighted random sampler, which performs sampling with replacement during each epoch. This mechanism ensures that rare styles receive a higher frequency of exposure.

At the algorithmic level, we further introduce a complementary loss-level reweighting mechanism, forming a unified sample-loss dual-balancing strategy. Specifically, in addition to oversampling rare categories, we compute an inverse frequency-based class weight:

$$W_{\text{loss},c} = \frac{\dfrac{N_{\text{total}}}{C \cdot n_c}}{\dfrac{1}{C}\sum_{i=1}^{C} \dfrac{N_{\text{total}}}{C \cdot n_i}} \tag{6}$$

where $C = 10$ denotes the total number of fashion style categories. All class weights are derived from a single global empirical frequency distribution computed over the entire training set, and are consistently applied across all training runs without per-class or per-split adjustment.

Integrating these normalized weights into the cross-entropy loss yields the weighted loss function:

$$\mathscr{L}_{\text{style}} = -\frac{1}{B}\sum_{i=1}^{B} W_{\text{loss},y_i} \log p_{i,y_i} \tag{7}$$

where B is the batch size, y_i is the ground truth label, and p_{i,y_i} is the predicted probability. By combining oversampling at the data level and inverse-frequency reweighting at the optimization level, the framework forms a unified dual-balancing mechanism for long-tail mitigation.

By construction, all hyperparameters involved in the imbalance handling strategy, including $T = 0.05$, $W_{\max} = 100$, and $C = 10$, are globally fixed and consistently applied across all experiments without category-specific or dataset-split-specific tuning.

3.5 Overall Model Architecture

The integrated architecture of our proposed method is synthesized in Figure 5, and the detailed algorithmic procedure is summarized in Algorithm 1.

Algorithm 1 Hierarchical Context Extraction with Unbiased Scene Graph Generation

Require: RGB image I, pretrained SGG model M_{sgg}, optional label set Q, annotation b_{gt}
Ensure: Entity context h_{ec}, global context h_{gc}, scene context h_{sc}

Extract multiscale feature maps $P = \{P_2, P_3, P_4, P_5\}$ using ResNeXt-101-FPN
$P \leftarrow \text{CooNet}(I)$
Compute global context:
$h_{gc} \leftarrow \text{AdaptiveAvgPool}(P)$
Extract region proposals:
$\{M, B, L\} \leftarrow \text{RoIAlign}(P, \text{RPN}(I, P))$
Scene Context Extraction
for each proposal i **do**
    Compute object feature: $x_i \leftarrow \text{BiLSTM}(m_i, b_i, l_i)$
    Predict refined label $o_i \leftarrow x_i$
end for
Predict pairwise predicates with TDE:
$R_x(I), R_x(I_{ref}) \leftarrow M_{sgg}$
$TDE = R_x(I) - R_x(I_{ref})$
Construct scene context:
$h_{sc} \leftarrow \{O, R\}$
if b_{gt} is available **then**
    Compute IoU between b_{gt} and proposals B
    $S_i \leftarrow \frac{\text{IoU}(b_i, b_{gt})}{\sum_j \text{IoU}(b_j, b_{gt})}$
    Select best proposal:
    $h_{ec} \leftarrow m_k,\ k = \arg\max(S)$
else
    Rank proposals by logits O
    $k \leftarrow 1$
    for $i = 1 \ldots n$ **do**
        if $o_i \in Q$ **then**
            $k \leftarrow i$
            break
        end if
    end for
    $h_{ec} \leftarrow m_k$
end if
return h_{ec}, h_{gc}, h_{sc}

The computational pipeline begins with image preprocessing, where input images are fed into the unbiased Scene Graph Generation module. Utilizing the TDE mechanism through counterfactual reasoning, the system actively eliminates contextual bias derived from object co-occurrence

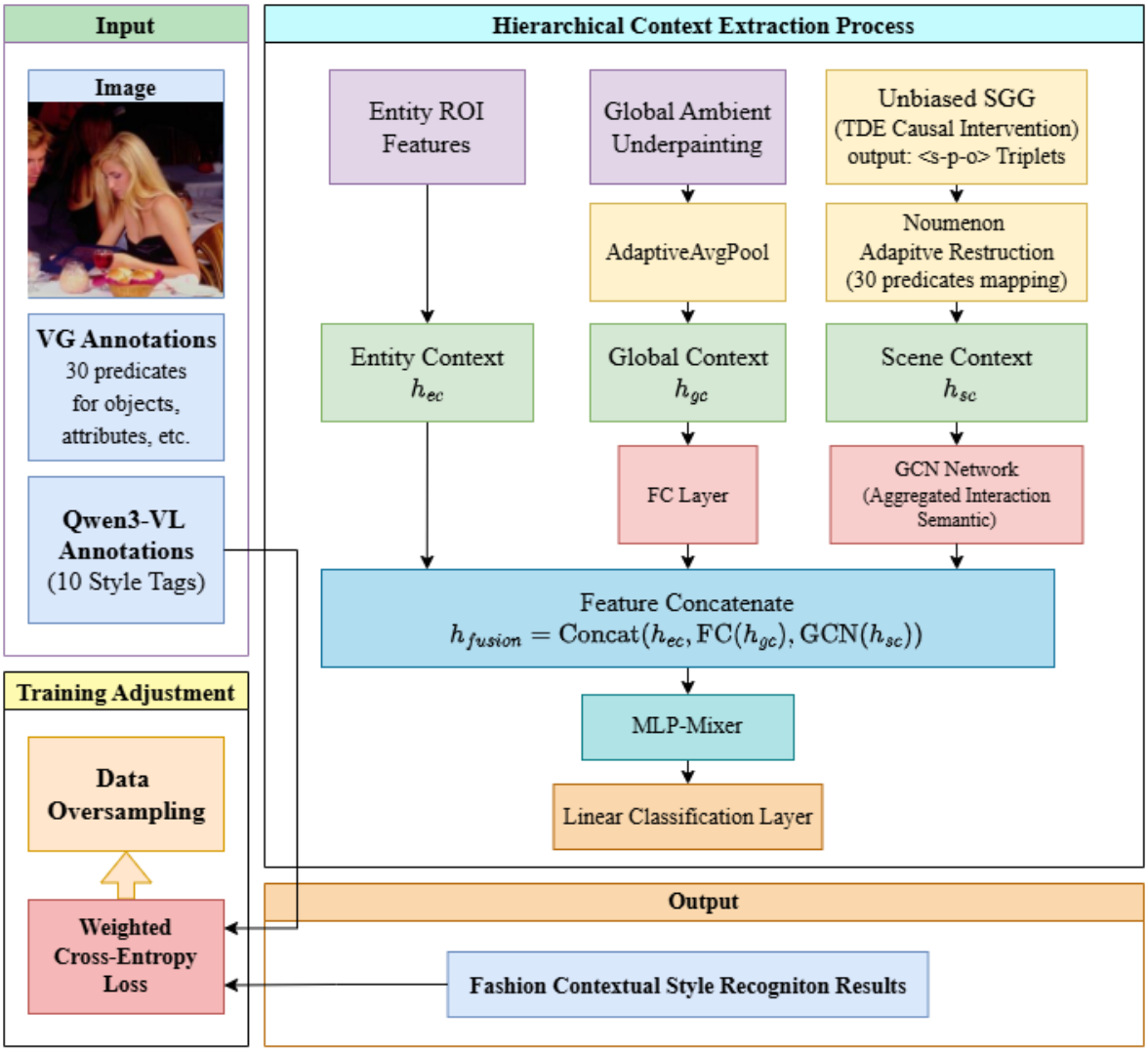

Figure 5: Schematic Diagram of the Fashion Contextual Style Recognition Model

frequencies. This results in a set of structured, debiased triplets. Subsequently, an adapter module performs ontological reconstruction, mapping the raw VG predicates into the 30 categories relevant to the "person-object-environment" framework.

The workflow then proceeds to the hierarchical context feature extraction stage. The entity context (h_{ec}) captures local RoI features of salient entities; the global context (h_{gc}) encodes macro attributes like tone and lighting via global pooling; and the scene context (h_{sc}) aggregates debiased node features through a Graph Convolutional Network (GCN). These hierarchical vectors are integrated via late fusion:

$$h_{\text{fusion}} = \text{Concat}\left(h_{ec}, \text{FC}(h_{gc}), \text{GCN}(h_{sc})\right) \tag{8}$$

The fused representation h_{fusion} is then processed by an MLP-Mixer module to enhance non-linear interactions and cross-dimensional feature expression [12]:

$$h'_{\text{fusion}} = \text{MixerBlock}\left(\text{LayerNorm}(h_{\text{fusion}})\right) \tag{9}$$

Finally, the transformed features h'_{fusion} are mapped to the 10-class style space through a linear classification layer. The entire model is optimized using the weighted cross-entropy loss described in Section 3.4.

4 Experiments

4.1 Experimental Configurations

4.1.1 Datasets

This study utilizes the Visual Genome (VG) dataset as the primary data source. As a large-scale benchmark constructed by Krishna *et al.*, VG comprises over 100,000 images, each meticulously annotated with entity bounding boxes, attributes, and relationship predicates. It is widely regarded as a standard for tasks such as Scene Graph Generation and visual relationship detection [8]. Unlike specialized fashion datasets restricted to e-commerce imagery with plain backgrounds, VG encompasses a rich array of natural settings, diverse entity categories, and intricate human-object interactions. This complexity aligns perfectly with our requirement for deep "contextual" information.

Building upon the methodology detailed in Section 3.3, we subjected the raw VG data to a two-stage refinement protocol. The first stage involved the convergence of relationship predicates toward fashion contexts, where an adapter module was employed to filter and map 30 predicates with high stylistic relevance. In the second stage, we generated supervision signals for style recognition. Leveraging the Qwen3-VL multi-modal large model, each image was annotated with one of 10 fashion style labels, including Candid/Humanistic, Kinetic/Dynamic, and Maximalist/Eclectic. Notably, we deliberately excluded ambiguous categories such as "Other/Unknown" to prevent the model from developing classification evasion behaviors during training.

Regarding data partitioning, we first applied a stratified sampling strategy to maintain a consistent distribution of the 10 style labels across the training, validation, and test sets. Subsequently, a relational validity filter was enforced to retain only those samples containing at least one refined fashion predicate. Following this pipeline, the final dataset consists of 75,653 images for training, 4,115 for validation, and 22,680 for testing.

4.1.2 Evaluation Metrics

To ensure a comprehensive appraisal of the proposed framework, we employ three primary metrics: Top-1 Accuracy, Macro-averaged Recall, and Macro-averaged F1-Score. The decision to prioritize macro-averaging over micro-averaging is driven by the pronounced long-tail distribution within the dataset. Macro metrics provide a more objective reflection of the model's performance on rare styles, ensuring that infrequent categories are not overshadowed by majority classes.
Top-1 Accuracy measures the proportion of images where the holistic style is correctly predicted:

$$\text{Accuracy} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \mathbb{1}(\hat{y}_i = y_i) \tag{10}$$

Macro Recall is computed by first determining the recall for each individual category c, $\text{Recall}_c = \text{TP}_c/(\text{TP}_c + \text{FN}_c)$, and subsequently calculating the arithmetic mean:

$$\text{Macro Recall} = \frac{1}{C} \sum_{c=1}^{C} \text{Recall}_c \tag{11}$$

Similarly, the Macro F1-Score is derived by calculating the precision

$$\text{Precision}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c} \tag{12}$$

and

$$F1_c = \frac{2 \times \text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{13}$$

for each class, followed by averaging:

$$\text{Macro F1} = \frac{1}{C} \sum_{c=1}^{C} F1_c \tag{14}$$

All metrics range from 0 to 1, with higher values signifying superior recognition performance.

4.1.3 Experimental Environment and Parameter Configurations

All computational experiments were conducted on a single NVIDIA GeForce RTX 3060 GPU with 12GB of VRAM. The training pipeline is bifurcated into two distinct phases: a main fine-tuning stage, where all learnable parameters (including SGG and hierarchical context extraction) are updated, followed by a Mixer-specific fine-tuning stage targeting only the MLP-Mixer module. Both phases utilize Stochastic Gradient Descent (SGD) as the optimizer, with momentum and weight decay set to 0.9 and 0.0001, respectively.

We initialized the base learning rate at 0.0005, employing a "warmup-and-decay" scheduling strategy. A linear warmup was applied during the first 500 iterations with a coefficient of 0.1, after which the learning rate underwent adaptive decay (with a factor of 0.3) based on validation loss. The batch size was set to 2, and input images were resized to a short-edge length of 800 pixels, not exceeding a maximum long-edge of 1333 pixels. The maximum iteration counts for the two stages were capped at 15,000 and 3,000, respectively.

4.2 Overall Performance and Comparative Analysis

4.2.1 Baseline Method Overview

To evaluate the efficacy of the proposed framework, we benchmarked our model against the DIG method developed by Yue *et al.* [3], a representative work in fashion style recognition. DIG extracts fashion attributes across four dimensions—texture, fabric, silhouette, and part-style—to construct a design issue graph, which is optimized through a joint training of Graph Convolutional Networks (GCN) and Convolutional Neural Networks (CNN). Notably, the DIG experiments were conducted on the DeepFashion dataset, which predominantly features garment items against plain e-commerce backgrounds with minimal environmental noise.

4.2.2 Quantitative Performance Results

The complete architecture of our model integrates unbiased Scene Graph Generation (via TDE causal intervention), hierarchical context feature extraction, and weighted loss optimization. Table 3 details the overall performance of the integrated model on the Visual Genome (VG) test set.

Table 3: Overall model performance

Metric	Value
Top-1 Accuracy	0.6596
Macro-averaged Recall	0.5001
Macro-averaged F1-Score	0.4998

The results indicate a Top-1 Accuracy of 65.96%, demonstrating the model's capacity to accurately identify holistic fashion contextual styles in the majority of test cases. It should be noted that the Macro-averaged Recall and F1-Score reflect the arithmetic mean across all 10 categories, providing a balanced view of performance across the style spectrum.

4.2.3 Comparative Analysis with DIG

Direct numerical comparison between DIG and our method is constrained by the significant disparity in dataset complexity and experimental settings. DIG was originally designed and evaluated on the DeepFashion dataset, which contains

relatively clean, product-centered images with minimal background interference. In contrast, our method is developed and tested on the Visual Genome dataset, which consists of unconstrained real-world scenes with complex interactions among people, objects, and environments. To provide a contextual reference for existing fashion graph-based reasoning methods, we summarize the performance results of DIG and our model in Table 4. It is important to emphasize that this comparison is not conducted under a strictly aligned benchmark setting, but rather serves to illustrate the relative behavior of both approaches under their respective evaluation environments.

Table 4: Contextual Reference Comparison with Prior Methods

Model	Dataset	Top-1 Acc	Macro Recall	Macro F1
DIG [3]	DeepFashion	0.7463	0.7335	0.7515
Ours	Visual Genome	0.6596	0.5001	0.4998

While DIG reports higher absolute performance on DeepFashion, this observation should be interpreted in the context of dataset characteristics rather than direct model superiority. DeepFashion images are typically well-structured with isolated garments, controlled backgrounds, and limited visual noise. Conversely, Visual Genome presents significantly higher complexity, including cluttered backgrounds, diverse object interactions, and varying environmental conditions. Despite these differences, the proposed model achieves competitive performance under substantially more challenging contextual scenarios. This suggests that integrating unbiased scene graph generation with hierarchical context modeling enables robust semantic reasoning in unconstrained environments, rather than purely item-level visual recognition.

4.2.4 Visualized Analysis via Confusion Matrix

To gain deeper insights into class-specific model behavior, we analyzed the confusion matrix and performance metrics illustrated in Figures 6 and 7. The following patterns emerged:

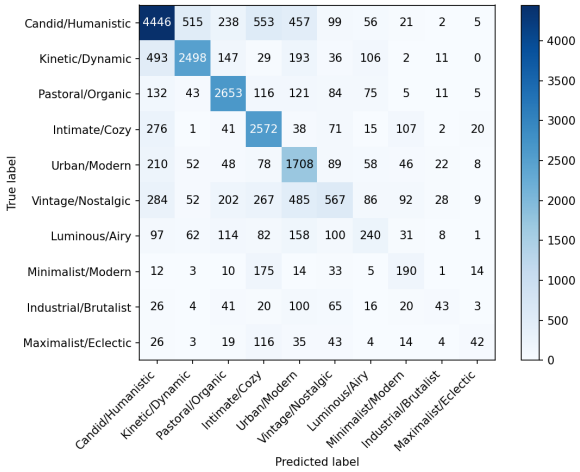

Figure 6: Confusion Matrix Heatmap (Complete Architecture)

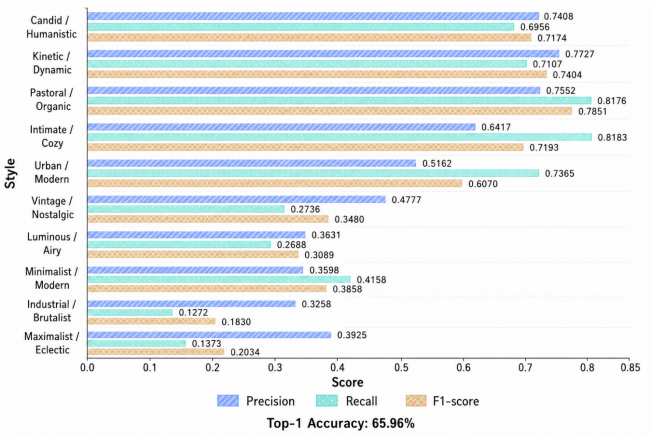

Figure 7: Detailed Recognition Metrics per Category

1) Identifiable Categories:

The Pastoral/Organic style achieved the highest Recall (0.8176) and F1-Score (0.7851). With 2,653 correct predictions, minor confusion occurred primarily with Intimate/Cozy (116 cases) and Candid/Humanistic (132 cases). These results suggest that stylistic cues such as natural lighting and floral elements are effectively captured by the global context branch, making this style highly distinguishable.

2) Moderately Recognized Categories:

Candid/Humanistic yielded a Recall of 0.6956 and an F1 of 0.7174. Errors were largely concentrated in Intimate/Cozy (553 cases) and Kinetic/Dynamic (515 cases). The visual overlap between these styles—often occurring in shared settings like cafes or residential interiors—poses a challenge for categorical separation. While Intimate/Cozy showed high Recall (0.8183), its lower Precision (0.6417) indicates a tendency for the model to "over-predict" this class when presented with ambiguous human-centric scenes.

3) Challenging Categories:

Industrial/Brutalist and Maximalist/Eclectic exhibited the weakest performance, with Recalls of 0.1272 and 0.1373, respectively. Confusion matrix data reveals that Industrial samples were frequently misclassified as Urban/Modern (100 cases). This stems from the extreme scarcity of these samples (approx. 1.5% of the dataset) and the fact that they share visual primitives—such as concrete and metallic textures—with the Urban style.

4) Summary of Confusion Patterns:

Two primary trends are observable. First, the bidirectional confusion between Candid, Intimate, and Kinetic suggests that "person-dominant" styles require even more granular interaction features for precise differentiation. Second, Urban/-Modern acts as a "semantic attractor," frequently appearing in the error distributions of other classes. This phenomenon is linked to its high representation (25.9%) in the dataset, illustrating the persistent distortion effects of the long-tail distribution on classification boundaries—a challenge partially mitigated by our weighted loss and oversampling strategies.

4.3 Ablation Study

4.3.1 Model Variants

To isolate and quantify the impact of individual modules on the overall recognition performance, we designed four ablation variants:

1) Removing Entity Context (w/o h_{ec}):

The local RoI features extracted by the object detector are masked, forcing the model to rely solely on the global atmosphere and relational context for style inference.

2) Removing Global Context (w/o h_{gc}):

The global pooling features are discarded, restricting the model to local entity details and situational relationships.

3) Removing Scene Context (w/o h_{sc}):

The Scene Graph Generation (SGG) branch is deactivated. The model reverts to a "purely vision-driven" paradigm, utilizing only entity and global contexts.

4) Removing TDE Causal Intervention (None group):

Building upon the full architecture, the TDE mechanism is replaced by raw likelihood outputs, thereby bypassing counterfactual bias removal.

4.3.2 Quantitative Analysis of Component Contributions

The quantitative results of the ablation study are summarized in Table 5, which reports the performance changes caused by removing each contextual component or disabling the TDE intervention.

1) Dominance of global context

Discarding the global context resulted in the most significant performance penalty, with accuracy dropping by 7.74 percentage points (from 0.6596 to 0.5822). This outcome empirically confirms that in fashion contextual style recognition, macro-attributes—such as color distribution, lighting, and compositional atmosphere—are more decisive than localized garment details. For instance, identifying a "Luminous/Airy" style depends heavily on high-brightness and low-saturation distributions across the entire image.

2) Auxiliary role of entity context

Removing h_{ec} led to a 4.53% decline in accuracy. Although local visual features are not the sole determinants of style, they provide the essential "material substrate" for recognition. For example, detecting specific entities like concrete walls or metallic pipes significantly bolsters the model's confidence when classifying the "Industrial/Brutalist" style.

3) Necessity vs. sufficiency of scene context

While removing the scene context yielded the smallest individual impact (-1.30%), its value should not be understated. This result indicates that while entity and global features possess strong baseline discriminative power, the scene context provides the structured relationships necessary for fine-grained differentiation, such as distinguishing "Intimate/Cozy" from "Candid/Humanistic" via specific interactions like <person–leaning on–sofa> versus <person–sitting on–sofa>.

4) Causal dependency of the TDE mechanism

With the scene context removed, the results for "With TDE" and "Without TDE" are identical (0.6466). This demonstrates that the TDE intervention relies entirely on the structural paths provided by the scene graph. Without predicate relationships between entities, the causal inference mechanism lacks an operational target, rendering TDE effectively inactive.

5) Trade-off between causal robustness and superficial accuracy

In the full architecture, the accuracy with TDE enabled (0.6596) is slightly lower than the None group (0.6664), a marginal decrease of approximately 0.68 percentage points. This phenomenon highlights a fundamental tension between causal intervention and statistical learning. While the None group may exploit spurious correlations within the dataset—such as a statistical bias linking "grass backgrounds" to the "Pastoral" style—the TDE mechanism actively prunes these non-causal co-occurrences. By forcing the model to rely on genuine visual interactions, TDE enhances generalization in unconstrained scenes, even at a slight cost to raw test set accuracy.

4.4 Analysis and Discussion

Synthesizing the experimental results yields several core insights:

4.4.1 Model Complexity and Computational Efficiency

To further analyze the computational characteristics of the proposed framework, we report a module-wise efficiency breakdown under a unified evaluation protocol. This includes the total number of trainable parameters, inference latency, and relative computational contribution of each component. The results provide a fine-grained understanding of where the computational bottlenecks arise and how the overall cost is distributed across different stages of the pipeline.

The results indicate that the majority of computational cost is concentrated in the object detection stage, accounting for more than 80% of the total inference time. In contrast, the proposed scene graph reasoning, hierarchical fusion, and causal intervention modules introduce minimal additional overhead, demonstrating the efficiency of the designed architecture for contextual modeling. Detailed computational profiling protocol is provided in the Supplementary Material.

4.4.2 Effective Recognition in Unconstrained Scenes

Our framework achieved a Top-1 Accuracy of 65.96% on the Visual Genome dataset, characterized by cluttered backgrounds and variable illumination. Given that this benchmark is significantly more challenging than DeepFashion, these results validate our strategy of capturing structured contextual

Table 5: Comparison of Top-1 Accuracy in Ablation Studies

Configuration	With TDE	Without TDE (None)	Performance Delta (Δ)*
Full Architecture	65.96%	66.64%	—
w/o h_{ec}	61.43%	61.71%	-4.53%
w/o h_{gc}	58.22%	59.97%	-7.74%
w/o h_{sc}	64.66%	64.66%	-1.30%

*Relative to the Full Architecture with TDE enabled.

semantics through scene graphs. While the model excels in categories with distinct visual cues (e.g., Pastoral/Organic with an F1 of 0.7851), performance gaps remain in rare categories like Industrial/Brutalist, primarily due to the long-tail distribution and semantic overlap between styles.

4.4.3 Disparate Contributions within the Hierarchical Context

The ablation study confirms that global context is the strongest predictor of fashion style, while entity context provides the necessary grounding. Crucially, scene context and TDE intervention form an indispensable causal link. Although TDE sacrifices a small fraction of superficial accuracy, it achieves superior causal robustness by systematically eliminating spurious associations.

4.4.4 Persisting Challenges of the Long-tail Distribution

The recognition of rare classes remains the primary bottleneck for model performance. Future work should explore data augmentation and meta-learning techniques to improve the sensitivity of the model toward infrequent styles. Furthermore, the relatively modest independent contribution of scene context suggests that there is room to improve the efficiency of structured relationship utilization. Incorporating Graph Attention Networks or more fine-grained predicate taxonomies could further enhance the model's relational modeling capabilities.

5 Conclusion

This research investigates the fundamental challenge of fashion contextual style recognition in unconstrained scenes, proposing a framework grounded in unbiased scene graphs and hierarchical context fusion. By shifting the definition of fashion style from isolated garment attributes to holistic situational semantics—encompassing interactions between people, objects, and environments—our approach enables precise style identification within complex visual landscapes through structured representations and causal intervention mechanisms

Specifically, the primary contributions of this work are fourfold: (1) we constructed a fashion-oriented Scene Graph Generation and filtering framework by distilling 30 stylistic interaction predicates from the Visual Genome dataset; (2) we designed a three-branch architecture comprising entity, global, and scene contexts, achieving multi-dimensional feature synergy through graph embeddings and adaptive fusion networks; (3) we implemented an unbiased SGG method based on Total Direct Effect (TDE) to prune statistical biases originating from object co-occurrence frequencies, with ablation studies confirming the causal dependency between the TDE mechanism and scene context; and (4) we introduced a dual-balancing strategy combining oversampling and weighted loss to effectively mitigate training deficiencies caused by the long-tail distribution of style labels.

Empirical results demonstrate that our model achieves a Top-1 Accuracy of 65.96% on the Visual Genome dataset, despite its inherent complexity and variable lighting conditions. Ablation analysis reveals disparate contributions across components: global context serves as the most potent predictor (7.74% performance drop upon removal), while entity context provides the necessary material grounding (4.53% drop). Most notably, the scene context and TDE mechanism form a causal loop that significantly enhances the model's capacity to capture genuine relational logic.

Despite these advancements, the experiment also identifies several limitations. Recognition performance for rare categories remains suboptimal—evidenced by low recalls for Industrial/Brutalist (0.1272) and Maximalist/Eclectic (0.1373)—primarily due to severe data imbalance and visual semantic overlap. Furthermore, the modest independent contribution of scene context (-1.30%) suggests that the efficiency of structured relationship utilization warrants further optimization.

Future research will be directed toward four primary areas: (1) integrating meta-learning or diffusion-based data augmentation to enhance the recognition of long-tail categories [17]; (2) exploring the synergy between Vision-Language Models (such as Qwen3-VL) and our framework to leverage cross-modal textual descriptions for deeper situational understanding; (3) employing Graph Attention Networks or finer-grained predicate taxonomies to boost the expressive power of scene contexts [18]; and (4) investigating model compression techniques like knowledge distillation to facilitate deployment on mobile devices [19]. Additionally, the proposed recognition method can serve as a robust context encoder within personalized recommendation systems. By providing high-level situational priors, this framework complements existing works such as PSGNet [14], an application we intend to validate in subsequent studies.

Supplementary Information

Computational Profiling Protocol

To evaluate the computational characteristics of each component in a reproducible and hardware-independent manner, we adopt a CPU-based profiling protocol under a deterministic evaluation setting. This design choice ensures that all modules are assessed under a unified execution environment, eliminating variability introduced by hardware acceleration or vendor-specific optimizations.

Specifically, all experiments are conducted with a fixed batch size of 1 under evaluation mode with gradient computation disabled. Execution time is measured using high-resolution wall-clock timing, averaged over multiple forward passes after an initial warm-up phase to ensure numerical stability.

This protocol focuses on measuring the intrinsic computational burden of each module, enabling a fair comparison of relative complexity across different stages of the proposed framework, including object detection, scene graph generation, causal intervention, and hierarchical feature fusion.

It is important to note that the reported latency values are intended for algorithmic complexity analysis rather than real-world deployment benchmarking, as they reflect a controlled and resource-constrained evaluation setting. Based on this protocol, the module-wise parameter counts, latency values, and relative computational contributions are reported in Table 6.

Table 6: Model Complexity and Module-wise Inference Cost Analysis

Module	Latency (ms)	Contribution (%)
Faster R-CNN + FPN Detector	1690.32	88.41
Scene Graph Generation (SGG)	224.87	11.76
Global Context Extraction	191.44	10.02
Entity Context Extraction	62.18	3.25
Scene Graph Embedding (Attention + MLP-Mixer)	48.56	2.54
Fusion + Classifier Head	9.73	0.51
TDE Causal Intervention	0.00	<0.1
Total Inference Cost	1919.10	100.00

Funding

This research received no external funding.

Author Contributions

Yuming Mai and Xin Zhao contributed equally to this work and should be regarded as co-first authors.

Conceptualization, Yuming Mai and Xin Zhao; Data Curation, Yuming Mai and Xin Zhao; Formal Analysis, Yuming Mai; Investigation, Yuming Mai; Methodology, Yuming Mai; Software, Yuming Mai and Xin Zhao; Validation, Yuming Mai; Visualization, Yuming Mai; Writing—Original Draft Preparation, Yuming Mai; Writing—Review and Editing, Yuming Mai and Xin Zhao; Project Administration, Xin Zhao and Yan Hong; Supervision, Yan Hong; Final Approval of the Manuscript, Yan Hong.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest

The authors declare that they have no conflict of interest.

Data Available

The data and materials used in this study are available upon request from the corresponding author.

Ethical Approval

Not applicable.

References

[1] Sun, G.-L., Cheng, Z.-Q., Wu, X., Peng, Q.: Personalized clothing recommendation combining user social circle and fashion style consistency. Multimedia Tools and Applications **77**, 17731–17754 (2018). https://doi.org/10.1007/s11042-017-5245-1

[2] Lee, H., Seol, J., Lee, S.-g.: Style2vec: Representation learning for fashion items from style sets. arXiv preprint arXiv:1708.04014 (2017). https://doi.org/10.48550/arXiv.1708.04014

[3] Yue, X., Zhang, C., Fujita, H., Lv, Y.: Clothing fashion style recognition with design issue graph. Applied Intelligence **51**(6), 3548–3560 (2021). https://doi.org/10.1007/s10489-020-01950-7

[4] Bossard, L., Dantone, M., Leistner, C., Wengert, C., Quack, T., Van Gool, L.: Apparel classification with style. In Proceedings of the 11th Asian Conference on Computer Vision, pp. 321–335. https://doi.org/10.1007/978-3-642-37447-0_25

[5] Ge, Y., Zhang, R., Wang, X., Tang, X., Luo, P.: DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5332–5340 (2019). https://doi.org/10.1109/CVPR.2019.00548

[6] Vaccaro, K., Shivakumar, S., Ding, Z., Karahalios, K., Kumar, R.: The elements of fashion style. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology, pp. 777–785. https://doi.org/10.1145/2984511.2984573

[7] Kiapour, M.H., Yamaguchi, K., Berg, A.C., Berg, T.L.: Hipster wars: Discovering elements of fashion styles. In European Conference on Computer Vision, pp. 472–488. https://doi.org/10.1007/978-3-319-10590-1_31

[8] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D.A.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017). https://doi.org/10.1007/s11263-016-0981-7

[9] Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D.: Graph R-CNN for Scene Graph Generation. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 670–685. https://doi.org/10.1007/978-3-030-01246-5_41

[10] Wu, S., Zhou, L., Hu, Z., Liu, J.: Hierarchical context-based emotion recognition with scene graphs. IEEE Transactions on Neural Networks and Learning Systems **35**(3), 3725–3739 (2022). https://doi.org/10.1109/

TNNLS.2022.3196831

[11] Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3716–3725. https://doi.org/10.1109/CVPR42600.2020.00377

[12] Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J.: Mlp-mixer: An all-mlp architecture for vision. In Proceedings of the 35th International Conference on Neural Information Processing Systems, pp. 24261–24272 (2021)

[13] Godart, F.C.: "Why is style not in fashion? Using the concept of "style" to understand the creative industries", in Frontiers of Creative Industries: Exploring Structural and Categorical Dynamics. Emerald Publishing Limited, Leeds, United Kingdom (2018). https://doi.org/10.1108/S0733-558X20180000055005

[14] Wang, K., Zhang, J., Zhang, P., Sun, K., Zhan, J., Wei, M.: Personal Style Guided Outfit Recommendation with Multi-Modal Fashion Compatibility Modeling. Journal of Donghua University(English Edition) **42**(02), 156–167 (2025). https://doi.org/10.19884/j.1672-5220.202404017

[15] Dey, A.U., Ghosh, S.K., Valveny, E., Harit, G.: Beyond visual semantics: Exploring the role of scene text in image understanding. Pattern Recognition Letters **149**, 164–171 (2021). https://doi.org/10.1016/j.patrec.2021.06.011

[16] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P.S.: A comprehensive survey on graph neural networks. IEEE transactions on neural networks and learning systems **32**(1), 4–24 (2020). https://doi.org/10.1109/TNNLS.2020.2978386

[17] Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the 34th International Conference on Machine Learning, pp. 1126–1135

[18] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017). https://doi.org/10.48550/arXiv.1710.10903

[19] Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015). https://doi.org/10.48550/arXiv.1503.02531