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

Article

Fashion Style Quantification and Clustering Network Based on Unsupervised Latent Space Representation Learning

Zhiyuan Yu[‡], Ziyi Guo[‡], Yan Hong[†]
College of Textile and Clothing Engineering, Soochow University, Suzhou 215021, China
[†]*E-mail: hongyan@suda.edu.cn*

Received: May 30, 2026 / Revised: June 22, 2026 / Accepted: June 24, 2026 / Published online: July 10, 2026

Abstract: Amidst the rapid evolution of fashion e-commerce, efficiently matching massive inventories with personalized demands requires the objective quantification of aesthetic styles. However, traditional manual annotations are subjective and costly, while low-level visual features struggle with the ”semantic gap” against high-level user preferences. To address these challenges, this study proposes LUCID (Latent Unsupervised Clustering with Interactive Display), an unsupervised latent space representation learning and clustering visualization network driven by vision-language priors. LUCID first utilizes a pre-trained Contrastive Language-Image Pre-training (CLIP) image encoder to extract high-dimensional semantic visual features. To overcome the ”curse of dimensionality,” it employs an auto-encoder integrated with a contrastive learning mechanism to compress these features into denoised, low-dimensional latent vectors. Subsequently, the K-means algorithm automatically clusters these representations, and Uniform Manifold Approximation and Projection (UMAP) generates an intuitive two-dimensional fashion style distribution map. This end-to-end workflow completely circumvents the reliance on manual labels, achieving effective feature dimensionality reduction and semantic refinement. Extensive experiments demonstrate that LUCID’s unsupervised clustering aligns highly with human consensus, achieving ACC, NMI, and ARI scores of 84%, 75%, and 69% on the DeepFashion dataset, with strong generalization on Fashion-MNIST (68%, 69%, 56%). Ultimately, this research effectively bridges the aesthetic semantic gap, providing robust technical support for intelligent fashion analysis systems.

Keywords: Latent Space Representation; Fashion Aesthetics; Deep Clustering; Data Visualization; Multimodal Fusion
https://doi.org/10.64509/jdi.13.119

1 Introduction

Driven by digitalization, intelligent design, and personalized consumption, the fashion industry is shifting from mass production toward data-driven style understanding and user-centered product personalization. Recent studies on metaverse-oriented design systems and automatic style generation indicate that human cognition modeling, intelligent interaction, and generative personalization are becoming important directions for future design intelligence [1, 2]. In this context, fashion products are no longer merely functional commodities; instead, they increasingly serve as carriers of aesthetic preference, lifestyle expression, and personalized identity. Therefore, accurately quantifying fashion styles and organizing large-scale visual fashion data into interpretable aesthetic groups have become critical tasks for intelligent fashion analysis, recommendation systems, and design decision support [3].

Existing fashion recommendation and style analysis methods have made considerable progress, but several limitations remain. Early studies mainly relied on manual labels, collaborative filtering, or user-item interaction histories to infer consumer preferences, which often neglected high-order visual attributes and aesthetic semantics [4, 5]. Although deep learning-based methods have improved automated visual feature extraction, many convolutional neural network-based frameworks still primarily capture low-level visual patterns, such as color, texture, contour, and local shape, while struggling to represent high-level aesthetic concepts such as “minimalist”, “romantic”, “vintage”, and “street style” [6, 7]. This semantic gap between low-level image representations

† Corresponding author: Yan Hong
‡ These authors contributed equally to this work.
* Academic Editor: Zhaohui Wang

and high-level human aesthetic cognition restricts the effectiveness of existing models in complex fashion analysis scenarios.

Another important challenge lies in the high cost and subjectivity of manual annotation. Supervised fashion analysis systems usually require large-scale and fine-grained labeled datasets, but aesthetic labels are inherently ambiguous, context-dependent, and influenced by individual cognition. This makes label construction expensive and potentially biased. In contrast, unsupervised deep clustering provides a promising way to discover latent fashion style structures without relying heavily on manual annotations. Representative methods such as Deep Embedded Clustering and Improved Deep Embedded Clustering integrate feature learning with clustering objectives and have demonstrated the potential of learning compact latent representations from high-dimensional data [8, 9]. However, when applied to real-world fashion images, traditional deep clustering models often overrely on pixel-level reconstruction losses, causing style-related features to become entangled with irrelevant visual noise, including background, lighting, human posture, and shooting environment [10]. Such feature entanglement may lead to clustering drift, where images are grouped according to non-aesthetic factors rather than meaningful fashion styles.

Recent advances in multimodal representation learning provide new opportunities for addressing these issues. Large-scale multimodal models have shown strong cross-modal semantic alignment capabilities and can extract visual representations containing richer high-level semantic priors [11, 12]. Meanwhile, contrastive representation learning encourages semantically similar samples to be closer in the latent space while pushing dissimilar samples apart, which is beneficial for filtering out irrelevant visual interference and improving the discriminative capacity of learned features [13]. Nevertheless, existing fashion clustering and recommendation systems still tend to output isolated recommendation lists or rely on conventional dimensionality reduction methods to generate static scatter plots [14, 15]. These presentation forms are insufficient for revealing continuous aesthetic gradients, transition patterns, and interpretable style boundaries, thereby limiting their practical value in intelligent design and personalized product development [16].

To overcome these limitations, this study proposes Latent Unsupervised Clustering with Interactive Display (LUCID), an unsupervised latent space representation learning and clustering visualization framework for fashion style quantification. Specifically, a pre-trained multimodal visual encoder is first employed to extract high-dimensional semantic visual features from fashion images. Then, a deep autoencoder integrated with a contrastive learning mechanism compresses these features into denoised and discriminative low-dimensional latent vectors. Subsequently, K-means clustering is conducted in the refined latent space to automatically discover fashion style groups. Finally, nonlinear dimensionality reduction is used to project the learned latent representations into a two-dimensional aesthetic distribution map, enabling intuitive visualization of fashion style boundaries and transition relationships.

The main contributions of this study are summarized as follows:

- A vision-prior-driven unsupervised clustering framework is proposed for fashion style quantification, which establishes an end-to-end pipeline from high-dimensional semantic feature extraction to latent representation refinement and aesthetic clustering.
- A joint representation learning mechanism based on autoencoders and contrastive learning is developed to alleviate feature entanglement and enhance the discriminability of latent fashion style representations.
- An interpretable nonlinear visualization strategy is introduced to transform refined high-dimensional latent representations into a two-dimensional aesthetic distribution map, providing intuitive support for fashion style analysis and design-oriented decision-making.
- Extensive experiments on Deep-Fashion and Fashion-MNIST demonstrate that the proposed LUCID framework achieves superior clustering performance and strong generalization capability compared with mainstream unsupervised clustering baselines.

The remainder of this paper is organized as follows. Section 2 reviews related work on fashion visual feature extraction, deep unsupervised clustering, and multimodal representation learning. Section 3 presents the proposed LUCID framework and its optimization mechanism. Section 4 reports the experimental setup, comparative results, ablation studies, and visualization analysis. Section 5 concludes this paper and discusses future research directions.

2 Related Work

This study focuses on the unsupervised quantification and visualization of fashion aesthetic styles, involving interdisciplinary research areas such as fashion visual analysis, deep clustering, multimodal representation learning, and interpretable design intelligence. Existing studies have provided valuable foundations for intelligent fashion recommendation and visual style understanding, but limitations remain in terms of semantic representation, annotation dependence, feature disentanglement, and interpretable visualization. Therefore, this section reviews related studies from three perspectives: fashion visual feature extraction and style quantification, deep unsupervised clustering mechanisms, and vision-language priors with contrastive representation learning.

2.1 Fashion Visual Feature Extraction and Style Quantification

The automated extraction of fashion visual attributes is a fundamental task for intelligent fashion analysis and recommendation systems. Early fashion recommendation methods mainly relied on collaborative filtering, user-item interaction histories, and manually designed features to match consumer preferences with fashion products [3, 4]. Although these methods are effective for capturing historical preference patterns, they often neglect the high-order visual attributes embedded in fashion images, such as silhouette, texture, fabric, style, and aesthetic atmosphere. Image-based recommendation methods further incorporated visual signals into

the recommendation process, improving the ability to identify visually similar or substitutable fashion items [5, 14]. However, these methods still mainly focus on product-level retrieval and recommendation rather than explicit fashion style quantification.

With the rapid development of deep learning, convolutional neural networks have become widely used in fashion image understanding. The success of deep convolutional models in large-scale image classification has provided an important foundation for visual representation learning [17]. In the fashion domain, deep learning-based fashion analysis frameworks can automatically learn visual representations and support tasks such as clothing classification, attribute prediction, and style recognition [6]. Studies on fashion style representation have also explored the modeling of visual clothing style through heterogeneous relationships and co-occurrence patterns [7]. Furthermore, fashion-specific visual systems, such as capsule wardrobe generation, demonstrate the potential of computational models to understand style compatibility and aesthetic consistency [18]. Large-scale fashion benchmarks, such as DeepFashion, have further promoted the development of robust clothing recognition, attribute prediction, and fashion retrieval models [19]. In addition, Fashion-MNIST has been widely used as a lightweight benchmark for evaluating machine learning algorithms on apparel image classification tasks [20]. These studies provide important technical support for intelligent fashion analysis.

Nevertheless, existing supervised fashion analysis paradigms still face two major limitations. First, the performance of supervised models depends heavily on large-scale manually annotated datasets, while fashion aesthetic labels are subjective, ambiguous, and often inconsistent across annotators. Second, traditional visual models tend to capture low-level image features such as color, texture, and shape, but struggle to fully represent high-level aesthetic concepts such as “minimalist”, “romantic”, “vintage”, or “avant-garde”. This semantic gap between low-level visual features and high-level human aesthetic cognition restricts the applicability of existing methods in complex fashion style analysis and design decision-making scenarios.

2.2 Deep Unsupervised Clustering Mechanisms

To reduce the dependence on manual labels, deep unsupervised clustering has become an important research direction for discovering latent structures from high-dimensional data. The core idea is to combine the nonlinear feature mapping capability of deep neural networks with unsupervised clustering objectives, enabling the model to learn compact and cluster-friendly latent representations. Classical deep representation learning studies have shown that auto-encoders can effectively reduce the dimensionality of high-dimensional data while preserving important structural information [21, 22]. Related theories such as sparse coding and independent component analysis also indicate that meaningful representations can be learned by extracting compact and disentangled latent factors from complex visual inputs [23, 24].

Auto-encoder variants have been extensively explored for robust latent representation learning. Denoising auto-encoders learn stable representations by reconstructing clean inputs from corrupted observations, which improves the robustness of feature extraction under noisy conditions [25]. Variational auto-encoders introduce probabilistic latent variables and provide a generative perspective for unsupervised representation learning [26]. Adversarial auto-encoders further regularize latent distributions through adversarial training, enhancing the flexibility of latent space modeling [27]. These methods provide useful foundations for learning compact representations from complex visual data.

Deep Embedded Clustering (DEC) is one of the representative methods in this field. It first uses an auto-encoder to map high-dimensional data into a low-dimensional latent space, and then optimizes clustering assignments by minimizing the Kullback-Leibler divergence between the soft assignment distribution and an auxiliary target distribution [8]. Improved Deep Embedded Clustering (IDEC) further introduces reconstruction loss during clustering optimization to preserve the local structure of the latent space [9]. Other studies have attempted to learn clustering-friendly spaces by jointly optimizing deep representation learning and K-means objectives [28], or by integrating representation learning with hierarchical image clustering in an end-to-end framework [29]. Variational Deep Embedding (VaDE) combines variational auto-encoders with Gaussian mixture modeling, showing that generative latent variable models can also be used for unsupervised clustering [30]. These studies demonstrate the effectiveness of deep clustering in discovering latent data structures without explicit supervision.

K-means remains a classical and efficient partitioning algorithm for multivariate observations and is still widely used as an initialization or baseline method in clustering studies [31]. However, directly applying distance-based clustering algorithms to high-dimensional feature spaces may suffer from distance concentration and reduced metric discrimination [32]. This issue is particularly serious for high-dimensional visual features, where redundant or noisy dimensions may weaken meaningful aesthetic similarity. Therefore, latent-space dimensionality reduction before clustering is necessary for learning compact and discriminative fashion style representations.

However, applying traditional deep clustering models directly to fashion images remains challenging. Real-world fashion images usually contain high-variance noise, including complex backgrounds, diverse human postures, illumination changes, and shooting angles. If the model over-relies on pixel-level reconstruction losses, it may allocate excessive capacity to reconstruct irrelevant visual details rather than learning intrinsic fashion style features. This may lead to feature entanglement and clustering drift, where samples are grouped according to background or photographic conditions rather than meaningful aesthetic styles [10]. Therefore, a more discriminative latent representation learning mechanism is required to filter out style-irrelevant interference and preserve core aesthetic information.

2.3 Vision-Language Priors and Contrastive Representation Learning

Recent advances in multimodal representation learning provide new opportunities for overcoming the semantic gap in fashion style quantification. Large-scale multimodal models learn cross-modal alignment between visual and textual information, enabling visual features to contain richer high-level semantic priors [11, 12]. Compared with conventional image encoders trained only on category labels, vision-language representation models are more capable of capturing abstract semantic concepts and user-oriented descriptions. In particular, Contrastive Language-Image Pre-training (CLIP) learns transferable visual representations from large-scale natural language supervision, making it suitable for extracting high-level semantic visual priors from fashion images [33]. Recent fashion-domain vision-language studies further show that general vision-language models need to be adapted to fine-grained fashion semantics. For example, FashionSAP introduces fashion symbols and attribute prompts to explicitly model fine-grained fashion attributes and improve fashion-domain vision-language representation learning [34]. This characteristic is particularly valuable for fashion analysis, where aesthetic styles are often expressed through subjective and language-related concepts.

Contrastive learning also plays an important role in improving representation quality. By constructing positive and negative sample relationships, contrastive learning encourages similar samples to be closer in the latent space while pushing dissimilar samples apart. Early contrastive mapping studies demonstrated that invariant representations can be learned by reducing the distance between similar pairs and enlarging the distance between dissimilar pairs [35]. Contrastive Predictive Coding further showed the potential of contrastive objectives for unsupervised representation learning [36]. More recent frameworks, such as SimCLR, have demonstrated that contrastive learning can learn effective visual representations from unlabeled data through carefully designed data augmentation and instance discrimination objectives [37]. This mechanism helps the model learn more discriminative and robust feature representations without relying heavily on manual labels.

In fashion recommendation, contrastive multimodal modeling has been used to improve personalized recommendation performance across diverse body shapes and user preferences [13]. From the perspective of design intelligence, human cognition modeling and advanced product personalization further indicate that future fashion systems need to connect user cognition, style generation, and interpretable design decision-making [1, 2]. These studies suggest that intelligent fashion systems should not only recognize existing visual patterns, but also support personalized aesthetic understanding and design-oriented decision support.

Despite these advances, existing fashion clustering and recommendation systems still have insufficient interpretability. Many systems output discrete recommendation lists, which are useful for retrieval but cannot reveal the continuous relationships among fashion styles [5, 14]. Some studies use dimensionality reduction methods such as t-SNE to visualize high-dimensional data, but static scatter plots may still be insufficient for explaining aesthetic gradients and style transition structures [15, 38]. Uniform Manifold Approximation and Projection (UMAP) provides another nonlinear manifold learning method that can preserve both local neighborhood structures and global data relationships, making it suitable for visualizing complex latent distributions [38]. In addition, the black-box nature of deep learning models limits their transparency in design-oriented decision-making processes [16]. Therefore, it is necessary to construct a complete technical pipeline that integrates high-dimensional semantic feature extraction, latent space refinement, unsupervised style clustering, and interpretable visual presentation.

As shown in Figure 1, this study aims to build an intelligent fashion analysis pipeline based on latent space aesthetic refinement. The proposed framework connects semantic feature extraction, deep auto-encoder-based latent representation learning, style clustering, and visual distribution mapping, thereby supporting the transformation from unstructured fashion images to interpretable aesthetic style groups.

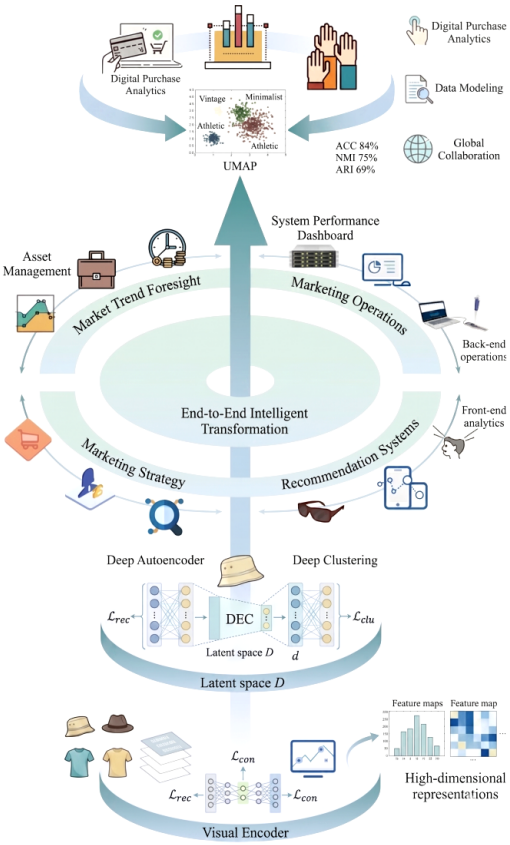

Figure 1: Panoramic Architecture of the Intelligent Fashion Business Pipeline Based on Latent Space Aesthetic Refinement.

3 Methodology

To construct the Latent Unsupervised Clustering with Interactive Display (LUCID) framework for fashion aesthetic quantification, this study develops an end-to-end unsupervised learning pipeline that integrates vision-language semantic priors, deep auto-encoder-based latent representation learning, contrastive feature refinement, clustering optimization, and aesthetic distribution visualization. The overall objective is to

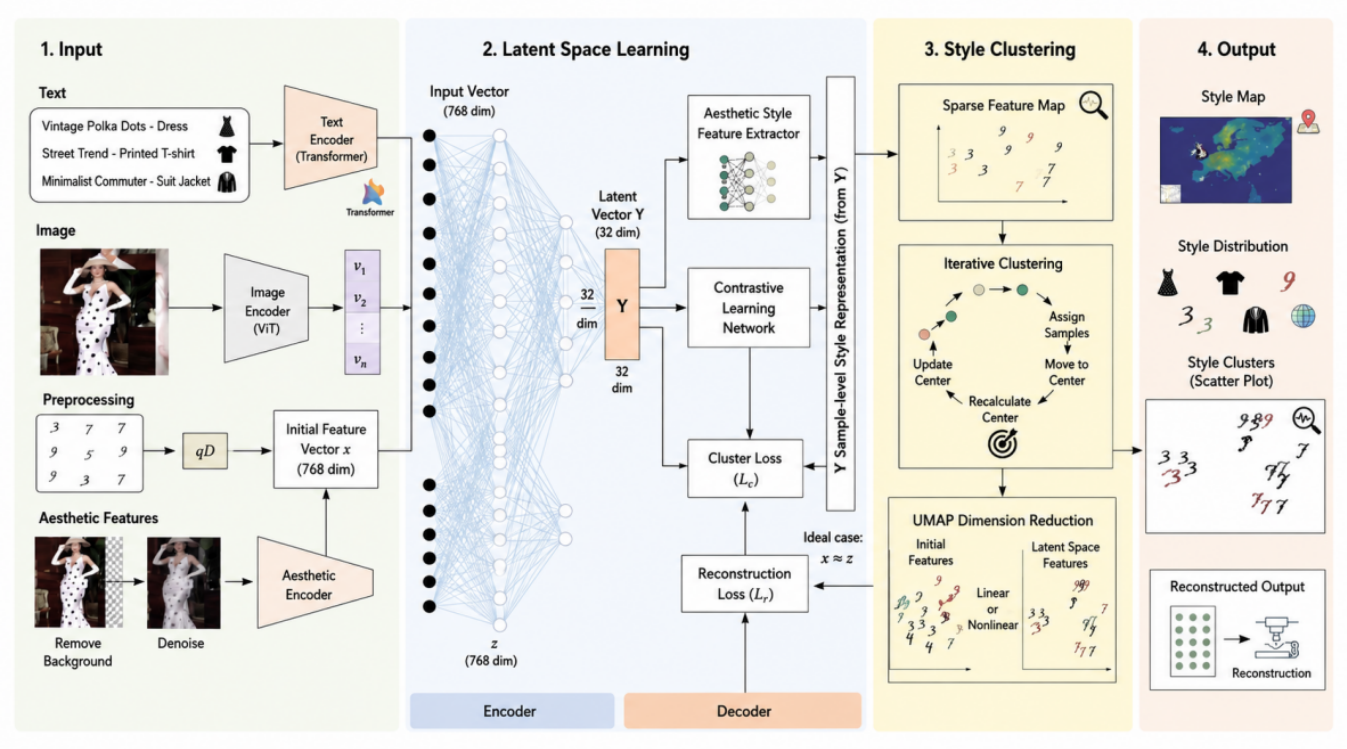

Figure 2: Comprehensive Methodological Workflow of the LUCID Framework

transform high-dimensional and unstructured fashion images into compact, discriminative, and interpretable latent style representations.

As shown in Figure 2, the proposed framework consists of four main modules: visual feature extraction, latent space representation learning, deep clustering, and two-dimensional style visualization. First, fashion images are processed by a pre-trained Contrastive Language-Image Pre-training (CLIP) visual encoder to obtain high-dimensional semantic features. Second, a deep auto-encoder compresses the original feature vectors into a low-dimensional latent space while preserving essential aesthetic information. Third, contrastive learning and clustering loss are jointly introduced to improve intra-cluster compactness and inter-cluster separability. Finally, the refined latent representations are clustered by K-means and projected into a two-dimensional visual map for interpretable fashion style analysis.

3.1 Framework of LUCID

Let $\mathscr{X} = \{x_i\}_{i=1}^{N}$ denote a set of fashion images, where N is the number of samples. Each image x_i is first transformed into a high-dimensional semantic visual feature vector by a pre-trained visual encoder. The extracted feature can be represented as

$$\mathbf{x}_i = f_{\mathrm{v}}(x_i), \quad \mathbf{x}_i \in \mathbb{R}^d, \tag{1}$$

where $f_{\mathrm{v}}(\cdot)$ denotes the visual feature extractor, and d is the dimensionality of the original feature representation. In this study, the input feature dimension is set to $d = 768$, following the high-dimensional semantic representation generated by the visual encoder.

The goal of LUCID is to learn a nonlinear mapping function that transforms the original feature $\mathbf{x}_i$ into a compact latent representation $\mathbf{z}_i$, which can be formulated as

$$\mathbf{z}_i = f_\theta(\mathbf{x}_i), \quad \mathbf{z}_i \in \mathbb{R}^m, \tag{2}$$

where $f_\theta(\cdot)$ denotes the encoder parameterized by θ, and m is the dimension of the latent representation. In the proposed architecture, m is set to 32 to obtain a compact and cluster-friendly representation. Such nonlinear dimensionality reduction follows the general principle of deep representation learning and auto-encoder-based feature compression [21, 22].

3.2 Dimensionality Reduction Based on Auto-Encoder

The auto-encoder is adopted to reduce the dimensionality of high-dimensional fashion features while preserving the essential structure of the original visual representation. It consists of an encoder and a decoder. The encoder maps the input feature into a low-dimensional latent vector, while the decoder reconstructs the original feature from the latent representation. The encoding process is defined as

$$\mathbf{z}_i = f_\theta(\widetilde{\mathbf{x}}_i) = \sigma\left(\mathbf{W}_e\widetilde{\mathbf{x}}_i + \mathbf{b}_e\right), \tag{3}$$

where $\widetilde{\mathbf{x}}_i$ denotes the perturbed input feature, $\mathbf{W}_e$ and $\mathbf{b}_e$ are the weight matrix and bias vector of the encoder, respectively, and $\sigma(\cdot)$ denotes the nonlinear activation function.

The decoder reconstructs the original feature representation from the latent vector, which is formulated as

$$\widehat{\mathbf{x}}_i = g_\phi(\mathbf{z}_i) = \sigma\left(\mathbf{W}_d\mathbf{z}_i + \mathbf{b}_d\right), \tag{4}$$

where $g_\phi(\cdot)$ denotes the decoder parameterized by ϕ, $\mathbf{W}_d$ and $\mathbf{b}_d$ are the weight matrix and bias vector of the decoder, respectively, and $\widehat{\mathbf{x}}_i$ is the reconstructed feature.

For a deep stacked auto-encoder, the layer-wise encoding process can be written as

$$\mathbf{h}_i^{(l)} = \sigma\left(\mathbf{W}^{(l)}\mathbf{h}_i^{(l-1)} + \mathbf{b}^{(l)}\right), \quad l = 1,2,\ldots,L, \tag{5}$$

where $\mathbf{h}_i^{(0)} = \widetilde{\mathbf{x}}_i$, $\mathbf{h}_i^{(L)} = \mathbf{z}_i$, and L denotes the number of encoding layers.

Correspondingly, the layer-wise decoding process is defined as

$$\widehat{\mathbf{h}}_i^{(l-1)} = \sigma\left(\mathbf{W}_d^{(l)}\widehat{\mathbf{h}}_i^{(l)} + \mathbf{b}_d^{(l)}\right), \quad l = L, L-1,\ldots,1. \tag{6}$$

The reconstruction loss is used to preserve the main information contained in the original visual feature. It is calculated as

$$\mathscr{L}_{\text{rec}} = \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i - \widehat{\mathbf{x}}_i\|_2^2. \tag{7}$$

This reconstruction objective helps the model retain useful aesthetic information while suppressing irrelevant noise. To further enhance robustness, the input feature is perturbed before encoding, following the idea of denoising representation learning. This design is consistent with the view that compact latent factors can improve the representation of complex visual inputs [23, 24].

In the proposed LUCID architecture, the encoder follows an expand-then-compress structure. The original 768-dimensional input is first expanded to 1536 dimensions to unfold the feature space, and then gradually compressed through 384 and 96 dimensions to a 32-dimensional latent vector. This structure avoids premature information loss and improves the discriminative capacity of the latent representation.

3.3 Contrastive Latent Representation Refinement

Although the auto-encoder can reduce feature dimensionality, reconstruction loss alone may cause the model to preserve style-irrelevant information such as background, lighting, and shooting environment. Therefore, a contrastive learning constraint is introduced to refine the latent representation. For each sample, two augmented views are generated and encoded into latent vectors $\mathbf{z}_i$ and $\mathbf{z}_i^+$, which form a positive pair. Other samples in the mini-batch are regarded as negative samples. The contrastive loss is defined as

$$\mathscr{L}_{\text{con}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}(\mathbf{z}_i,\mathbf{z}_i^+)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}(\mathbf{z}_i,\mathbf{z}_j)/\tau\right)}, \tag{8}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, and τ is a temperature parameter. By minimizing Equation (8), visually and semantically similar fashion samples are encouraged to be closer in the latent space, while dissimilar samples are pushed apart. This mechanism improves latent feature discrimination and helps alleviate feature entanglement in fashion style clustering [13].

3.4 Joint Clustering Mechanism

After obtaining the latent representations, K-means is first applied to initialize the cluster centers. Let $\boldsymbol{\mu}_j$ denote the center of the j-th cluster, where $j = 1,2,\ldots,K$. Following deep embedded clustering methods [8, 9], the similarity between latent representation $\mathbf{z}_i$ and cluster center $\boldsymbol{\mu}_j$ is measured using Student's t-distribution:

$$q_{ij} = \frac{\left(1+\|\mathbf{z}_i - \boldsymbol{\mu}_j\|_2^2\right)^{-1}}{\sum_{j'=1}^{K}\left(1+\|\mathbf{z}_i - \boldsymbol{\mu}_{j'}\|_2^2\right)^{-1}}, \tag{9}$$

where q_{ij} represents the soft assignment probability of sample i belonging to cluster j.

To emphasize high-confidence assignments and improve intra-cluster compactness, an auxiliary target distribution is constructed as

$$p_{ij} = \frac{q_{ij}^2/\sum_{i=1}^{N}q_{ij}}{\sum_{j'=1}^{K}\left(q_{ij'}^2/\sum_{i=1}^{N}q_{ij'}\right)}. \tag{10}$$

The clustering loss is then defined as the Kullback-Leibler divergence between the target distribution P and the soft assignment distribution Q:

$$\mathscr{L}_{\text{clu}} = \text{KL}(P\|Q) = \sum_{i=1}^{N}\sum_{j=1}^{K}p_{ij}\log\frac{p_{ij}}{q_{ij}}. \tag{11}$$

By minimizing Equation (11), ambiguous samples are gradually assigned to more appropriate clusters, and the latent space becomes more suitable for fashion style grouping.

3.5 Joint Optimization Objective

The final objective of LUCID integrates reconstruction loss, contrastive loss, and clustering loss into a unified optimization framework:

$$\mathscr{L} = \mathscr{L}_{\text{rec}} + \alpha\mathscr{L}_{\text{con}} + \beta\mathscr{L}_{\text{clu}}, \tag{12}$$

where α and β are trade-off coefficients used to balance contrastive representation refinement and clustering optimization.

The model parameters are updated through backpropagation. The cluster centers are updated according to

$$\boldsymbol{\mu}_j \leftarrow \boldsymbol{\mu}_j - \eta\frac{\partial\mathscr{L}_{\text{clu}}}{\partial\boldsymbol{\mu}_j}, \tag{13}$$

where η denotes the learning rate.

The encoder parameters are optimized by considering reconstruction, contrastive, and clustering constraints:

$$\theta \leftarrow \theta - \eta\frac{\partial\mathscr{L}}{\partial\theta}. \tag{14}$$

The decoder parameters are mainly optimized by the reconstruction objective:

$$\phi \leftarrow \phi - \eta \frac{\partial \mathscr{L}_{\text{rec}}}{\partial \phi}. \tag{15}$$

Through this joint optimization process, the proposed framework simultaneously preserves essential visual information, suppresses irrelevant noise, and improves the clustering structure of the latent space. Compared with two-stage clustering strategies, this end-to-end learning mechanism can obtain more compact and discriminative fashion style representations [28, 29].

3.6 Aesthetic Distribution Visualization

After optimization, the learned latent representations are used for fashion style clustering and visualization. The final cluster label of sample i is obtained by

$$y_i = \arg\max_j q_{ij}. \tag{16}$$

To make the clustering results more interpretable, the refined latent vectors are further projected into a two-dimensional space. Conventional linear dimensionality reduction methods may fail to preserve complex nonlinear relationships in high-dimensional aesthetic spaces. Therefore, nonlinear manifold visualization methods are more suitable for presenting fashion style distributions and transition patterns [15]. The visualization process can be expressed as

$$\mathbf{u}_i = \mathscr{V}(\mathbf{z}_i), \quad \mathbf{u}_i \in \mathbb{R}^2, \tag{17}$$

where $\mathscr{V}(\cdot)$ denotes the nonlinear visualization mapping function, and $\mathbf{u}_i$ represents the two-dimensional coordinate of the i-th fashion sample.

In this way, the proposed LUCID framework constructs a complete technical pipeline from high-dimensional semantic visual features to compact latent representations, unsupervised fashion style clusters, and interpretable aesthetic distribution maps. This enables fashion style quantification to support both computational analysis and design-oriented decision-making.

4 Experiments

4.1 Dataset

To evaluate the effectiveness, robustness, and generalization capability of the proposed LUCID framework, experiments were conducted on two fashion-related datasets with different visual characteristics. The first dataset is Deep-Fashion, which is used as the main experimental dataset. The second dataset is Fashion-MNIST, which is adopted as a supplementary validation dataset to test the generalization capability of the proposed framework under a simpler visual distribution.

Deep-Fashion is used as the core dataset in this study, as shown in Figure 3. Specifically, the Category and Attribute Prediction Benchmark subset is adopted for experimental evaluation. After preprocessing operations, including unified cropping, denoising, and image normalization, high-quality fashion images are used for feature extraction and clustering analysis. The dataset contains diverse clothing categories and rich visual attributes, such as texture, fabric, cut, part, and style, which makes it suitable for evaluating fashion style quantification and unsupervised aesthetic clustering.

Fashion-MNIST is used as an additional validation dataset. It contains grayscale fashion images from ten basic apparel categories. Compared with Deep-Fashion, Fashion-MNIST has lower image resolution and simpler visual content. Therefore, it is suitable for testing whether the proposed LUCID framework can maintain stable clustering performance across different data distributions.

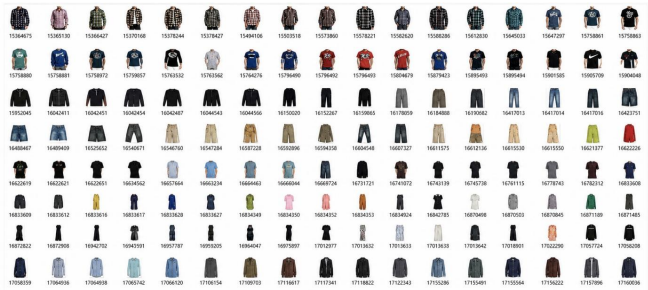

Figure 3: Random samples from the Deep-Fashion dataset

4.2 Baseline Models

To demonstrate the superiority of the proposed LUCID framework, several representative traditional and deep clustering methods are selected as baselines. These models include DEC, IDEC, DCN, VaDE, and JULE. DEC learns low-dimensional representations by using an auto-encoder and optimizes clustering assignments through Kullback-Leibler divergence [8]. IDEC further improves DEC by preserving local structure during clustering optimization [9]. DCN combines deep representation learning with K-means clustering to learn clustering-friendly latent spaces [28]. JULE integrates image representation learning and agglomerative clustering in an end-to-end framework [29]. These baselines are widely used in unsupervised representation learning and clustering analysis, providing a meaningful comparison for evaluating the effectiveness of LUCID.

4.3 Image Feature Extraction Model

Unlike traditional image encoders trained only on general image classification tasks, this study uses a pre-trained multimodal visual encoder to extract semantic-rich fashion image features. The extracted feature vector has a dimensionality of 768. To filter style-irrelevant local noise and generate compact latent representations suitable for clustering, these high-dimensional features are further processed by the proposed auto-encoder.

As shown in Table 1, the encoder adopts an expand-then-compress structure. The 768-dimensional input feature is first expanded to 1536 dimensions to unfold the semantic feature space, and is then gradually compressed through 384 and 96 dimensions to a 32-dimensional latent vector. This structure helps retain important fashion style information while alleviating the curse of dimensionality.

Table 1: Encoder Feature Mapping Network Architecture

Layer	Input Feature	FC 1	FC 2	FC 3	Latent Vector
Output Dim.	768	1536	384	96	32
Activation	–	ReLU	ReLU	ReLU	Linear/ReLU
Regularization	–	Dropout (0.5)	Dropout (0.5)	Dropout (0.5)	–

4.4 Model Training Parameters

The experimental environment is configured with Ubuntu 22.04, an NVIDIA GeForce RTX 4090 GPU with 24 GB memory, Python 3.10, PyTorch 2.1.2, and CUDA 11.8. During training, the 768-dimensional visual features are normalized and then fed into the symmetric auto-encoder. The network is optimized using stochastic gradient descent with a learning rate of 1 and a momentum of 0.9. The model is trained for 100 epochs with a mini-batch size of 250.

The detailed parameter configuration of the auto-encoder is shown in Table 2. The model contains 3,623,648 trainable parameters in total. The encoder gradually compresses the input feature into a 32-dimensional latent vector, while the decoder reconstructs the original 768-dimensional representation from the latent space.

Table 2: Autoencoder Network Parameter Overview

Layer Name	Output Shape	Parameters
input (Input Layer)	(None, 768)	0
encoder_0 (Dense)	(None, 1536)	1,181,184
encoder_1 (Dense)	(None, 384)	590,208
encoder_2 (Dense)	(None, 96)	36,960
encoder_3 (Dense)	(None, 32)	3,104
decoder_3 (Dense)	(None, 96)	3,168
decoder_2 (Dense)	(None, 384)	37,248
decoder_1 (Dense)	(None, 1536)	591,360
decoder_0 (Dense)	(None, 768)	1,180,416
Total Parameters	–	3,623,648
Trainable Parameters	–	3,623,648
Non-trainable Parameters	–	0

4.5 Evaluation Metrics

To objectively evaluate clustering performance, three commonly used metrics are adopted: Clustering Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI) [39, 40]. ACC and NMI range from 0 to 1, where a larger value indicates better clustering performance. ARI ranges from -1 to 1, where a positive value indicates that the clustering result is better than random assignment.

Clustering Accuracy measures the best matching between predicted cluster labels and ground-truth labels. It is defined as

$$\mathrm{ACC} = \max_{m} \frac{\sum_{i=1}^{N} \mathbf{1}\left(y_i = m(\widehat{y}_i)\right)}{N}, \tag{18}$$

where y_i denotes the ground-truth label, $\widehat{y}_i$ denotes the predicted cluster label, $m(\cdot)$ is the optimal mapping function between predicted labels and ground-truth labels, N is the total number of samples, and $\mathbf{1}(\cdot)$ is the indicator function.

Normalized Mutual Information evaluates the mutual dependence between the predicted clustering distribution and the ground-truth label distribution. It is calculated as

$$\mathrm{NMI} = \frac{I(Y,\widehat{Y})}{\left[H(Y)+H(\widehat{Y})\right]/2}, \tag{19}$$

where $I(Y,\widehat{Y})$ denotes the mutual information between the ground-truth label set Y and the predicted label set $\widehat{Y}$, and $H(\cdot)$ denotes information entropy.

Adjusted Rand Index evaluates the consistency between two partitions while correcting for random chance. It is defined as

$$\mathrm{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}}{\binom{N}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}}{\binom{N}{2}}}. \tag{20}$$

Here, n_{ij} denotes the number of samples shared by the i-th ground-truth class and the j-th predicted cluster, while a_i and b_j denote the corresponding marginal sums.

4.6 Comparative Experiments

To comprehensively assess the performance of LUCID, comparative experiments are conducted against mainstream clustering baselines on Deep-Fashion and Fashion-MNIST. The results are shown in Table 3 and Figure 4. Missing baseline results are denoted by "–".

As shown in Table 3 and Figure 4, LUCID achieves the best overall performance on the challenging Deep-Fashion dataset, with an ACC of 0.839, an NMI of 0.754, and an ARI of 0.693. Compared with DEC and IDEC, LUCID obtains higher clustering accuracy and better cluster consistency. On Fashion-MNIST, LUCID also achieves stable generalization performance, with an ACC of 0.683, an NMI of 0.692, and an ARI of 0.560. These results demonstrate that the proposed framework can learn compact and discriminative latent representations for fashion style clustering.

Table 3: Clustering Performance Across Datasets

Model	Deep-Fashion			Fashion-MNIST		
	ACC	NMI	ARI	ACC	NMI	ARI
DEC	0.780	0.678	0.590	0.516	0.546	0.419
IDEC	0.791	0.716	0.633	0.529	0.557	0.428
DCN	0.723	0.699	0.614	0.531	0.571	0.386
VaDE	0.801	0.745	0.662	0.578	0.630	–
JULE	0.809	0.752	0.679	0.562	0.608	–
LUCID	0.839	0.754	0.693	0.683	0.692	0.560

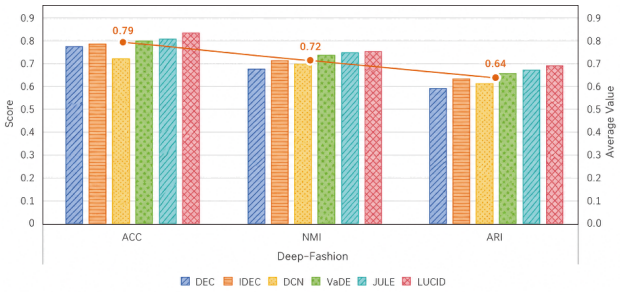

Figure 4: Performance comparison of clustering algorithms on Deep-Fashion

4.7 Ablation Studies

To analyze the contribution of each key component in LUCID, ablation studies are conducted on the Deep-Fashion dataset. The controlled variables include optimizer, learning rate, momentum, training epochs, and batch size. The compared variants are as follows.

Pure K-means is used as a no-dimensionality-reduction baseline, where K-means is directly applied to the 768-dimensional visual features. Auto-Encoder (AE)+K-means first compresses the original features into 32-dimensional latent vectors using a standard auto-encoder and then applies K-means clustering. Variational Auto-Encoder (VAE)+K-means replaces the standard auto-encoder with a variational auto-encoder. Adversarial Auto-Encoder (AAE)+K-means replaces the standard auto-encoder with an adversarial auto-encoder. LUCID represents the full model with Denoising Auto-Encoder (DAE)-based latent representation learning and end-to-end joint optimization.

The ablation results in Table 4 show that dimensionality reduction is necessary for high-dimensional fashion feature clustering. Directly applying K-means to 768-dimensional features only achieves an ACC of 0.587. After auto-encoder-based dimensionality reduction, AE+K-means improves ACC to 0.749, indicating that nonlinear latent representation learning is beneficial for clustering. Compared with two-stage methods, LUCID achieves the best ACC of 0.839 and ARI of 0.693, demonstrating the effectiveness of joint optimization. This result confirms that integrating reconstruction, contrastive representation refinement, and clustering objectives can produce a more discriminative latent space.

Table 4: LUCID Core Component Ablation Results

Model Variant	Dimensionality	Optimization	ACC	ARI
K-means	None	Traditional	0.587	0.481
AE+K-means	Standard AE	Two-stage	0.749	0.617
VAE+K-means	VAE	Two-stage	0.743	0.615
AAE+K-means	AAE	Two-stage	0.776	0.632
LUCID	DAE	Joint	0.839	0.693

Note: The corresponding NMI values are 0.705, 0.765, 0.685, 0.672, and 0.754, respectively.

4.8 Hyperparameter Sensitivity Analysis

The preset number of clusters is an important hyperparameter in unsupervised clustering. An excessively large K may cause over-clustering, where semantically similar fashion samples are split into several small clusters. Conversely, an excessively small K may cause under-clustering, where visually and semantically different styles are incorrectly merged. To ensure fair comparison with baseline models, the objective category number of the datasets is used, and K is set to 10 in the main experiments.

To further examine the distribution structure of the learned latent representations, dimensionality reduction visualization is conducted on the Deep-Fashion dataset. As shown in Figure 5, the latent representations learned by LUCID form relatively clear style regions. Samples belonging to similar fashion styles tend to be close to each other in the projected space, indicating that the proposed model can effectively preserve aesthetic similarity in the latent representation.

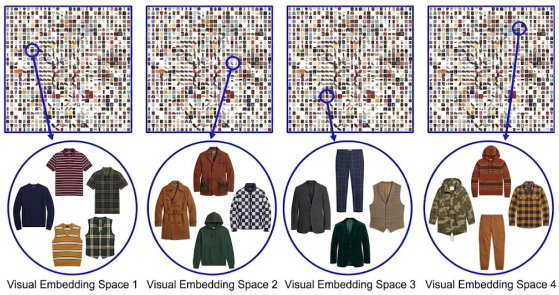

Figure 5: UMAP dimensionality reduction visualization of the Deep-Fashion dataset

4.9 Confusion Matrix Analysis

The confusion matrix is used to further analyze the relationship between predicted clustering labels and ground-truth labels. As shown in Figure 6, since unsupervised learning does not use ground-truth labels during training, predicted cluster indices do not necessarily correspond directly to the original label order. However, the confusion matrix still shows strong local aggregation patterns. Most samples from the same category are concentrated in specific predicted clusters, indicating that LUCID can effectively capture deep fashion style features and generate meaningful clustering results.

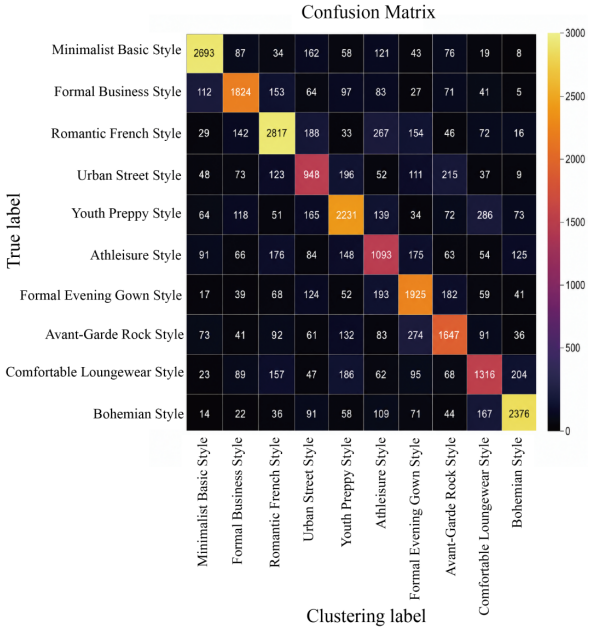

Figure 6: Confusion matrix of the Deep-Fashion dataset

4.10 Failure Case Analysis and Limitations

Although LUCID achieves superior overall clustering performance, several failure cases can still be observed from the confusion matrix and visualization results. As shown in Figure 6, some fashion styles are more likely to be confused when they share similar visual attributes. For example, casual street-style samples may be assigned to athleisure-style clusters, because both categories often contain loose silhouettes, sports-inspired details, dark or neutral colors, and similar wearing scenarios. Similarly, minimalist basic-style samples may overlap with comfortable loungewear-style samples when the garments share simple contours, low-decoration surfaces, and soft fabric appearances.

These failure cases indicate that the proposed model still has limitations in distinguishing fine-grained fashion styles with highly similar visual cues. Since LUCID mainly relies on visual semantic features, it may not fully capture contextual information such as wearing occasion, user intention, fabric hand feel, or cultural style interpretation. In addition, styles in fashion are naturally continuous rather than strictly discrete. Therefore, partial overlap between adjacent categories does not necessarily indicate a completely incorrect representation, but rather reflects the inherent ambiguity and transitional nature of fashion aesthetics.

The above analysis suggests that future improvements should incorporate richer multimodal information, such as textual product descriptions, user comments, material attributes, and purchase behavior, to further enhance fine-grained style discrimination. Moreover, adaptive clustering strategies could be introduced to reduce the dependence on a fixed number of clusters and better capture emerging or hybrid fashion styles.

4.11 Visualization Verification

To further verify the interpretability of the clustering results, Figure 7 visualizes the distribution of ten fashion style clusters in a two-dimensional latent space. The visualization result shows that fashion styles are not completely isolated from each other. Instead, different styles present smooth transitions and partial overlaps, which is consistent with the nature of fashion aesthetics. For example, casual, street, and athleisure styles may share visual elements, while formal and business styles may form adjacent regions in the latent space.

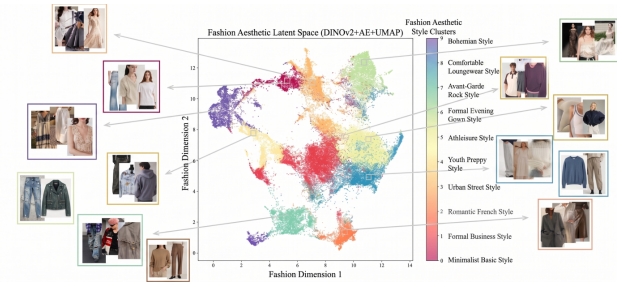

Figure 7: Dimensionality reduction visualization of clustering results

Overall, the experimental results demonstrate that LUCID can effectively bridge high-dimensional fashion visual representations and interpretable aesthetic style structures. The combination of multimodal semantic priors, auto-encoder-based latent space learning, contrastive feature refinement, and clustering optimization enables the proposed model to achieve both strong quantitative performance and intuitive visual interpretability.

5 Conclusions

This study proposes LUCID, an unsupervised latent space representation learning and clustering visualization framework for fashion style quantification. By integrating multimodal semantic priors, auto-encoder-based dimensionality reduction, contrastive representation refinement, and clustering optimization, the proposed framework transforms high-dimensional fashion visual features into compact and interpretable latent style representations. Experimental results on Deep-Fashion and Fashion-MNIST demonstrate that LUCID achieves superior clustering performance and stable generalization compared with mainstream unsupervised clustering baselines. In addition, the visualization results show that the learned latent space can effectively reveal style boundaries and aesthetic transition patterns, providing useful support for intelligent fashion analysis and design-oriented decision-making.

Future research will further investigate adaptive clustering strategies, multimodal user preference modeling, and human-in-the-loop design intelligence systems to enhance the practical value of fashion style quantification in personalized product development.

Funding

This research received no external funding.

Author Contributions

Zhiyuan Yu and Ziyi Guo contributed equally to this work and share first authorship. Concetualization, Zhiyuan Yu; Methodology, Zhiyuan Yu, Ziyi Guo and Yan Hong; Validation, Ziyi Guo; Formal analysis, Zhiyuan Yu and Ziyi Guo; Investigation, Zhiyuan Yu; Data curation,Zhiyuan Yu and Ziyi Guo; Writing—original draft preparation, Zhiyuan Yu; Writing—review and editing, Ziyi Guo and Yan Hong; Visualization, Zhiyuan Yu and Ziyi Guo; Supervision, Yan Hong; Project administration, Yan Hong. All authors have read and agreed to the published version of the manuscript.

Conflict of Interest

All the authors declare that they have no conflict of interest.

Data Avaliable

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

References

[1] Hong, Y., Guo, S., Zeng, X., Zhang, J.: Human cognition modeling for the metaverse-oriented design system.

IEEE Network **38**(6), 243–251 (2024). https://doi.org/10.1109/MNET.2024.3377909

[2] Li, M., Zhang, J., Hong, Y., Xie, X., Zhang, M., Guo, S.: Advanced product personalization in blockchain-enabled metaverse: A diffusion model for automatic style generation. IEEE Internet Things Journal **12**(8), 10304–10315 (2025). https://doi.org/10.1109/JIOT.2024.3511667

[3] Chakraborty, S., Hoque, M.S., Jeem, N.R., Biswas, M.C., Bardhan, D., Lobaton, E.: Fashion Recommendation Systems, Models and Methods: A Review. Informatics **8**(3), 49 (2021). https://doi.org/10.3390/informatics8030049

[4] Hu, Z.-H., Li, X., Wei, C., Zhou, H.-L.: Examining Collaborative Filtering Algorithms for Clothing Recommendation in E-Commerce. Textile Research Journal **89**(14), 2821–2835 (2019). https://doi.org/10.1177/0040517518801200

[5] McAuley, J., Targett, C., Shi, Q., Hengel, A.: Image-Based Recommendations on Styles and Substitutes. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 43–52 (2015). https://doi.org/10.1145/2766462.2767755

[6] Jeon, Y., Jin, S., Han, K.: FANCY: Human-Centered, Deep Learning-Based Framework for Fashion Style Analysis. In Proceedings of the Web Conference 2021, pp. 2367–2378 (2021). https://doi.org/10.1145/3442381.3449833

[7] Veit, A., Kovacs, B., Bell, S., McAuley, J., Bala, K., Belongie, S.: Learning Visual Clothing Style with Heterogeneous Dyadic Co-Occurrences. In 2015 IEEE International Conference on Computer Vision (ICCV), pp. 4642–4650 (2015). https://doi.org/10.1109/ICCV.2015.527

[8] Xie, J., Girshick, R., Farhadi, A.: Unsupervised Deep Embedding for Clustering Analysis. In Proceedings of the 33rd International Conference on International Conference on Machine Learning, pp. 478–487 (2016)

[9] Guo, X., Gao, L., Liu, X., Yin, J.: Improved Deep Embedded Clustering with Local Structure Preservation. In Proceedings of the International Joint Conference on Artificial Intelligence, pp. 1753–1759 (2017)

[10] Jiao, Y., Xie, N., Gao, Y., Wang, C.-C., Sun, Y.: Fine-Grained Fashion Representation Learning by Online Deep Clustering. In European Conference on Computer Vision, pp. 19–35 (2022). https://doi.org/10.1007/978-3-031-19812-0_2

[11] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A Survey on Multimodal Large Language Models. National Science Review **11**(12), 403 (2024). https://doi.org/10.1093/nsr/nwae403

[12] Zhao, F., Zhang, C., Geng, B.: Deep Multimodal Data Fusion. ACM Computing Surveys **56**(9), 1–36 (2024). https://doi.org/10.1145/3649447

[13] Ma, J., Sun, H., Yang, D., Zhang, H.: Personalized Fashion Recommendations for Diverse Body Shapes with Contrastive Multimodal Cross-Attention Network. ACM Transactions on Intelligent Systems and Technology **15**(4), 1–21 (2024). https://doi.org/10.1145/3637217

[14] He, R., McAuley, J.: VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, pp. 144–150 (2015)

[15] Maaten, L., Hinton, G.: Visualizing Data Using t-SNE. Journal of Machine Learning Research **9**(86), 2579–2605 (2008)

[16] Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., Du, M.: Explainability for Large Language Models: A Survey. ACM Transactions on Intelligent Systems and Technology **15**(2), 1–38 (2024). https://doi.org/10.1145/3639372

[17] Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks. Communications of the ACM **60**, 84–90 (2017). https://doi.org/10.1145/3065386

[18] Hsiao, W.-L., Grauman, K.: Creating Capsule Wardrobes from Fashion Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7161–7170 (2018). https://doi.org/10.1109/CVPR.2018.00748

[19] Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1096–1104 (2016). https://doi.org/10.1109/CVPR.2016.124

[20] Xiao, H., Rasul, K., Vollgraf, R.: Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms. arXiv preprint arXiv:1708.07747 (2017). https://doi.org/10.48550/arXiv.1708.07747

[21] Hinton, G.E., Salakhutdinov, R.R.: Reducing the Dimensionality of Data with Neural Networks. Science **313**(5786), 504–507 (2006). https://doi.org/10.1126/science.1127647

[22] Krizhevsky, A., Hinton, G.E.: Using Very Deep Autoencoders for Content-Based Image Retrieval. In Proceedings of the European Symposium on Artificial Neural Networks (2011)

[23] Olshausen, B.A., Field, D.J.: Sparse Coding with an

Overcomplete Basis Set: A Strategy Employed by V1? Vision Research **37**(23), 3311–3325 (1997). https://doi.org/10.1016/S0042-6989(97)00169-7

[24] Bell, A.J., Sejnowski, T.J.: The Independent Components of Natural Scenes Are Edge Filters. Vision Research **37**(23), 3327–3338 (1997). https://doi.org/10.1016/S0042-6989(97)00121-1

[25] Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.-A.: Extracting and Composing Robust Features with Denoising Autoencoders. In Proceedings of the 25th International Conference on Machine Learning, pp. 1096–1103 (2008). https://doi.org/10.1145/1390156.1390294

[26] Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In International Conference on Learning Representations (2014)

[27] Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial Autoencoders. arXiv preprint arXiv:1511.05644 (2015). https://doi.org/10.48550/arXiv.1511.05644

[28] Yang, B., Fu, X., Sidiropoulos, N.D., Hong, M.: Towards K-Means-Friendly Spaces: Simultaneous Deep Learning and Clustering. In Proceedings of the 34th International Conference on Machine Learning, pp. 3861–3870 (2017)

[29] Yang, J., Parikh, D., Batra, D.: Joint Unsupervised Learning of Deep Representations and Image Clusters. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5147–5156 (2016). https://doi.org/10.1109/CVPR.2016.556

[30] Jiang, Z., Zheng, Y., Tan, H., Tang, B., Zhou, H.: VaDE: Variational Deep Embedding: An Unsupervised and Generative Approach to Clustering. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, pp. 1965–1972 (2017). https://doi.org/10.24963/ijcai.2017/273

[31] MacQueen, J.: Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297 (1967)

[32] Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the Surprising Behavior of Distance Metrics in High Dimensional Space. In Proceedings of the 8th International Conference on Database Theory, pp. 420–434 (2001)

[33] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning, pp. 8748–8763 (2021)

[34] Han, Y., Zhang, L., Chen, Q., Chen, Z., Li, Z., Yang, J., Cao, Z.: FashionSAP: Symbols and Attributes Prompt for Fine-Grained Fashion Vision-Language Pre-Training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15028–15038 (2023). https://doi.org/10.1109/CVPR52729.2023.01443

[35] Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality Reduction by Learning an Invariant Mapping. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1735–1742 (2006). https://doi.org/10.1109/CVPR.2006.100

[36] Oord, A., Li, Y., Vinyals, O.: Representation Learning with Contrastive Predictive Coding. arXiv preprint arXiv:1807.03748 (2018). https://doi.org/10.48550/arXiv.1807.03748

[37] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the 37th International Conference on Machine Learning, vol. 119, pp. 1597–1607 (2020)

[38] McInnes, L., Healy, J., Saul, N., Grossberger, L.: UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software **3**(29), 861 (2018). https://doi.org/10.21105/joss.00861

[39] Cai, D., He, X., Han, J.: Locally Consistent Concept Factorization for Document Clustering. IEEE Transactions on Knowledge and Data Engineering **23**(6), 902–913 (2010). https://doi.org/10.1109/TKDE.2010.165

[40] Santos, J.M., Embrechts, M.: On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification. In International Conference on Artificial Neural Networks, pp. 175–184 (2009). https://doi.org/10.1007/978-3-642-04277-5_18