ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 11, No. 6, 1-17 (2026)
www.astesj.com
Special Issue on Emerging Multidisciplinary Directions in Engineering, Computing, and Applied Sciences

ASTES Journal
ISSN: 2415-6698

DualPath-SSA: Separate Prediction of Strength and Importance on Plural Eccentricity Standards for Explainable and Diverse Outfit Recommendation

Kousuke Yamashita*[1], Hiromitsu Shimakawa[2], Fumiko Harada[2]

[1]*Graduate School of Information Science and Engineering, Ritsumeikan University, Ibaraki, 567-8570, Japan*

[2]*College of Information Science and Engineering, Ritsumeikan University, Ibaraki, 567-8570, Japan*

Email(s): is0640hp@ed.ritsumei.ac.jp (K. Yamashita), simakawa@cs.ritsumei.ac.jp (H. Shimakawa), harada@de.is.ritsumei.ac.jp (F. Harada)

*Corresponding Author: Kousuke Yamashita, Graduate School of Information Science and Engineering, Ritsumeikan University, 2-150 Iwakura-cho, Ibaraki, Osaka 567-8570, Japan, Email: is0640hp@ed.ritsumei.ac.jp

ARTICLE INFO

Article history:
Received: 03 October, 2026
Revised: 03 November, 2026
Accepted: 06 November, 2026
Online: 18 November, 2026

Keywords:
Fashion recommendation
Subjective impression
Vision-language model
Label distribution learning
Recommendation diversity

ABSTRACT

In this paper, we propose DualPath-SSA, a fashion-recommendation method that realizes diversity recommendation based on subjective impression while preserving preference fit. Because existing diversification methods rely on differences in objective features, their diversity appears as a blend of many factors and cannot control the continuous change of subjective impression a user perceives. In contrast, focusing on eccentricity as one such subjective impression, the proposed method quantifies the overall eccentricity of an outfit from the visual elements composing it, and thereby expresses a range of subjective impression even among outfits that share a common item.

In experiments on 100 outfits sharing a blue shirt, rated for eccentricity by 82 raters as ground truth, MMR that maximizes distance in the embedding space kept the human-rated eccentricity range below random selection, whereas the proposed method, selecting by the predicted eccentricity, widened this range to about twice that of similarity-based recommendation and reached the level of an ideal selection using human ratings. The method also reproduced the population's ordering of eccentricity on significantly different image pairs at 82.6% (human reproducibility ceiling 88.2%), showing that subjective impression can be quantified at a level approaching human ratings.

These results show that DualPath-SSA is a new fashion-recommendation framework integrating the quantification and diversity control of subjective impression.

1. Introduction

This paper is an extension of work originally presented at the 2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA) [1].

1.1. Background

Recommender systems have developed around collaborative filtering and content-based filtering, reducing users' search costs while presenting items that fit their preferences well [2]. However, excessive optimization toward past behavior and similarity measures causes "filter bubbles" and "over-specialization", in which homogeneous items are over-presented. This lowers the diversity of recommendations, and the loss of opportunities for discovery (serendipity) and the resulting decline in satisfaction have been pointed out as problems.

The impact of this loss of diversity is particularly pronounced in the fashion domain [3]. Clothing and outfits are not merely functional goods but media for self-expression and social impression, so users' behavior combines a desire to "conform" to existing preferences with a desire to "differentiate" themselves by seeking novelty. If a system clings to current preferences and presents only similar outfits, users lose opportunities to explore and extend new styles, which tends to reduce long-term engagement.

Conventional recommendation research has proposed methods that increase the diversity of a recommendation list through re-ranking or topic diversification [4, 5]. However, most of the diversity handled by these methods is "objective diversity" based on item attributes or distances in a **feature space**. Here, a feature space is the space into which an image recognition model maps an image as a numeric vector of several hundred dimensions, in which images with similar visual properties such as color tone and shape are placed close together. Two outfits that are far apart in this space certainly differ visually, but that difference does not necessarily correspond to a difference in the impression a person receives. Because the diversity that users perceive in fashion recommendation depends heavily on the subjective impression received from the outfit as a whole, diversification of attributes or features does not necessarily correspond to diversification of subjective impression. Indeed, a comparative experiment that presented users with recommendation lists constructed under different diversity metrics reported that metric-level diversity and perceived diversity did not agree, and that users found it hard to distinguish fine differences in diversity [6]. Therefore, to realize encounters with new styles while maintaining preference fit, explicitly defining and controlling this "subjective diversity" becomes an important challenge.

1.2. Objective and Three Requirements

The objective of this study is to realize recommendation diversity based on subjective sensibility. Recommendation that satisfies this objective must meet the following three requirements simultaneously.

1. **Preference fit:** Recommendations must be made within the range that fits the user's preferences, including the item the user has specified. This study assumes a candidate pool sharing the specified item, and the narrowing-down itself is out of scope. What is asked is whether the range of impressions can be predicted and presented even inside this range, where candidates are visually similar.

2. **Subjective diversity:** The impressions contained in the recommendation list must be spread in the user's perception. An impression is formed from the combination of the whole outfit, and raters disagree about it. Therefore, rather than reducing the impression to a single deterministic value, the degree to which ratings split must be retained together with the representative value. Although this study does not use the degree of disagreement to control recommendation directly, it indicates where individual differences in perception appear and is information that serves as a starting point for future personalization.

3. **Explainability:** It must be possible to present to the user why an outfit was predicted to give that impression. Presenting the reasons for a recommendation has been shown to increase the transparency of a recommender system, improve users' acceptance and satisfaction, and help users evaluate the validity of the recommendation themselves [7, 8, 9]. To enhance explainability, this study aims to present, in addition to the predicted eccentricity, the six visual elements that grounded the judgment together with the strength of deviation and the importance to the impression of each element. This allows the user to understand which visual elements the system used as grounds for judging the outfit eccentric.

To satisfy these three requirements simultaneously, it is not enough to quantify subjective impression; the quantification process must have a structure that corresponds to human interpretation. Furthermore, it is precisely the cases on which ratings split that leave room for raising satisfaction per user, and these form the focal point of personalization.

1.3. Hypothesis, Approach, and Contributions

As one subjective impression, this study focuses on the **eccentricity** of an outfit—the degree of visual deviation from a safe, standard style. Eccentricity expresses the degree of deviation from the norm and serves as an impression axis for directly controlling novelty and serendipity in recommendation. Sociologically, eccentricity is a phenomenon that emerges in the tug of war between social conformity and individual differentiation [10], and carries the context of deliberate deviation from the norm [11].

Our hypothesis is that **eccentricity can be decomposed into visual axes and interpreted**. That is, we consider that the eccentricity of an outfit can be explained as a combination of the "amount of deviation from the norm" on each of six visually observable axes—color, pattern, silhouette, material, accessory, and styling—and "which axis mainly determines the impression". This multi-axis decomposition is consistent with recent computational studies that treat the aesthetic evaluation of outfits in multiple dimensions [12].

Based on this hypothesis, we propose DualPath-SSA, which separately predicts the **strength** (amount of deviation) and **importance** (allocation of the impression) of each axis and aggregates them to output eccentricity as a five-level probability distribution. The aggregation of strength and importance quantifies the impression (Requirement 2), importance presents the grounds of judgment (Requirement 3), and the distribution output retains the degree of disagreement. The recommendation list is constructed to cover the predicted range in the candidate pool at equal intervals, giving a spread of impressions inside the range that fits the user's preferences (Requirement 1).

The contributions of this study are the following three points, each corresponding to one requirement.

- **Realizing subjective diversity (Requirements 1 and 2):** Through a recommendation selection experiment on a candidate pool sharing a specified item, we show that maximizing the distance in embedding space does not widen the range of human-rated eccentricity and falls below random selection. In contrast, the proposed method, which covers the range of predicted eccentricity, widens that range to about twice that of MMR and reaches the level of an oracle that uses human ratings directly.

- **Accuracy of separate prediction (Requirement 2):** We show that the configuration that predicts strength and importance separately and aggregates them outperforms regression

to a single scalar in the ordering agreement on discriminable image pairs and approaches the human reproducibility ceiling. There is no cost in accuracy for retaining an explainable intermediate representation.

- **Providing explainability (Requirement 3):** We show that axis strength and axis importance are each independently valid with respect to their teacher labels, and that for a single image the method can present "which axis", "how much it deviates", "which concept", and "where in the image" it reacted to.

In addition, we define an evaluation protocol that partitions the data into bands according to inter-rater agreement and empirically measures, from repetitions differing only in random seed, the lower bound of differences worth discussing, and we make the absolute values of the metrics interpretable by presenting an uninformative reference point and the human reproducibility ceiling alongside.

The remainder of this paper is organized as follows. Section 2 reviews related work, Section 3 describes the proposed method, Section 4 the datasets and the analysis of inter-rater agreement, Section 5 the evaluation protocol, Section 6 the experimental results and discussion, and Section 7 concludes.

2. Related Work

2.1. Definition of Eccentricity in Outfits

Impression evaluation and aesthetic judgment of outfits carry social and psychological meaning beyond the evaluation of the practical function of clothing. In [10], the authors viewed fashion as a balance between "social conformity" and "individual differentiation" and argued that deviation from norms plays an important role in expressing individuality. In [11], likewise showed the process by which particular outfits come to be regarded as "deviant" in a social context, pointing out that eccentricity in outfits is not a mere visual feature but a subjective impression formed in relation to social norms. The eccentricity treated in this study is likewise defined not as the change of fashion trends themselves but as the degree of deviation from the current standard outfit.

However, eccentricity is not determined by a single visual element. For example, an outfit may look eccentric because of a loud color scheme, or because of an extreme silhouette, decoration, or way of wearing. Furthermore, the study in [13] showed that the relation between the degree of color coordination and fashionableness is maximized not at complete matching or clashing but at "moderate coordination", reporting that aesthetic impression is a nonlinear (inverted-U) judgment with respect to the degree of coordination.

To handle such complex subjective impressions computationally, it is effective to decompose the factors of impression formation into multiple visual dimensions. In [12], the authors proposed an evaluation framework based on fine-grained attributes such as color and silhouette in the fashion aesthetic benchmark A100 and showed the effectiveness of multi-axis evaluation. Eccentricity is therefore appropriately represented not by a single scale but by multiple visual axes; our decomposition of eccentricity into six axes is based on these findings.

2.2. Diversification Methods in Recommender Systems and Their Limits

Research on diversification has been conducted widely to prevent the monotonization and over-specialization of recommendation lists and to increase serendipity. Representative methods include Maximal Marginal Relevance (MMR) [4], which combines relevance and dissimilarity, topic diversification [5], and Intra-List Diversity (ILD) [14], which evaluates the average dissimilarity within a list.

In fashion recommendation, the study in [3] showed, in a field experiment with 20,000 customers at physical stores, significant improvements in purchase response and repurchase rates by combining different recommendation algorithms per customer type to raise catalog coverage. However, most of the diversity handled in existing studies is "objective diversity" based on category information, distances in embedding space, or catalog coverage as in [3]. The authors of [3] themselves note as a limitation that the customer's sensitivity and subjective experience are not reflected in the recommender system and leave this as future work. That is, the gap of quantifying subjective impression has already been pointed out as an unresolved problem even in empirical studies at practical scale.

Meanwhile, the diversity that people perceive from outfits depends strongly not only on differences in item attributes but on differences in the impression received from the whole. Therefore, a recommendation list diversified in feature space is not necessarily perceived by the user as a list of different impressions. A large gap remains between objective diversification and diversification of subjective impression, and a framework that quantifies subjective impression itself and uses it for recommendation control is needed.

2.3. Learning Representations of Sensibility and Impression with Vision-Language Models

Vision-language models (VLMs), which learn a joint embedding space of images and language, have been reported to acquire representations that are effective not only for image recognition but also for aesthetic evaluation and sensibility analysis. Representative models include CLIP [15] and SigLIP [16]; SigLIP in particular learns high-quality semantic representations using a sigmoid loss.

In [17], the authors reported that VLM-derived features show higher predictive performance than conventional image features in aesthetic evaluation tasks. In [18], the authors showed that the texture and impression of an image can be evaluated without training by contrasting similarities to antonymous text prompts, supporting the view that similarity to text prompts is a handle for interpreting and controlling the semantic attributes of an image. On the other hand, image features and text features are known to be distributed in separate regions of the joint space (modality gap [19]), so using prompt similarities as they are requires handling the bias arising from this distributional difference.

2.4. Modeling Uncertainty in Subjective Evaluation

Human subjective evaluation inherently contains inter-rater variability and ambiguity. Label distribution learning (LDL) [20],

which treats evaluation results not as a single label but as a probability distribution, has therefore attracted attention.

For ordinal rating scales in particular, the Earth Mover's Distance (EMD) [21], which accounts for the distance between distributions, and the ordinal regression method CORAL [22] have been reported to be effective. In [23], the authors predicted the aesthetic evaluation of images as a distribution of rater scores and showed that using squared EMD as the loss yields higher accuracy than regression to a single value. These methods are characterized by expressing not only the representative evaluation but also inter-rater disagreement and uncertainty.

2.5. Interpretable Machine Learning and Concept Bottlenecks

The design of this study, which presents the grounds of eccentricity judgment per axis, belongs to the lineage of interpretable machine learning. In [24], the authors argues that in high-stakes decision making one should use models that are interpretable from the outset rather than explaining black-box models post hoc, and that there is no inherent trade-off between interpretability and accuracy as widely believed. The same paper further points out that post-hoc explanations can in principle be unfaithful: if an explanation were completely faithful to the model's computation, the explanation itself would be equivalent to the model, and there would be no point in preparing a separate explanation.

This concern has developed into Concept Bottleneck Models [25], which route predictions through human-readable concepts as an explicit bottleneck. In forcing an interpretable intermediate representation before the final output, this corresponds to our design, which passes through the intermediate representations of strength and importance before obtaining the representative value μ.

In the recommender systems field as well, the survey by [9] organizes explanations along the important axis of whether they are model-intrinsic or post hoc. Our axis strength $\mathbf{s}$ and axis importance $\hat{\mathbf{w}}$ are the inputs of the very computation that yields the representative value and are not post-hoc approximations.

In [26], the authors also position interpretability not merely as a means of explaining outputs but as a scientific methodology for verifying the validity of a model. The verification of the validity of each mechanism in this study (Section 6.2) is a diagnosis that becomes possible only with a design that passes through per-axis intermediate representations, and is consistent with this position.

Viewed against the three requirements of Section 1, the prior work can be organized as follows. Diversification methods for recommendation have aimed to reconcile preference fit and diversity, but the diversity they can handle is limited to objective diversity in attributes or feature space, and the diversity of subjective impression is not controlled (Requirements 1 and 2). Label distribution learning, which treats subjective evaluation as a distribution, and impression representation by vision-language models provide the means to quantify impression itself, including inter-rater variability (Requirement 2). Interpretable designs that pass through concepts provide a framework for presenting the grounds of judgment in human-readable units (Requirement 3). Each of these has been shown to be effective individually, but, to the best of our knowledge, no study has decomposed subjective impression into interpretable elements, predicted it as a distribution, and applied it to controlling the subjective diversity of a recommendation list. This study aims to fill that gap.

3. Proposed Method: DualPath-SSA

3.1. Problem Setting

Given an outfit image x, this study addresses the problem of predicting the eccentricity rating distribution $\mathbf{y} = (p_0, \ldots, p_4)$ of a population of raters for that outfit. Here p_k is the probability that a rater assigns eccentricity score $k \in \{0, \ldots, 4\}$, and $\sum_{k=0}^{4} p_k = 1$.

To express the label distribution by a representative value and a spread, we let μ be the mean and σ the standard deviation of the rater distribution and use the (μ, σ) representation. The proposed method predicts the representative value μ of eccentricity from the input image x and constructs σ from the characteristics of the inter-rater distribution. A continuous distribution is then generated from (μ, σ) and converted into the five-level discrete label distribution $\mathbf{y}$. A set of texts describing, for each visual axis, visual concepts of graded eccentricity (the prompt bank, Section 3.5.1) is also assumed to be given in advance.

For recommendation, we assume a candidate pool $C = \{x_1, \ldots, x_N\}$ that shares the item specified by the user. Narrowing down the candidate pool itself is outside the scope of this study; what we address is the problem of constructing a recommendation list $\mathcal{R} \subset C$ of K items, based on the eccentricity distributions predicted for the images in the pool, such that the range of impressions covers the range of the pool (Section 3.7).

3.2. Overview of DualPath-SSA

Rather than predicting eccentricity directly from image features, this study proposes DualPath-SSA (Dual-Path Strength–Semantic Attribution), which separately predicts the **strength**, indicating how far each visual element deviates from the norm, and the **importance**, indicating to what extent the impression of eccentricity stems from each element (Figure 1). This separation enables eccentricity prediction that accounts not only for the state of each element in isolation but also for the relations among elements. For example, even when outfits contain patterns of similar distinctiveness, the magnitude of the pattern's effect on overall eccentricity differs depending on its combination with color and silhouette. This study aims to predict final eccentricity while incorporating such combination effects.

The proposed method consists of the following three mechanisms.

The axis strength prediction mechanism (strength path) extracts ordinal features for each of the six axes—color, pattern, silhouette, material, accessory, and styling—from a LoRA-adapted SigLIP2-base and predicts the axis strength $\mathbf{s} \in [0, 2]^6$ by CORAL ordinal regression (Section 3.4). The teacher label of each axis has three levels: 0 (standard), 1 (slightly deviant), and 2 (strongly deviant). Strength carries the quantity "how far each axis deviates".

The axis importance prediction mechanism (importance path) predicts the axis importance $\hat{\mathbf{w}} \in \Delta^5$ by per-axis ridge regression whose inputs are both the similarity $\mathbf{d}_i \in \mathbb{R}^{20}$ to a prompt bank of 20 text concepts per axis and the visual feature $\mathbf{x}_i$ extracted by the strength path for the same axis (Section 3.5). Importance

carries the allocation "which axis the impression stems from" and is normalized to sum to one. That importance is determined neither by the amount of deviation alone nor by the concepts present alone is confirmed in Section 6.2.

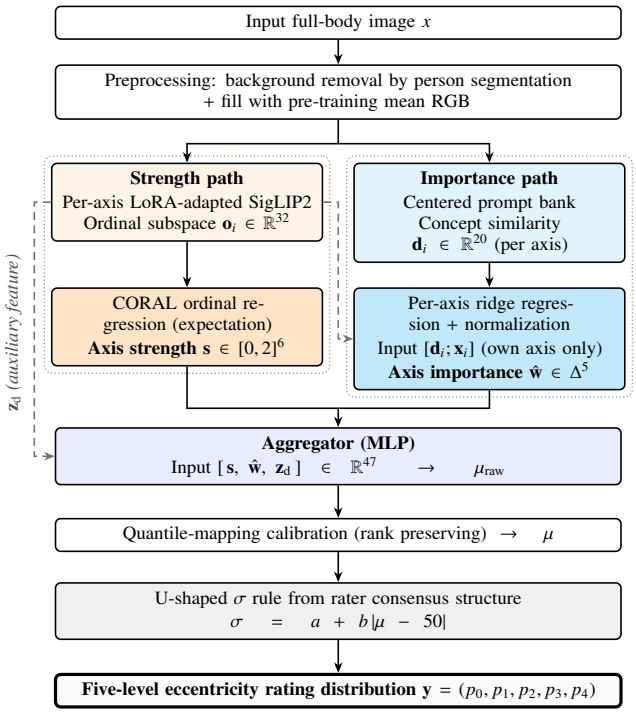


Figure 1: Overall structure of DualPath-SSA. The dotted frames mark the extent of each path. The strength path (left) predicts the axis strength $\mathbf{s}$, the amount of deviation of each axis. The importance path (right) predicts the axis importance $\hat{\mathbf{w}}$ by per-axis ridge regression whose inputs are the similarity $\mathbf{d}_i$ to the per-axis prompt bank and the visual feature $\mathbf{x}_i$ of the same axis passed from the strength path (center dashed line). The aggregator obtains the representative value from $\mathbf{s}$, $\hat{\mathbf{w}}$, and the auxiliary visual feature $\mathbf{z}_d$ derived from the reconstruction-regularized subspace of the strength path (left dashed line).

The eccentricity distribution prediction mechanism integrates $\mathbf{s}$, $\hat{\mathbf{w}}$, and the auxiliary visual feature $\mathbf{z}_d$ derived from the strength path in an aggregator to predict the representative value μ, assigns the distribution width σ by a rule grounded in the consensus structure of human ratings, and outputs a five-level probability distribution (Section 3.6). The distribution carries "how the ratings split".

Finally, the recommendation list is constructed so as to cover the range of predicted values over the candidate pool (Section 3.7).

3.3. Image Preprocessing

Image preprocessing is common to the strength path and the importance path. The input image x is a full-body outfit image including its background. Background colors and incidental objects may affect the eccentricity prediction as visual information other than clothing. We therefore extract and remove the background region with a person segmentation model and replace the removed region uniformly with the mean RGB value used in pre-training the backbone vision model, SigLIP2-base. The preprocessed image is resized to the input resolution of SigLIP2-base. Hereafter, the preprocessed image is denoted by x'.

3.4. Axis Strength Prediction Mechanism (Strength Path)

The strength path predicts, from the preprocessed image x', the visual strength of eccentricity on each visual axis. For the six axes of color, pattern, silhouette, material, accessory, and styling, it learns ordinal representations corresponding to the per-axis eccentricity level $y_i \in \{0, 1, 2\}$ (0: standard, 1: slightly deviant, 2: strongly deviant) and produces the axis strength $\mathbf{s} \in [0, 2]^6$.

3.4.1. Per-Axis Feature Extraction by Subspace Partitioning and Training-Time Regularization

For each visual axis $i \in \{1, \ldots, 6\}$, we apply an independent low-rank adaptation (LoRA [27], rank 8) to the vision encoder of the backbone SigLIP2-base (`siglip2-base-patch16-224`), constructing six axis-specialized vision encoders.

The preprocessed image x' is fed to each encoder, and on its pooler output $\mathbf{h}_i \in \mathbb{R}^{768}$ we place two projection heads with different purposes to partition the representation space. Both projection heads are two-layer fully connected networks (hidden width 256, GELU activation) that take $\mathbf{h}_i$ as input; they do not share the hidden layer and are trained with independent parameters. The **ordinal subspace** ($768 \rightarrow 256 \rightarrow 32$) preserves the ordinal structure of the per-axis teacher label $y_i \in \{0, 1, 2\}$; the projected feature $\mathbf{o}_i \in \mathbb{R}^{32}$ is fed to CORAL ordinal regression to predict the axis strength $\mathbf{s}$. The **reconstruction-regularized subspace** ($768 \rightarrow 256 \rightarrow 96$) prevents the per-axis LoRA from over-adapting to ordinal discrimination and losing fine-grained visual information such as saturation and texture. A decoder reconstructs the frozen original SigLIP pooled embedding $\mathbf{f}$, and the cosine dissimilarity $\mathcal{L}_{\text{distill}} = 1 - \cos(\text{Dec}(\mathbf{e}_i), \mathbf{f})$ is used as a loss. At inference, the bottleneck features $\mathbf{e}_i \in \mathbb{R}^{96}$ of the six axes are concatenated and compressed by principal component analysis to 90% cumulative explained variance, yielding $\mathbf{z}_d \in \mathbb{R}^{35}$, which is used as an auxiliary visual feature for the aggregator.

The per-axis encoder is trained by simultaneously optimizing the LoRA adapter and the two projection heads with the following loss consisting of three terms: ordinal constraint, class discrimination, and preservation of visual representation.

$$\mathcal{L} = \bar{w}\,(\mathcal{L}_{\text{CORAL}} + \lambda_{\text{CE}}\,\mathcal{L}_{\text{CE}}) + \lambda_{\text{d}}\,\mathcal{L}_{\text{distill}} \tag{1}$$

Here $\lambda_{\text{CE}} = 0.5$ and $\lambda_{\text{d}} = 1.0$. The first term $\mathcal{L}_{\text{CORAL}}$ is the ordinal constraint. From the ordinal-subspace feature $\mathbf{o}_i$, a linear layer yields the logits $z_{i,1}, z_{i,2}$ corresponding to the two cumulative boundaries "level 1 or higher" and "level 2 or higher", and the binary cross-entropy against the cumulative binary targets $\mathbf{1}[y_i \geq 1]$, $\mathbf{1}[y_i \geq 2]$ derived from the teacher label y_i is taken.

$$\mathcal{L}_{\text{CORAL}} = -\frac{1}{2}\sum_{k=1}^{2}\Big[\,\mathbf{1}[y_i \geq k]\log\sigma(z_{i,k}) + \mathbf{1}[y_i < k]\log\left(1 - \sigma(z_{i,k})\right)\Big] \tag{2}$$

Here σ is the sigmoid function, and $p_{i,k} = \sigma(z_{i,k})$ is the cumulative probability used in Section 3.4.2. The second term $\mathcal{L}_{\text{CE}}$ is a class-weighted cross-entropy for a linear layer that directly discriminates the three classes from the same $\mathbf{o}_i$; the class weights are proportional to the inverse of the class frequencies in the training

data to stabilize learning of the minority class, level 2. The third term $\mathcal{L}_{\text{distill}}$ is the reconstruction loss $1 - \cos(\text{Dec}(\mathbf{e}_i), \mathbf{f})$ described above. $\bar{w}$ is the mean sample weight within a mini-batch; training images containing a blue striped shirt, the same domain as the evaluation data, are given weight 2 and the others weight 1. This training is completed within the per-axis vision encoder and is separate from the training of the importance path (Section 3.5) and the aggregator (Section 3.6).

3.4.2. Computing Axis Strength by CORAL Ordinal Regression

To the feature $\mathbf{o}_i$ obtained from the ordinal subspace we apply the CORAL (Consistent Rank Logits [22]) ordinal head trained with (1). CORAL decomposes the three-level ordinal classification into two cumulative binary decisions corresponding to the category boundaries, and applying the sigmoid function to the logit of each boundary gives the probability that the eccentricity level is at or above that boundary.

$$p_{i,1} = P(y_i \geq 1) = \sigma(z_{i,1}), \qquad p_{i,2} = P(y_i \geq 2) = \sigma(z_{i,2}) \tag{3}$$

The eccentricity strength s_i of each axis is defined as the expectation given by the cumulative probabilities.

$$s_i = \mathbb{E}[y_i] = p_{i,1} + p_{i,2} \in [0, 2], \qquad i = 1, \ldots, 6 \tag{4}$$

These are stacked into $\mathbf{s} = (s_1, \ldots, s_6)^\top$, the input to the downstream aggregator.

3.5. Axis Importance Prediction Mechanism (Importance Path)

The axis strength $\mathbf{s}$ obtained from the strength path (Section 3.4) expresses how far each visual axis deviates from the norm. It does not, however, retain the allocation of which visual elements are the main factors in judging the eccentricity of the outfit as a whole.

The importance path therefore extracts, for each axis, the semantic similarity $\mathbf{d}_i$ to text prompts corresponding to visual concepts, concatenates it with the visual feature $\mathbf{x}_i$ extracted by the strength path for the same axis, and feeds it to per-axis ridge regression to predict the axis importance $\hat{\mathbf{w}} \in \Delta^5$, which differs for each input outfit (Section 3.5.2). That is, this path comprises everything from the extraction of prompt similarities to the per-axis regression, and its output is $\hat{\mathbf{w}}$.

3.5.1. Prompt Bank and Centering

Eccentricity is not a single attribute but is formed by the combination of multiple visual elements such as color, pattern, silhouette, material, accessory, and styling.

We therefore design, for each visual axis, 20 visual concept prompts covering states from standard to highly eccentric, constructing a prompt bank of 120 text prompts across the six axes. Table 1 lists the features targeted by each axis and representative prompts.

From each prompt t_j, the text feature $\mathbf{t}_j \in \mathbb{R}^{768}$ is extracted with the text encoder of SigLIP2-base.

Table 1: Composition of the prompt bank. For each axis, one representative example each from the standard side and the deviant side is shown out of the 20 prompts.

Axis	Targeted features	Example prompts (standard / deviant)
Color	hue, saturation, color scheme	"an outfit in muted earth tones" / "an outfit with highly saturated neon colors"
Pattern	presence, type, and scale of patterns	"plain solid clothing with no patterns" / "clothing with a large floral print"
Silhouette	fit, volume, shape	"a slim fitted outfit with standard proportions" / "an extremely oversized baggy outfit"
Material	material, texture, gloss	"plain cotton fabric clothing" / "glossy leather clothing"
Accessory	amount, type, and prominence of accessories	"an outfit with no visible accessories" / "large statement earrings or necklace"
Styling	way of wearing, combination	"a conventionally coordinated everyday outfit" / "avant-garde experimental fashion styling"

In CLIP-family models, however, image features and text features are distributed in separate regions of the joint space (modality gap [19]). If similarities are used as they are in this state, the values are dominated not by differences among concepts but by a baseline response common to all prompts, and the comparison "which concept reacted strongly" does not hold. Because our method presents prompt similarities as the grounds of explanation, removing this common component is a precondition for the explanation to be meaningful, prior to any question of accuracy.

We therefore apply a centering operation that subtracts the mean of the 120 text features, $\bar{\mathbf{T}} = \frac{1}{120}\sum_j \mathbf{t}_j$, from each feature and renormalizes. On the image side, the preprocessed image x' is encoded by the original SigLIP2-base image encoder without LoRA adaptation into the feature $\mathbf{f}_{img}$, and the cosine similarity r_j to each prompt is computed.

$$\mathbf{t}'_j = \frac{\mathbf{t}_j - \bar{\mathbf{T}}}{\|\mathbf{t}_j - \bar{\mathbf{T}}\|_2}, \qquad r_j = \cos(\mathbf{f}_{img}, \mathbf{t}'_j) \tag{5}$$

Among the 120 resulting similarities, the 20 belonging to axis i are denoted $\mathbf{d}_i \in \mathbb{R}^{20}$.

The effect of centering should be evaluated not by downstream prediction accuracy but by whether the polarity of the prompts is correctly separated. Each axis of the prompt bank contains both sentences describing the eccentric side (e.g., "an outfit with many clashing colors") and sentences describing the non-eccentric side (e.g., "an outfit unified in a single color"). The condition for the bank to capture the concepts correctly is that the similarities of the two groups correlate in opposite directions with overall eccentricity. We therefore measured, on the 926 teacher images, the difference $\rho_Q - \rho_P$ between the mean correlation ρ_Q of the eccentric-side prompts and the mean correlation ρ_P of the non-eccentric-side prompts as the polarity separation (Table 2).

Table 2: Separation of prompt polarity by centering. ρ_Q and ρ_P are the mean rank correlations between overall eccentricity and the similarities of eccentric-side and non-eccentric-side prompts, respectively. Separation is their difference.

	Before centering			After centering		
Axis	ρ_Q	ρ_P	Sep.	ρ_Q	ρ_P	Sep.
Color	+0.44	−0.16	0.60	+0.44	−0.22	**0.66**
Pattern	+0.34	−0.09	0.43	+0.29	−0.26	**0.55**
Silhouette	+0.31	−0.20	0.51	+0.26	−0.34	**0.60**
Material	+0.30	−0.20	0.50	+0.22	−0.29	**0.51**
Accessory	+0.38	+0.12	0.27	+0.26	−0.15	**0.41**
Styling	+0.50	−0.18	0.69	+0.45	−0.48	**0.93**
Mean	+0.40	−0.08	0.48	+0.32	−0.29	**0.61**

Polarity separation improved on all six axes, and the mean rose from 0.48 to 0.61. On the accessory axis in particular, before centering the non-eccentric side had $\rho_P = +0.12$, the same sign as the eccentric side, so the bank could not distinguish polarity; after centering it becomes −0.15 and the sign is correctly reversed. Meanwhile, the prediction accuracy of axis importance itself does not change with or without centering (with standardized features and cross-validated regularization strength, the cross-validated Spearman rank correlations are 0.546 and 0.547). The justification for centering is therefore not accuracy improvement but the guarantee of explainability that prompt polarity functions correctly.

3.5.2. Predicting Axis Importance by Per-Axis Ridge Regression

The importance of axis i is considered to be determined by both the semantic cues concerning that axis and how far that axis deviates from the norm. We therefore use an independent ridge regression per axis and restrict the input to information concerning the axis in question. Specifically, for axis i, the similarities $\mathbf{d}_i \in \mathbb{R}^{20}$ to the 20 prompts of axis i in the prompt bank and the 128-dimensional feature $\mathbf{x}_i = [\mathbf{o}_i; \mathbf{z}_i] \in \mathbb{R}^{128}$ extracted by the strength path for axis i ($\mathbf{o}_i$ from the ordinal subspace and $\mathbf{z}_i$ from the distillation subspace) are concatenated, and

$$w_{\mathrm{raw},i} = \mathbf{a}_i^\top \begin{bmatrix} \mathbf{d}_i \\ \mathbf{x}_i \end{bmatrix} + b_i$$

is predicted. Because the similarities of other axes are not included in the input, the explanation "axis i is important" is guaranteed to be derived from the information of axis i itself. That this restriction does not harm accuracy is confirmed in Section 6.2.

Since the regression output $w_{\mathrm{raw},i}$ may contain negative or extremely small values, we normalize to sum one while leaving a minimum contribution for each axis.

$$\hat{w}_i = \frac{\max(w_{\mathrm{raw},i}, \epsilon)}{\sum_{j=1}^{6} \max(w_{\mathrm{raw},j}, \epsilon)} \tag{6}$$

The resulting $\hat{\mathbf{w}} = (\hat{w}_1, \ldots, \hat{w}_6) \in \Delta^5$ is the allocation expressing to what extent the impression of the outfit stems from each visual axis.

3.6. Eccentricity Distribution Prediction Mechanism (Aggregator and Distribution Generation)

The eccentricity of an outfit as a whole cannot be expressed by the linear weighted sum $\sum_i s_i \hat{w}_i$ of axis strengths alone. Even at the same color strength, the perceived eccentricity differs between the case where color is the main factor of the impression and the case where pattern or silhouette determines the impression. We therefore use an aggregator that takes $\mathbf{s}$ and $\hat{\mathbf{w}}$ as independent features and learns their combination nonlinearly. A preliminary study confirmed that replacing this with the weighted sum $\sum_i s_i \hat{w}_i$ degrades performance.

Adding the auxiliary visual feature, the aggregator input vector $\mathbf{z}_{\mathrm{agg}} \in \mathbb{R}^{47}$ is composed as follows.

$$\mathbf{z}_{\mathrm{agg}} = [\mathbf{s}, \hat{\mathbf{w}}, \mathbf{z}_{\mathrm{d}}] \tag{7}$$

Here $\mathbf{z}_{\mathrm{d}} \in \mathbb{R}^{35}$ is the auxiliary visual feature obtained by concatenating the reconstruction-regularized (distillation) subspaces of the six axes in the strength path and compressing them by principal component analysis to 90% explained variance; it complements fine-grained visual information that per-axis ordinal features alone do not retain well.

From the aggregated feature, the raw representative eccentricity $\mu_{\mathrm{raw}} = g(\mathbf{z}_{\mathrm{agg}})$ is predicted by a multilayer perceptron $g(\cdot)$. g can learn nonlinear combinations across axes and represent the eccentricity formation process that differs from outfit to outfit.

3.6.1. Calibration of the Representative Value and Generation of the Rating Distribution

To correct the scale difference between the teacher distribution at training time and the prediction distribution at inference, quantile-mapping calibration $\mu = Q_{\mathrm{target}}(F_{\mathrm{raw}}(\mu_{\mathrm{raw}}))$ is applied to μ_{raw}. F_{raw} is the cumulative distribution function constructed from the out-of-fold predictions on the training data, and Q_{target} is the inverse cumulative distribution function of the teacher distribution. Because the mapping is monotonically increasing, the rank order among samples is preserved.

The standard deviation σ, which expresses differences in perception among raters, is then determined from the calibrated representative value μ.

$$\sigma = a + b|\mu - 50| \tag{8}$$

Here a, b are coefficients estimated by least squares from the standard deviations of the rating distributions in the teacher set. This relation reflects the consensus structure in which ratings converge for extreme cases and disperse for intermediate cases; the analysis of human ratings that grounds it is given in Section 4.3. We give σ by a fixed two-coefficient rule rather than learning it per image because a preliminary study confirmed that, given the true μ, this rule alone brings the distribution prediction error below the human reproducibility ceiling (EMD 0.046 vs. 0.049), leaving little room for per-image prediction of σ.

Finally, the continuous distribution $f(t; \mu, \sigma)$ parameterized by (μ, σ) is discretized into the five rating intervals to obtain the final output $\mathbf{y} = (p_0, \ldots, p_4)$.

$$p_k = \int_{20k}^{20(k+1)} f(t;\mu,\sigma)\,dt, \qquad k = 0,\ldots,4 \tag{9}$$

3.7. Constructing the Recommendation List

Using the representative value μ predicted for each image in the candidate pool C, we select K items that cover its range at equal intervals. Specifically, we take as targets K equally spaced quantiles $q_1 < \cdots < q_K$ of the distribution of μ over the pool (q_1 the minimum and q_K the maximum) and select, in order, the unselected candidate closest to each quantile. The recommendation list thus contains, in steps, from the safest to the most eccentric outfit in the pool. Because every image in the pool shares the user's specified item, this selection presents a spread of impressions while staying within the range that fits the user's preferences. The distribution width σ is not used for recommendation selection in this study; it is output as supplementary information indicating how far ratings split for each recommended image.

3.8. Presentation of Explanations

Together with the predict, the proposed method presents the grounds of judgment in the following three units.

Per-axis strength and importance. The axis strength $\mathbf{s}$ of Section 3.4 and the axis importance $\hat{\mathbf{w}}$ of Section 3.5 are presented as they are for each of the six axes. Strength is an absolute quantity expressing how far each axis is from the norm, and importance is a ratio expressing how much of the image's impression is allocated to each axis.

Grounding prompts. For axes of high importance, the sentences with high similarity r_j among the 20 prompts of that axis are shown, presenting in language what the axis reacted to. Because the prompt bank contains equal numbers of standard-side and deviant-side sentences, the top sentences indicate not the magnitude of deviation but what the axis captured; the magnitude of deviation is carried separately by strength.

Referenced regions. Where in the image a grounding prompt reacted is visualized by gradient-times-input attribution [28]. Taking the patch representations at the last layer of the SigLIP2 vision encoder as the input, we differentiate the similarity r_j to the prompt with respect to the patch representations, take the inner product of gradient and representation as the contribution of each patch, and highlight the patches with the largest contributions. This is an existing technique of the same family as attribution for Transformers [29] and text-based decomposition of CLIP representations [30]; by combining it with per-axis prompts, our method presents in one sequence "which axis", "which concept", and "where in the image" reacted.

4. Datasets

This study uses the three kinds of data listed in Table 3. The collected images and the teacher set are used for training DualPath-SSA, and eval100 for evaluation. Their details are given below.

Table 3: Data composition and label sources. Per-axis strength labels are generated by the lightweight model, and the overall distribution and axis importance by the large reasoning model. Although axis importance is a per-axis label, it is assigned by the latter because it requires a holistic judgment involving the interaction of multiple elements. No label is assigned to the evaluation data.

Data	Images	Axis strength	Distribution / importance	Use
Collected images	7,416	qwen3.6	—	training of strength path
Teacher set	926	qwen3.6	Fable 5	training of importance path and aggregator
eval100	100	—	—	fixed-domain evaluation (human)

4.1. Training Data and Two Kinds of Teacher Signals

To train DualPath-SSA we collected 7,416 full-body outfit images. The images are publicly available outfit photographs collected through web image search, using search queries designed for each combination of visual axis and deviation level. Collecting by axis and level yields training data that covers, for each of the six axes, the range from standard to strongly deviant. The collected images are used solely for the computational analysis of this study (training and evaluation of the model); copies of the images themselves are neither published nor redistributed. In preprocessing, regions other than the person are removed (Section 3.3), and no processing aimed at identifying individuals is performed. Because eccentricity ratings cannot be obtained at scale by hand, this study uses two vision-language models of different character, with separate roles, to generate teacher signals.

- **Per-axis strength label generator**: a locally run vision-language model, `qwen3.6:35b-a3b-q4_K_M`. For each of the six axes—color, pattern, silhouette, material, accessory, and styling—it assigns the degree of deviation from the standard state on a three-level scale. Each axis judgment is a local decision based on observable visual features such as the number of colors or the type of pattern, and can be processed in bulk with a prompt of fixed criteria; it is therefore applied to all 7,416 collected images.
- **Overall distribution and axis importance label generator**: the large multimodal reasoning model Claude Fable 5 (Anthropic, released June 2026; accessed July 2026). It assigns a five-level eccentricity rating distribution for the outfit as a whole and the contribution of each axis to the overall impression (axis importance). Because these require holistic judgments that account for the interaction of multiple visual elements and for inter-rater variability, a model with high reasoning ability is used and the target is limited to the 926-image teacher set.

The two roles are complementary. The per-axis strength labels are used for ordinal training of the strength path (Section 3.4) and define "how far each visual element deviates". The overall distribution and axis importance labels are used for training the

importance path (Section 3.5) and the aggregator (Section 3.6) and define "which element grounded the judgment" and "how the whole is evaluated". The 926-image teacher set is a subset of the 7,416 collected images and carries both kinds of labels.

The instructions given to both models are summarized in Appendix A.

Neither teacher signal is assigned to the evaluation data eval100, which is never used for training or model selection. The evaluation images are likewise used solely for computational analysis and are not reproduced in this paper. For the example in Section 6.3, an image under a redistributable license is used separately.

4.1.1. Per-Axis Three-Level Labels

For each axis, $y_i \in \{0, 1, 2\}$ (0 = standard, 1 = slightly deviant, 2 = strongly deviant) is assigned. The judgment is based on observable visual features such as the number of colors, the type of pattern, material properties, and the deviation of the silhouette. The criteria are given in Table 4.

Table 4: Criteria for the per-axis three-level labels, summarized from the instruction given to the label-generating model (Appendix A). For each axis, three levels are defined: standard, slightly deviant, and strongly deviant.

Visual axis	Criteria (0 / 1 / 2)
Color	0 = only achromatic colors or a single neutral tone / 1 = exactly one chromatic color / 2 = two or more chromatic colors, or any neon color
Pattern	0 = completely solid / 1 = one pattern type on one garment / 2 = two or more pattern types, or patterns covering more than half of the outfit
Silhouette	0 = all garments in standard proportions / 1 = one garment deliberately non-standard (e.g., oversized) / 2 = two or more garments with extreme proportions, or one dramatically unusual shape
Material	0 = only opaque matte fabrics / 1 = a textured or lustrous fabric (e.g., leather, satin, lace) / 2 = transparent, metallic, PVC, fur, or feathers, or heavy embellishment
Accessory	0 = none or only plain items / 1 = one statement accessory / 2 = two or more statement accessories, or one avant-garde item
Styling	0 = all items in one style category / 1 = items from two style categories combined / 2 = three or more style categories, or a strong deliberate contrast

4.1.2. Overall Rating Distribution and Axis Importance Labels

The following two kinds of labels are assigned to the 926-image teacher set.

1. **Five-level eccentricity rating distribution:** the label-generating model is given, as anchors of the rating scale, reference images ranging from standard to extremely distinctive, and is then made to generate a probability distribution over the five-level rating space for each image. The distribution width is intended to express **differences in perception among human raters**, not the model's own uncertainty; the policy was to assign concentrated distributions to clearly standard or clearly eccentric cases and wide distributions to intermediate cases on which judgments split.
2. **Per-axis importance:** the contribution of each visual axis to the overall judgment is rated on a four-level scale of 0–3. The criterion is the counterfactual "how much would the overall impression change if only the target axis were replaced by a standard element". This allows the deviation of the axis itself, $\mathbf{s}$, and its contribution to impression formation, $\hat{\mathbf{w}}$, to be learned separately.

4.1.3. Label Reliability Check

The per-axis importance labels are subjective judgments without an explicit ground truth. We therefore re-annotated 90 images randomly drawn from the teacher set in a separate session under blind conditions, without reference to the initial labels, and compared the agreement against two external criteria (Table 5). First, we estimated the **chance level of agreement**. Keeping the frequency of each level in the re-annotated labels unchanged, we randomly permuted only their correspondence to the images, computed the agreement between the permuted labels and the initial labels, and repeated this 20,000 times. This corresponds to the agreement that a rater who assigns each level in the same proportion as the re-annotation, but without looking at the images, would reach by chance. Second, we computed the quadratically weighted Cohen's κ [31], a standard coefficient for ordinal scales. κ is corrected for chance agreement, and the interpretation scale in [32], which regards 0.61–0.80 as substantial agreement, is widely used.

Table 5: Reliability check of the per-axis importance labels by blind re-annotation (90 images × 6 axes = 540 judgments). The chance baseline is estimated by permutation preserving the marginal distribution of the re-annotated labels.

Metric	Observed	Chance	Verdict
Mean absolute difference (MAD)	**0.380**	0.944	$p < 10^{-4}$
Exact agreement	**65.6%**	35.1%	—
Agreement within one level	**96.9%**	76.7%	—
Agreement on the most important axis	**63.3%**	16.7%	—
Quadratically weighted κ	**0.725**	0 (by definition)	substantial

The mean absolute difference is 0.380, far from the chance baseline of 0.944 (95% interval 0.887–0.998; permutation test $p < 10^{-4}$, a 59.8% reduction relative to chance). Most disagreements stay within adjacent levels; judgments differing by more than one level account for only 3.1% of the total. The judgment of which axis contributed most agrees on 63.3% of images, about four times the chance level of 16.7%.

The quadratically weighted κ is 0.725, which falls in the substantial-agreement range (0.61–0.80) of [32]. We therefore judged that the labels have reproducibility at the level of "re-judging the same image leads to broadly the same conclusion" and can be used as teacher signals.

Reproducibility is, however, not uniform across axes. The per-axis κ is high for pattern (0.842), color (0.815), and silhouette (0.756) but decreases for accessory (0.637), material (0.561), and styling (0.401). Styling requires the holistic judgment of "mixing of styles" and is an axis on which criteria tend to waver even within a single annotator. This ordering corresponds well to the ordering of axis importance prediction performance shown in Section 6.2 (styling lowest among the six), suggesting that part of the downstream prediction accuracy is bounded not by the representational

power of the model but by the reproducibility of the teacher labels themselves.

4.2. Evaluation Data

For performance evaluation we use human rating data that is never used in training. **eval100** is an evaluation dataset of 100 outfit images. Ratings were collected in five surveys with 24, 15, 17, 13, and 13 raters, respectively, each survey presenting 20 images. Each rater answered eccentricity on a continuous scale of 0–100, and from the 13–24 ratings obtained per image we computed the representative value μ_{human}, the inter-rater standard deviation σ_{human}, and the histogram over five bins. This histogram is used as the ground-truth distribution when evaluating distribution prediction.

All 100 images in eval100 are outfits that include a blue striped shirt, making it an in-domain evaluation with a fixed common element. This setting lets us evaluate whether the model predicts eccentricity from fine differences in visual elements within an outfit, rather than relying on global cues such as the background or the presence of a specific item.

σ_{human} expresses the degree of agreement among raters: the smaller its value, the more the judgments agree; the larger, the more the ratings split. This study uses band partitioning based on σ_{human} as the basic unit of evaluation. The definition of the bands and the metrics used in each are given in Section 5.

4.3. Analysis of the Consensus Structure of Annotations

The rule for the distribution width σ in Section 3.6 is grounded in the consensus structure observed in human ratings. This section analyzes the human ratings of eval100 to show that structure. Note that the coefficients a, b of the rule are estimated only from the 926-image teacher set; the evaluation data is used to confirm the validity of the **form** of the rule.

4.3.1. Relation between Inter-Rater Variability and Eccentricity

Figure 2 shows the relation between the representative value μ_{human} and the inter-rater standard deviation σ_{human} of each outfit. Inter-rater variability does not change monotonically with eccentricity but shows a U-shaped structure, largest for intermediate cases and decreasing at both extremes. The Spearman rank correlation between the extremeness index $e = |\mu_{\text{human}} - 50|$ and the standard deviation is $\rho(e, \sigma_{\text{human}}) = -0.659$; the more extreme the case, the higher the agreement among raters. That is, the uncertainty of the ratings is not random noise but a structural property dependent on the representative value.

4.3.2. Reflection of the Consensus Structure in the Teacher Distributions

The same analysis on the teacher distributions generated under the policy of Section 4.1 gives

$$\rho(e, \sigma_{\text{teacher}}) = -0.832 \tag{10}$$

reproducing a trend in the same direction as the U-shaped consensus structure observed in the human ratings.

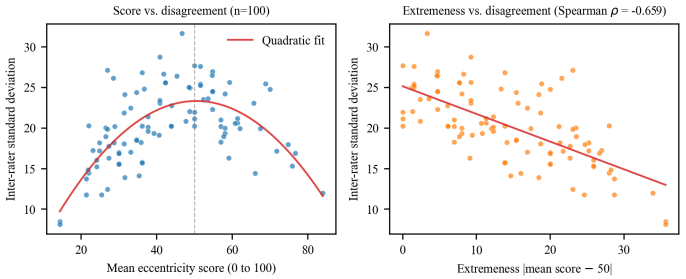


Figure 2: Inter-rater consensus structure on eval100 (100 images). Left: relation between the mean eccentricity score and the inter-rater standard deviation. Right: relation between extremeness $|\mu_{\text{human}} - 50|$ and the standard deviation. A negative correlation between extremeness and standard deviation is observed (Spearman $\rho = -0.659$).

This correlation reflects the design policy used when generating the teacher distributions and is not a property discovered independently from human rating data. However, since the same trend is confirmed on the independent eval100, the adopted distribution generation policy is considered consistent with actual human rating characteristics.

5. Evaluation Protocol

Before the experiments of Section 6, this section defines what is measured and by which metrics. In predicting subjective impressions, the human ratings used as ground truth themselves contain inter-rater variability, so the value of a metric is meaningful only when that variability is taken as the reference. Below we describe, in order, the definition of bands based on inter-rater agreement (Section 5.1), the metrics used (Section 5.2), the magnitude of differences worth discussing (Section 5.3), and the reference points and comparison methods (Section 5.4). Section 6 refers to them in the same order.

5.1. Evaluation Data and Definition of Bands

Evaluation uses the human rating data eval100 (100 images), which is not used for training or model selection (Section 4.2). From the human ratings of each image, the representative value μ_{human} and the inter-rater standard deviation σ_{human} are computed, and the images are partitioned into the following four bands.

- **Low-HA band** ($\sigma_{\text{human}} \leq 17$ and $\mu_{\text{human}} < 45$, 19 images): raters agree that the outfit is “not eccentric”.
- **High-HA band** ($\sigma_{\text{human}} \leq 17$ and $\mu_{\text{human}} \geq 45$, 5 images): raters agree that the outfit is “eccentric”.
- **MA band** ($17 < \sigma_{\text{human}} \leq 24$, 50 images): judgments split moderately.
- **LA band** ($\sigma_{\text{human}} > 24$, 26 images): judgments split widely.

The high-agreement region is further divided by the representative value because the two parts play different roles in recommendation. The high-HA band corresponds to “cases that many users clearly regard as eccentric” and defines the upper end of the recommendation list; the low-HA band corresponds to “safe cases” and

defines the lower end. However, since the high-HA band contains only 5 images, its values are shown in the tables for every metric below but differences in this band are not interpreted. The overall (ALL) value of every metric is computed over all 100 images.

5.2. Metrics

We use the ordering agreement for the prediction of the representative value, EMD for the prediction of the distribution, and the range of human-rated eccentricity for the recommendation list. No metric other than these three is used.

Ordering agreement on discriminable pairs (primary metric). In subjective evaluation, measurement error is also present on the ground-truth side. Correlation coefficients and RMSE ignore this and count differences that lie within rater variability as "hits" or "misses". In [33], the authors pointed out this problem in image quality assessment and proposed a framework that asks for the prediction of ordering only on stimulus pairs that differ significantly. Following them, with μ_i the mean human rating of image i and se_i its standard error computed from the number of raters, only pairs satisfying

$$|\mu_i - \mu_j| > \sqrt{\mathrm{se}_i^2 + \mathrm{se}_j^2} \tag{11}$$

are regarded as discriminable, and we measure the proportion of such pairs on which the ordering of the predicted values agrees with that of the human ratings. Because indistinguishable pairs are excluded, the metric does not depend on the spread within a band and can be compared directly across bands. The numbers of discriminable pairs are 99 in the low-HA band, 9 in the high-HA band, 899 in the MA band, 181 in the LA band, and 3,646 overall.

EMD (distribution metric). The quality of reproduction as a distribution is measured by the Earth Mover's Distance (EMD) [21, 23]; the smaller the value, the closer the predicted distribution is to the human rating distribution. Our EMD is defined as the mean absolute difference of cumulative distributions. Because the distribution width σ of our method is determined from the representative value by a fixed two-coefficient rule (Section 3.6), differences in EMD mainly reflect differences in the representative value μ.

Range of human-rated eccentricity (recommendation metric). The difference between the maximum and minimum μ_{human} of the images in a recommendation list of K items is measured as the diversity of subjective impression. In addition, the mean dissimilarity in SigLIP embedding space, Intra-List Diversity (ILD) [14], is reported as objective diversity, and the mean similarity to the query as relevance.

5.3. Magnitude of Differences Worth Discussing

Subjective impressions such as eccentricity fluctuate between and within raters, and human ratings cannot be obtained at scale; our evaluation data too consists of 100 images. Before stating results, we therefore empirically measure the lower bound of differences worth discussing. Repeating the same evaluation eight times with the configuration unchanged and only the set of random seeds of the aggregator replaced, the standard deviation of the ordering agreement was 0.2 points overall, and 2.0, 0.6, 1.7, and 11.9 points in the low-HA, MA, LA, and high-HA bands, respectively. The standard deviation of EMD was 0.001 overall and 0.001–0.006 per band. In the discussion below we take twice the standard deviation as the lower bound for each band and do not attribute differences below it to differences in configuration: 0.4 points for the overall ordering agreement and 0.002 for the overall EMD. Since the ordering agreement in the high-HA band moves by 11.9 points merely by changing seeds, no difference in that band is interpreted, as stated in the previous section.

5.4. Reference Points and Comparison Methods

Neither the ordering agreement nor EMD can be judged good or bad from its absolute value alone. We therefore present the following two reference points for both metrics and read each value by where it lies between them.

- **Reference: mean distribution**: an uninformative lower bound that outputs the mean of the teacher distributions for every image. It gives no ordering, so the ordering agreement is undefined.
- **Reference: human reproducibility ceiling**: the reproducibility limit of the human ratings themselves. For the ordering agreement, the raters of each image are split into two independent halves, and the extent to which the mean of one half reproduces the ordering of the mean of the other half is averaged over 400 splits. For EMD, the same number of raters is independently resampled from the human rating histogram of each image and the distance to the original distribution is measured by parametric bootstrap. Because sampling error is present on both the prediction and ground-truth sides, this underestimates the attainable upper bound, and attainment rates are optimistic.

The human reproducibility ceiling differs from band to band; the more the judgments split, the lower the ceiling itself. This is why the absolute values of the metrics must not be compared directly across bands.

To isolate the contribution of each component of the proposed method, the following methods are compared under the same calibration and the same distribution generation.

- **Baseline: single-scalar ridge**: linear ridge regression that regresses the representative value directly from the 192-dimensional feature obtained by concatenating the ordinal-subspace features of the six axes. It does not separate strength and importance and has no intermediate representation.
- **DualPath-SSA (strength only)**: an ablated configuration that removes the axis importance $\hat{\mathbf{w}}$ from the aggregator input and uses the 41 dimensions $[\mathbf{s}, \mathbf{z}_{\mathrm{d}}]$.
- **DualPath-SSA (strength + importance)**: the proposed method using the 47-dimensional input of Section 3.6.

For the comparison of recommendation lists, for the candidate pool sharing the specified item "blue striped shirt" (eval100, $N = 100$), the 19 images of the low-HA band are used as queries, and a recommendation list of $K = 8$ items is generated for each

query. Queries are taken from the low-HA band because the situation of widening recommendations starting from a safe outfit is the application scenario assumed in this study. To avoid circularity, only the **predicted** values of the proposed method are used for selection, and the human **ground-truth** values are used for evaluation. The comparison methods are similarity top-k, which selects the K items most similar to the query; MMR [4], which balances relevance and diversity with a mixing coefficient λ; and random selection. MMR is compared at two settings: $\lambda = 0.0$, which maximizes diversity only, and $\lambda = 0.5$, its original setting weighting both equally. A comparison that chooses λ by consulting the evaluation metric is not performed because it would be unfair. As a reference point, an oracle that performs quantile coverage with the human ground-truth values is also shown.

6. Results and Discussion

This section is organized to correspond to the three contributions of Section 1. Section 6.1 presents the accuracy of eccentricity distribution prediction (Contribution 2), Section 6.2 the validity of the axis strength and axis importance mechanisms (Contribution 3), Section 6.3 an example of the explanations built from them (Contribution 3), Section 6.4 the subjective diversity of recommendation lists (Contribution 1), and Section 6.5 the limitations and future work.

6.1. Accuracy of Eccentricity Distribution Prediction

This section supports the claim that "predicting strength and importance separately costs nothing in accuracy". Table 6 shows the per-band ordering agreement and EMD alongside the reference points, the baseline, and the ablated configuration. All values come from a single run in which calibration, distribution generation, and the set of random seeds are shared.

We first confirm the differences that exceed the bounds set in Section 5.3. The overall ordering agreement of the proposed method (strength + importance) is 82.6%, reaching 94% of the human reproducibility ceiling of 88.2%. Its EMD is 0.101, and the difference of 0.055 from the mean distribution (0.156) far exceeds the bound of 0.002. That is, eccentricity, a subjective impression, is quantified at a level that reproduces the judgment of the rater population.

Next, the proposed method outperforms the single-scalar ridge beyond the bound of differences worth discussing. The ordering agreement is 82.6% versus 77.8% (a difference of 4.8 points against a bound of 0.4), and the EMD is 0.101 versus 0.107 (0.006 against 0.002). Per band, the low-HA band shows 64.6% versus 73.7% (bound 4.0) and the MA band 74.7% versus 81.3% (bound 1.2), both exceeding the bound. Compared with regressing to a single scalar without separating strength and importance, the present configuration with per-axis ordinal regression and nonlinear aggregation is superior in accuracy.

In the comparison with the ablated configuration, adding importance to the strength-only 81.4% gives 82.6% for the proposed method, and the difference of 1.2 points exceeds the overall bound of 0.4. Per band, the MA band (79.4% versus 81.3%, bound 1.2) exceeds the bound, whereas the low-HA band (71.7% versus 73.7%, bound 4.0) and the LA band (70.7% versus 73.5%, bound 3.3) are within it. Adding axis importance to the aggregator thus at least does not harm the accuracy of the representative value and slightly improves it overall. There is no cost in accuracy for retaining the intermediate representation used for explanation. The contribution of axis importance to accuracy is thus small; its main role is to provide the per-axis allocation used for explanation (Section 6.3), not to improve accuracy.

In the LA band, the ordering agreement of the proposed method is 73.5% against a human reproducibility ceiling of 74.3%, an attainment rate of 99%. Ordering discrimination in the LA band is thus bounded not by the model's ability but by the agreement of

Table 6: Accuracy of eccentricity prediction on eval100. (a) Ordering agreement on discriminable pairs (%, higher is better); (b) EMD (lower is better). The first row of each block gives the number of items aggregated (image pairs for (a), images for (b)). "Strength only" is the ablated configuration that removes the axis importance $\hat{\mathbf{w}}$ from the aggregator input; "strength + importance" is the proposed method. The high-HA band has only 5 images and 9 pairs, and its ordering agreement moves by 11.9 points merely by changing random seeds (Section 5.3); its values are shown but not interpreted. ALL is computed over all 100 images. Bold marks the best value excluding reference points.

	Low-HA	High-HA	MA	LA	ALL
(a) Ordering agreement (%) ↑					
Discriminable pairs	99	9	899	181	3646
Reference: human ceiling	79.5	93.5	87.9	74.3	88.2
Baseline: single-scalar ridge	64.6	66.7	74.7	**75.7**	77.8
DualPath-SSA (strength only)	71.7	77.8	79.4	70.7	81.4
DualPath-SSA (strength + importance)	**73.7**	88.9	**81.3**	73.5	**82.6**
(b) EMD ↓					
Images	19	5	50	26	100
Reference: mean distribution	0.248	0.225	0.144	0.099	0.156
Reference: human ceiling	0.033	0.032	0.049	0.064	0.049
Baseline: single-scalar ridge	**0.091**	0.201	0.105	**0.104**	0.107
DualPath-SSA (strength only)	**0.091**	0.182	0.096	0.112	0.104
DualPath-SSA (strength + importance)	0.094	0.176	**0.093**	0.108	**0.101**

the human ratings themselves. The EMD in the LA band, however, is 0.108, above the mean distribution's 0.099. In this band human judgments themselves split, and the human reproducibility ceiling of 0.064 is also larger than in other bands. Since ordering discrimination has reached the ceiling, the remaining difference stems from the prediction of the distribution width σ; this point is discussed in Section 6.5.

6.2. Validity of Each Mechanism

This section supports the claim that "axis strength and axis importance are each independently valid". For the intermediate representations presented as explanations not to be arbitrary, the output of each mechanism must be valid with respect to its own teacher. Agreement with the teacher labels shows that each mechanism is consistent with its training signal; it does not by itself show agreement with human perception of each axis. Agreement with human perception is evaluated at the level of overall eccentricity, against the ratings of 82 human raters (Section 6.1).

Axis strength. The axis strength prediction mechanism was evaluated on 1,112 held-out images not used for training (6,672 labels over six axes; Table 7). The balanced accuracy over all labels is 87.7%. The per-level recall is 93.9% for standard and 88.9% for strongly deviant, whereas the intermediate level, slightly deviant, is the lowest at 80.4%. Per axis, pattern is the highest at 92.4% and styling the lowest at 79.6%. Styling requires a holistic judgment of the appropriateness of the combination and is harder than pattern or color, which are determined by local visual cues.

Table 7: Held-out accuracy of the axis strength prediction mechanism (1,112 images, six axes). Per axis: balanced accuracy; per level: recall. Overall is the mean of the per-level recalls.

	Color	Pattern	Silh.	Material	Access.	Styling	Overall
Bal. acc. (%)	89.9	**92.4**	84.8	82.5	89.9	79.6	87.7
	Standard (3,170)			Slightly dev. (2,596)		Strongly dev. (906)	
Recall (%)	93.9			80.4		88.9	

Axis importance. The axis importance prediction mechanism was evaluated by five-fold cross-validation on the 926-image teacher set, repeated 10 times with different fold seeds, using the Spearman rank correlation with the per-axis importance labels (Table 8). The adopted configuration (20-dimensional own-axis prompt similarities + 128-dimensional own-axis visual features) attains a six-axis mean of +0.563 (standard deviation over the 10 repetitions 0.004), far exceeding on every axis the 95th percentile of the permutation null distribution (0.063–0.066). Per axis, pattern is the highest at +0.67 and styling the lowest at +0.44.

Table 8: Input configurations of the axis importance prediction mechanism and cross-validated accuracy within the teacher set (per-axis Spearman rank correlation, mean of 10 repetitions). All inputs are restricted to the information of the axis in question. Bold marks the adopted configuration. The 95th percentile of the permutation null distribution is 0.063–0.066.

Input	Color	Pattern	Silhouette	Material	Accessory	Styling	Mean
Prompt similarity 20d	+0.60	+0.63	+0.53	+0.44	+0.45	+0.35	+0.500
Visual feature 128d	+0.61	+0.65	+0.59	+0.54	+0.48	+0.44	+0.552
20d + 128d	**+0.62**	**+0.67**	**+0.60**	**+0.55**	**+0.50**	**+0.44**	**+0.563**

It is also confirmed that both semantic cues and visual deviation are necessary. Prompt similarities alone give +0.500 and visual features alone +0.552, while their combination, +0.563, exceeds visual features alone in 10 out of 10 repetitions (mean difference +0.011, standard deviation 0.002). In terms of the norm of the standardized regression coefficients, the share of prompt similarities is 28–48% depending on the axis, so they are not absorbed by the visual features. Because the input is restricted to the axis in question, the explanation "axis *i* is important" is derived solely from the information of axis *i*.

6.3. Example of Presented Explanations

This section shows how the intermediate representations validated in the previous section are presented for a single image in the form of Section 3.8. Figure 3 shows an example.

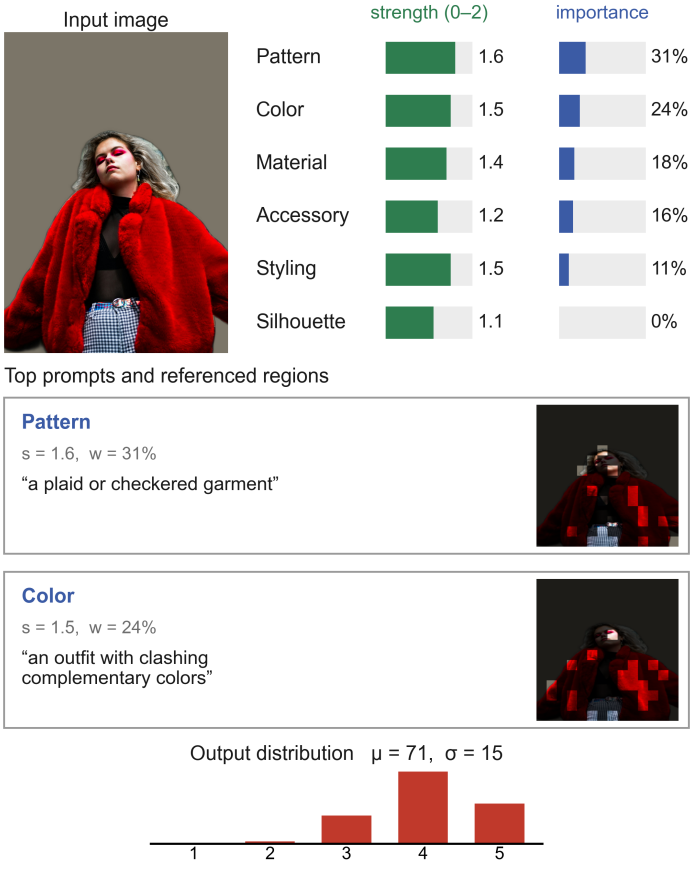


Figure 3: Example of the inference process and presented explanation for one image. Top: the input image and the axis strengths (0–2) and axis importances (sum 1) of the six axes. Middle: the grounding prompts and referenced regions of the two most important axes. Bottom: the output distribution (μ, σ). The input image is a photograph from Unsplash (photo: Romane Gautun, Unsplash License) after the preprocessing of our method.

For this image, the strength of pattern is predicted at 1.6 with an importance of 31%, and the strength of color at 1.5 with an importance of 24%; the representative eccentricity is 71 and the distribution width 15. From this presentation the user can read that the outfit was "judged eccentric mainly because of pattern and color". As grounding prompts, "a plaid or checkered garment" for the pattern axis and "an outfit with clashing complementary colors"

for the color axis appear with high similarity, and both correspond to the image content (the checkered trousers, and the contrast of the red coat against blue and black). Their referenced regions also match: the pattern axis highlights the checkered trousers and the color axis the red fur coat, so the referenced regions differ between axes. In this way, for each axis a prompt that correctly describes the image appears with high similarity and its reacted region corresponds to it, confirming that the explanation functions. In this way, for a single image, the method presents in one sequence "which axis", "how much it deviates", "which concept", and "where in the image" reacted.

6.4. Subjective Diversity of Recommendation Lists

This section supports the claim that "diversification in embedding space does not diversify subjective impression, whereas selection based on predicted eccentricity does". Table 9 shows the results of the recommendation selection experiment under the setting of Section 5.4.

Table 9: Subjective and objective diversity of recommendation lists ($K = 8$), averaged over 19 queries. Selection uses only predicted values; evaluation uses human ground-truth values. Range and standard deviation are on the 0–100 scale; coverage is the number of the five bins covered. The rightmost column is the difference in range from the proposed method; * marks cases where the 95% interval of the paired per-query difference, bootstrapped 2,000 times over queries, excludes zero.

	Subjective (human GT)			Objective		
Selection method	Range↑	SD↑	Cov.↑	ILD↑	Rel.	Diff.
Similarity top-k	33.7	10.9	2.42	0.179	**0.839**	−35.5*
MMR ($\lambda = 0.0$)	33.2	10.1	2.00	**0.326**	0.715	−36.0*
MMR ($\lambda = 0.5$)	42.5	14.2	2.95	0.240	0.813	−26.7*
Random	43.1	14.3	2.91	0.253	0.752	−26.1*
DualPath-SSA (proposed)	**69.2**	**19.9**	3.95	0.267	0.753	—
Reference: oracle (GT quantiles)	69.6	20.3	**4.00**	0.275	0.749	+0.4

The proposed method exceeds all comparison methods with a range of human-rated eccentricity of 69.2. The differences are +26.1 against random selection (95% interval [24.9, 27.1]), +26.7 against MMR at its original setting $\lambda = 0.5$ ([21.4, 31.7]), and +35.5 against similarity top-k ([31.1, 39.8]); none of the intervals includes zero. The difference from the oracle that uses the ground-truth values directly is −0.4 ([−1.1, +0.0]), so selection by predicted values is statistically indistinguishable from selection with knowledge of the ground truth.

The more important observation is that objective diversity and the diversity of subjective impression do not point in the same direction. MMR at $\lambda = 0.0$, which maximizes diversity only, has the largest ILD of all methods, 0.326, so diversification in embedding space works as intended. Yet at this setting the range of human-rated eccentricity is 33.2, below the 43.1 of random selection, and the number of covered bins, 2.00, is the smallest. Even at the original setting $\lambda = 0.5$, the range of 42.5 does not reach random selection. Looking at individual queries as well, the eight images chosen by MMR ($\lambda = 0.5$) concentrate on calm outfits close to the query, and those of MMR ($\lambda = 0.0$), although scattered in the embedding, span only a narrow range of human ratings. In contrast, the eight images of the proposed method spread in steps from an outfit as safe as the query to a flamboyant one. Because the candidate pool is fixed by the common element of a blue striped shirt, distances in embedding space mainly reflect differences such as background and small accessories that do not contribute to impression formation. Widening distances in embedding space does not coincide with widening the range of impressions that humans perceive.

In ILD itself, MMR ($\lambda = 0.0$) exceeds the proposed method (0.267). The advantage of the proposed method is obtained not on the metric of objective diversity but in the range of human-rated impression. Relevance is 0.753 for the proposed method against 0.839 for similarity top-k, so the general trade-off in which diversification lowers fit exists in our method as well. However, because every image in the candidate pool shares the specified item, this decrease stays within the range that fits the user's preferences. Note that a wider range of impressions does not by itself imply better recommendations or higher user satisfaction. What these results show is that the range of impressions in a list can be controlled through the predicted eccentricity; whether users prefer such lists remains to be examined in a user study.

6.5. Limitations and Future Work

This study has four limitations; for each we state its scope and which conclusions are unaffected.

Size of the evaluation data. Eccentricity is a subjective impression, and as long as its ground truth is obtained by human surveys, the data is bound to be small. Our evaluation data consists of 100 images, of which the high-agreement, high-score band contains 5. Against this constraint, we adopted the policy of empirically measuring, from repetitions differing only in random seed, the lower bound of differences worth discussing and treating only differences beyond it as results (Section 5.3). The findings shown under this policy—the range of impressions in recommendation lists, the superiority of separate prediction over single-scalar regression, and the validity of each mechanism—all far exceed that bound. This bound, however, reflects only the variation due to random seeds, not the sampling variation of the evaluation images; with a different set of 100 images, the values could change by more than this bound. For the comparison of recommendation lists, the differences are also supported by bootstrap intervals over queries (Table 9). The relative merits of methods within the high-HA band, however, cannot be determined with these data.

The evaluation data are limited to a single item domain. All 100 evaluation outfits share a blue striped shirt. This setting was chosen deliberately: with a common item fixed, eccentricity must be discriminated from differences in the other visual elements rather than from the garment category, which matches the application scenario of recommending outfits that share a specified item. It also means, however, that we have not verified whether the accuracy obtained here holds for outfits with other garments, colors, and styles. The training data cover diverse outfits (Section 4.1), but evaluation on a more diverse set of human-rated outfits is future work.

Prediction of the distribution width σ is constrained by the range of the teacher labels. The distribution widths of the teacher set are concentrated in a narrow range (mean 15.9, standard deviation 2.6) and contain almost none of the high-variance cases that

exist in the LA band. For this reason the EMD in the LA band is larger than that of the mean distribution (0.108 versus 0.099). The cause lies not in the model structure but in the range of the teacher labels, and can be resolved by a survey design that includes high-variance cases. Our experiments, however, do not directly demonstrate the value of outputting a distribution. Given the true μ, the fixed two-coefficient rule alone falls below the human reproducibility ceiling (Section 3.6), so the accuracy to be gained by refining σ is small. The role of the distribution width lies elsewhere than accuracy. The MA and LA bands, in which judgments split, account for 76% of the evaluation data, and for these cases, hitting the representative value of the population does not fit individual users. Being able to identify a case as "prone to split" is itself the starting point for adapting recommendations to each user's sensibility. Because our method retains per-axis strength and importance as intermediate representations, re-predicting axis importance per individual would extend it to presenting diversity aligned with preferences such as "I want eccentricity in the direction of color". Realizing and validating this personalization is future work.

The validity of the explanations is verified against teacher labels. We confirmed that axis strength and axis importance are valid with respect to their respective teacher labels, but whether the presented explanations agree with human grounds of judgment must be verified by collecting human per-axis annotations. Among the presented elements, the axis strength and axis importance are the inputs of the computation that yields the representative value (Section 2.5), so the statement of which axis contributed and by how much reflects the actual computation of the model. The grounding prompts and referenced regions, in contrast, present afterwards what each axis responded to. The visualization of referenced regions relies on an existing attribution technique, and its faithfulness is not verified in this study.

7. Conclusion

The objective of this study was to realize recommendation diversity based on subjective sensibility, and to that end we set the task of simultaneously satisfying three requirements: preference fit, subjective diversity, and explainability. Under the hypothesis that eccentricity can be decomposed into visual axes and interpreted, we proposed DualPath-SSA, which separately predicts the strength and importance of each of six axes and aggregates them to output a five-level probability distribution.

For evaluation, taking into account that the human ratings used as ground truth themselves fluctuate, we adopted as the primary metric the ordering agreement restricted to image pairs that differ significantly given rater variability, and compared methods after empirically measuring, from repetitions differing only in random seed, the lower bound of differences worth discussing. The findings, corresponding to the three requirements, are the following three points.

First, in a candidate pool sharing a specified item, diversification in embedding space does not diversify subjective impression (Table 9). The range of human-rated eccentricity in the lists produced by MMR ($\lambda = 0.0$), which maximizes intra-list dissimilarity, was 33.2, below the 43.1 of random selection. In contrast, the range of the proposed method, which covers the range of predicted eccentricity, was 69.2, exceeding the 42.5 of MMR ($\lambda = 0.5$) by 26.7 points and reaching the level of the oracle using human values directly (69.6). A spread of impressions can be presented inside the range that fits the user's preferences.

Second, there is no cost in accuracy for predicting strength and importance separately (Table 6). The ordering agreement on discriminable image pairs was 82.6%, exceeding the 77.8% of ridge regression to a single scalar by 4.8 points (lower bound of differences worth discussing: 0.4 points). The human reproducibility ceiling, obtained by splitting the raters into two independent groups and measuring how far one group reproduces the ordering of the other, is 88.2%, and the proposed method corresponds to about 94% of it. The distribution prediction error (EMD) is 0.101, between the uninformative reference point of 0.156 and the human reproducibility ceiling of 0.049.

Third, the presented intermediate representations are not arbitrary. Axis strength attained a balanced accuracy of 87.7% on 1,112 images not used for training, and axis importance a Spearman rank correlation of +0.563 in cross-validation within the teacher set (Table 8). Restricting the input of importance to the axis in question does not reduce accuracy, and a configuration in which the grounds of explanation do not leak from other axes holds at the same time. This allows the method to present, for a single image, which axis deviates by how much, and to which concept and where in the image that axis reacted. In the example of Figure 3, the pattern axis references the checkered trousers and the color axis the red fur coat, suggesting that, through the combination of per-axis prompts and attribution, each axis acquires features from the regions relevant to the judgment on that axis.

In summary, by decomposing eccentricity into per-axis strength and importance and predicting them, it was shown that the diversity of recommendation can be controlled at the level of subjective impression and that its grounds can be explained per axis.

On the other hand, the value of the distribution width output by our method is not directly demonstrated by the experiments in this paper. Because the widths of the teacher distributions are concentrated in a narrow range, the predicted distributions in the LA band have a larger EMD than the uninformative mean-distribution reference (0.108 versus 0.099; Section 6.5). The distribution width expresses the degree to which raters disagree, and the bands in which judgments split account for 76% of the evaluation data. In these bands, users differ in how they perceive eccentricity, and which axes they weight and by how much is presumably individual. Because our method retains per-axis strength and importance as intermediate representations, re-predicting axis importance per individual would extend it to presenting diversity aligned with that person's perception. The distribution width is an indicator of the cases in which this extension is needed, and the decomposition into six axes provides a handle for receiving preferences such as "I want eccentricity in the direction of color" on a per-axis basis. Moreover, whether presenting explanations to users leads to higher acceptance and satisfaction with recommendations has not been verified in this study. Together with expanding the evaluation data, future work will pursue the realization of this personalization and a user evaluation of recommendation including the presentation of explanations.

Conflict of Interest The authors declare no conflict of interest.

Funding This paper was supported by the Research Institute Support Fund of the Art Research Center, Ritsumeikan University.

Data Availability The analysis code and the trained model weights are available from the authors on reasonable request. The evaluation data cannot be shared because it contains personal information obtained from the human participants of the rating surveys. The outfit images used for training and evaluation are publicly available web images used solely for information analysis (training and evaluation of the model); the images themselves are neither reproduced nor redistributed, and their rights remain with the respective rights holders.

Use of AI Tools Vision–language models were used to generate the training labels of this study, namely the axis-wise strength labels and the overall distribution and axis importance labels; the specific models are described in Section 4.1.

Acknowledgment The authors thank all the participants who took part in the rating surveys.

Appendix A. Summary of the Instructions for Teacher Label Generation

This appendix summarizes the instructions given to the label-generating models described in Section 4.1.

A.1 Per-axis three-level labels (Qwen3.6). The model received one outfit image and a fixed instruction. The instruction asked the model to:

- act as a fashion expert labeler;
- score each of the six axes independently, using countable and observable criteria rather than subjective impression;
- reply only with the six scores (0, 1, or 2).

The criteria for each axis are given in Table 4. Images were resized to at most 512×512 pixels, and the temperature was set to 0.1.

A.2 Per-axis importance labels (Claude Fable 5). For each axis, the model rated how much the overall eccentricity would change if only the elements of that axis were replaced with ordinary ones:

- 0: almost no change;
- 1: a small change (a supporting factor);
- 2: a clear change (one of the main factors);
- 3: a large change (the dominant factor).

Only one axis could receive 3, or two when they were clearly comparable. A completely basic outfit could receive 0 on all axes. When uncertain, the lower level was to be chosen.

A.3 Five-level rating distributions (Claude Fable 5). For each image, the model assigned a probability distribution over five bins of the eccentricity scale: 0–20, 20–40, 40–60, 60–80, and 80–100. The instructions were as follows.

- *Scale.* Reference images accompanied by their mean ratings fixed the scale. They ranged from a completely basic outfit, such as a plain T-shirt with jeans, to costume-level outfits in which three or more axes are extreme at the same time.
- *Width.* The width of a distribution expresses how much human raters would disagree, not the model's own uncertainty. Clearly basic and clearly costume-level images receive concentrated distributions. Wide distributions are used only for intermediate images on which judgments are expected to split.
- *Procedure.* Images were labeled in batches of 12 to 15. Before labeling, previously labeled images were re-judged without seeing their labels to confirm that the scale had not drifted.

References

[1] K. Yamashita, F. Harada, H. Shimakawa, "Clustering-Based Analysis of the Relationship Between Clothing Attributes and Ease of Styling," in 2025 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA), 1–6, IEEE, 2025, doi:10.1109/ACDSA65407.2025.11166535.

[2] B. M. Sarwar, G. Karypis, J. A. Konstan, J. Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms," in Proceedings of the 10th International Conference on World Wide Web (WWW '01), 285–295, ACM, 2001, doi:10.1145/371920.372071.

[3] H. Kwon, J. Han, K. Han, "ART (Attractive Recommendation Tailor): How the Diversity of Product Recommendations Affects Customer Purchase Preference in Fashion Industry?" in Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), 2573–2580, ACM, 2020, doi:10.1145/3340531.3412687.

[4] J. G. Carbonell, J. Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries," in Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 335–336, ACM, 1998, doi:10.1145/290941.291025.

[5] C.-N. Ziegler, S. M. McNee, J. A. Konstan, G. Lausen, "Improving Recommendation Lists Through Topic Diversification," in Proceedings of the 14th International Conference on World Wide Web (WWW '05), 22–32, ACM, 2005, doi:10.1145/1060745.1060754.

[6] P. Dokoupil, L. Boratto, L. Peska, "User Perceptions of Diversity in Recommender Systems," in Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '24), 212–222, ACM, 2024, doi:10.1145/3627043.3659555.

[7] J. L. Herlocker, J. A. Konstan, J. Riedl, "Explaining Collaborative Filtering Recommendations," in Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00), 241–250, ACM, 2000, doi:10.1145/358916.358995.

[8] N. Tintarev, J. Masthoff, "A Survey of Explanations in Recommender Systems," in 2007 IEEE 23rd International Conference on Data Engineering Workshop (ICDEW), 801–810, IEEE, 2007, doi:10.1109/ICDEW.2007.4401070.

[9] Y. Zhang, X. Chen, "Explainable Recommendation: A Survey and New Perspectives," Foundations and Trends in Information Retrieval, **14**(1), 1–101, 2020, doi:10.1561/1500000066.

[10] G. Simmel, "Fashion," American Journal of Sociology, **62**(6), 541–558, 1957, doi:10.1086/222102, originally published in International Quarterly, vol. 10, pp. 130–155, 1904.

[11] J. Turney, editor, Fashion Crimes: Dressing for Deviance, Bloomsbury Visual Arts, London, 2019.

[12] X. Zou, K. Pang, W. Zhang, W. Wong, "How Good Is Aesthetic Ability of a Fashion Model?" in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 21168–21177, IEEE, 2022, doi:10.1109/CVPR52688.2022.02052.

[13] K. Gray, P. Schmitt, N. Strohminger, K. S. Kassam, "The Science of Style: In Fashion, Colors Should Match Only Moderately," PLOS ONE, **9**(7), e102772, 2014, doi:10.1371/journal.pone.0102772.

[14] B. Smyth, P. McClave, "Similarity vs. Diversity," in Case-Based Reasoning Research and Development (ICCBR 2001), volume 2080 of *Lecture Notes in Computer Science*, 347–361, Springer, 2001, doi:10.1007/3-540-44593-5_25.

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, "Learning Transferable Visual Models from Natural Language Supervision," in Proceedings of the 38th International Conference on Machine Learning (ICML), volume 139 of *Proceedings of Machine Learning Research*, 8748–8763, PMLR, 2021.

[16] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, "Sigmoid Loss for Language Image Pre-Training," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 11941–11952, IEEE, 2023, doi:10.1109/ICCV51070.2023.01100.

[17] S. Hentschel, K. Kobs, A. Hotho, "CLIP Knows Image Aesthetics," Frontiers in Artificial Intelligence, **5**, 976235, 2022, doi:10.3389/frai.2022.976235.

[18] J. Wang, K. C. K. Chan, C. C. Loy, "Exploring CLIP for Assessing the Look and Feel of Images," Proceedings of the AAAI Conference on Artificial Intelligence, **37**(2), 2555–2563, 2023, doi:10.1609/aaai.v37i2.25353.

[19] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, J. Y. Zou, "Mind the Gap: Understanding the Modality Gap in Multi-Modal Contrastive Representation Learning," in Advances in Neural Information Processing Systems (NeurIPS), volume 35, 17612–17625, 2022, doi:10.52202/068431-1280.

[20] X. Geng, "Label Distribution Learning," IEEE Transactions on Knowledge and Data Engineering, **28**(7), 1734–1748, 2016, doi:10.1109/TKDE.2016.2545658.

[21] L. Hou, C.-P. Yu, D. Samaras, "Squared Earth Mover's Distance-Based Loss for Training Deep Neural Networks," arXiv:1611.05916, 2016, doi:10.48550/arXiv.1611.05916.

[22] W. Cao, V. Mirjalili, S. Raschka, "Rank Consistent Ordinal Regression for Neural Networks with Application to Age Estimation," Pattern Recognition Letters, **140**, 325–331, 2020, doi:10.1016/j.patrec.2020.11.008.

[23] H. Talebi, P. Milanfar, "NIMA: Neural Image Assessment," IEEE Transactions on Image Processing, **27**(8), 3998–4011, 2018, doi:10.1109/TIP.2018.2831899.

[24] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," Nature Machine Intelligence, **1**(5), 206–215, 2019, doi:10.1038/s42256-019-0048-x.

[25] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, P. Liang, "Concept Bottleneck Models," in Proceedings of the 37th International Conference on Machine Learning (ICML), volume 119 of *Proceedings of Machine Learning Research*, 5338–5348, PMLR, 2020.

[26] F. Doshi-Velez, B. Kim, "Towards a Rigorous Science of Interpretable Machine Learning," arXiv:1702.08608, 2017, doi:10.48550/arXiv.1702.08608.

[27] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in International Conference on Learning Representations (ICLR), 2022.

[28] A. Shrikumar, P. Greenside, A. Shcherbina, A. Kundaje, "Not Just a Black Box: Learning Important Features Through Propagating Activation Differences," arXiv:1605.01713, 2016.

[29] H. Chefer, S. Gur, L. Wolf, "Transformer Interpretability Beyond Attention Visualization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 782–791, IEEE, 2021, doi:10.1109/CVPR46437.2021.00084.

[30] Y. Gandelsman, A. A. Efros, J. Steinhardt, "Interpreting CLIP's Image Representation via Text-Based Decomposition," in International Conference on Learning Representations (ICLR), 2024.

[31] J. Cohen, "Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit," Psychological Bulletin, **70**(4), 213–220, 1968, doi:10.1037/h0026256.

[32] J. R. Landis, G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," Biometrics, **33**(1), 159–174, 1977, doi:10.2307/2529310.

[33] L. Krasula, K. Fliegel, P. Le Callet, M. Klíma, "On the Accuracy of Objective Image and Video Quality Models: New Methodology for Performance Evaluation," in 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), 1–6, IEEE, 2016, doi:10.1109/QoMEX.2016.7498936.

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (https://creativecommons.org/licenses/by-sa/4.0/).