How Good Is Aesthetic Ability of a Fashion Model?

Xingxing Zou[1,3], Kaicheng Pang[1,3], Wen Zhang[2], Waikeung Wong[3,1*]
[1]Laboratory for Artificial Intelligence in Design, The Hong Kong Polytechnic University, [2]Amazon.com
[3]Institute of Textiles and Clothing, The Hong Kong Polytechnic University
[1]{aemika.zou,16106013g}@connect.polyu.hk, calvin.wong@polyu.edu.hk, [2]wenzhaw@amazon.com

Abstract

*We introduce **A100** (Aesthetic 100) to assess the aesthetic ability of the fashion compatibility models. To date, it is the first work to address the AI model's aesthetic ability with detailed characterization based on the professional fashion domain knowledge. A100 has several desirable characteristics: 1. **Completeness**. It covers all types of standards in the fashion aesthetic system through two tests, namely **LAT** (Liberalism Aesthetic Test) and **AAT** (Academicism Aesthetic Test); 2. **Reliability**. It is training data agnostic and consistent with major indicators. It provides a fair and objective judgment for model comparison. 3. **Explainability**. Better than all previous indicators, the A100 further identifies essential characteristics of fashion aesthetics, thus showing the model's performance on more fine-grained dimensions, such as Color, Balance, Material, etc. Experimental results prove the advance of the A100 in the aforementioned aspects. All data can be found at https://github.com/AemikaChow/AiDLab-fAshIon-Data.*

1. Introduction

Fashion compatibility learning is a task to measure the compatibility among a set of fashion items [2, 3, 11, 12, 18, 28, 30, 31]. Utilizing the aesthetic ability of these methods for cross-selling is the most common strategy for online retailers. Naturally, a good indicator of a model's aesthetic ability is vital for both method improvement and real-world fashion applications. Current practices to evaluate the fashion compatibility models most focus on retrieving or ranking performance such as Recall [8], mAP [15], MRR [17], etc. AUC [26] is the commonly adopted metric to evaluate the compatibility classification accuracy. FITB (Fill-in-the-black) accuracy [9] is introduced to evaluate fashion recommendation methods. However, none of the existing indicators focus on reflecting the model's aesthetic ability [21,32].

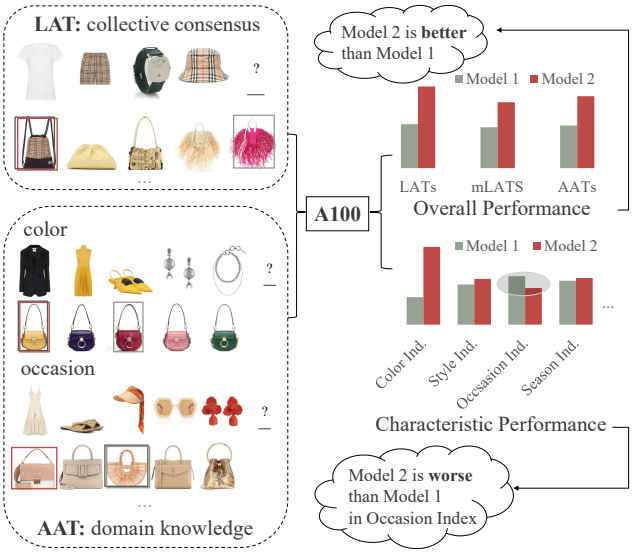


Figure 1. **Overview.** We introduce the A100 (Aesthetic 100) for fashion compatibility model evaluation. The LAT and AAT reflect the overall performance of fashion compatibility models covering all types of aesthetic standards in fashion. Meanwhile, the A100 can also indicate the characteristic performance, which is not always aligned with the overall performance.

The aesthetic ability here refers to how the model can understand fashion items' compatibility and perception of their beauty. Generally, aesthetic system in fashion consists of two types standard: **Bottom-up** and **Top-down** [6,7,10]. The Bottom-up aesthetic means that the fashion from the crowd further affects the mainstream. *This type of standard is a collective consensus that will be formed when the number of people agreeing on one thing is large enough.* We emphasize that it is a large number of people have a consensus on the **same** thing. On the contrary, the Top-down aesthetic means fashion from professional knowledge and will be widely accepted by the crowd from its essence. *This kind of standard faithfully follows the created beauty according to domain knowledge.* The Bottom-up aesthetic is a kind of collective consensus, and thus, the public will accept it by nature, while the Top-down aesthetic can be regarded as

* Corresponding author.

luxury fashion, i.e., is pre-defined and then exerts a substantial influence to lead the public to accept it. Furthermore, we argue that there are three considerations to examine when designing the evaluation protocol: 1. **Completeness**. A relatively objective consensus to serve as the basis of reference is essential to the quantitative assessment, while the systematic standard is key to a comprehensive evaluation. 2. **Practicality**. It refers to choosing a feasible way to perform the evaluation. 3. **Reliability**. The content of the evaluation should be professional and reliable.

To this end, as shown in Figure 1, unlike previous evaluations only focused on overall performance, we propose **A100** (Aesthetic 100), with a more comprehensive evaluation. Specifically, we introduce two tests with multiple-choice questions, namely **LAT** (Liberalism Aesthetic Test) and **AAT** (Academicism Aesthetic Test), to cover Bottom-up and Top-down standards in the fashion aesthetic system. The LAT represents the aesthetic standard of Bottom-up. The source images are collected from mainstream outfit datasets [9, 23, 25]. The questions are automatically generated following the proposed Outfit Generation Principle and then manually verified by experts with a fashion background. We ensure that, for each question, there has already formed a collective consensus in a small group of people, i.e., has only one correct answer. Finally, we build a website and release the LAT to the fashion community for obtaining the ground truth. It is worth noting that the answers of each participant are not the same. Thus, the LAT has two scores: 1. LATs (LAT score). The hard score follows the majority, i.e., the most selected choice of each question will be scoring 1 and others are 0. 2. mLATs (mean LATs). The soft score considers the minority, i.e., the score of each choice equals the probability of it being selected.

In addition, AAT represents the aesthetic standard of Top-down. The creation of the AAT has significant highly professional requirements, and thus we introduce the domain power from the fashion community (all participating designers will be claimed in the acknowledgment). After detailed investigation and discussion, we conclude six dimensions that should be examined when judging the aesthetic ability of the model, including Color, Style, Occasion, Season, Material, and Balance. Then, the questions and choices in the AAT are rigorously designed following these dimensions and their sub-dimensions. Each question is limited to focusing on testing the model's performance from only one dimension. This strategy enables A100 to uniquely show the **characteristic performance** of the model on a fine-grained level in addition to the overall performance. The accuracy is denoted as AATs (AAT score), while the accuracy of each dimension set is called detailed index, e.g., Color Index, Style Index, Occasion Index, etc. We perform analysis across quantitative and qualitative results to demonstrate that our evaluations are more reliable than the previous indicators. Meanwhile, we present the results showing the explainability of the proposed protocol. The main contributions are summarized as the following:

- We do the first work to evaluate fashion compatibility quality based on the professional fashion domain knowledge.
- We introduce A100 covering systematic aesthetic standards, which can provide charateristical performance in addition to overal performance.
- We demonstrate the reliability and explainability of the new evaluations through experiments.

2. Related Work

2.1. Fashion Compatibility Evaluation

There are a total of 14 indicators adopted in previous works. AUC [26] is the most popular indicator which evaluates the item-item recommendation based on the compatibility score. Similar to AUC, mAP [15] and NDCG [13] are indicators of ranked retrieval while Recall [8] reflects items that are not in order. In addition, F1 score [27], MRR [17], and Lift@K [22] reflect the ranking performance of a model. ER [20] is used to predict the "also-bought" relationship in the Amazon dataset [20]. Agreeable [26] is to measure how agreeable the recommendation algorithm's results are across solid and patterned queries. The N-best accuracy [14] represents the rate of recommending the right top/bottom with N recommendations given a test bottom/top set. The Similarity evaluation [14] is the average similarity between the recommended clothing and the held-out paired clothing. FITB accuracy [9] was introduced to evaluate fashion recommendation methods. All in all, we summarize that most of the previous indicators focus on either the recommending or the retrieval performance of the fashion compatibility models. Details can be found in the supplementary material Section 1.

2.2. Fill-In-The-Blank Test

We briefly review the FITB tests used in previous fashion compatibility modeling works. Maryland dataset is the first public outfit dataset which was proposed in 2017 by Han *et al.* [9]. There is a total of 3,076 outfits collected from Polyvore for testing. For each outfit, three wrong FITB choices are selected randomly from all remaining products. Vasileva *et al.* [25] introduced FITB test set with 10,000 questions. The incorrect choices in each question of this FITB task are sampled from the items having the same category as the correct choice. Polyvore-U [19] is an outfit dataset containing user information, which also uses FITB tests for evaluation. iFashion dataset has 1.01 million outfits created by Taobao's fashion experts [1]. For obtaining

the FITB accuracy, they split 10% data as the test set. For each masked item, they randomly select three items from other outfits along with the ground truth item to obtain a multiple-choice set. Previous FITB test sets still have problems in terms of 1. Completeness. The aesthetic standard contained in these FITB tests lacks uniformity. All of those outfit datasets are contributed by different online users. The outfit is "compatible" or not, and the "correct" answer for each question most likely does not reach the consensus. Meanwhile, the aesthetic standards contained in these FITB tests are not systematic enough. 2. Reliability. The quality of these FITB questions is questionable. The way to obtain the choice set is to randomly select several fashion items from the rest data according to the same category [25] or not [9]. The problem is that one cannot ensure the randomly selected items are not correct. The randomly selected items may also be compatible or even more compatible than the ground truth item. Details can be found in the supplementary material Section 2.

3. A100 Evaluation Protocol

In this section, we introduce the details of how to build the A100. Before that, for better demonstration, we provide background knowledge in our task.

3.1. Domain Background

Principles to construct a valid outfit. We summarize the general categories of fashion items and provide the details in Table 1. As a complete outfit, there needs a pair of shoes and clothing items that at least cover the whole body (as shown in Figure 2(a), e.g., top and bottom, one-piece, etc.) The optional items include bags and accessories (Figure 2(b)). Note that Tops, Skirts, Pants, Outwear, Dresses, and Jumpsuits are collectively called clothing. Earrings, Necklaces, Rings, Bracelets, Watches, Hats, Eyewear, Gloves, Legwear, Neckwear, Hair wear, and Brooch are collectively called accessories. Clothing and shoes are affirmatively needed items. There should be only one pair of shoes in an outfit. Skirts, Pants, Dresses, and Jumpsuits are mutually exclusive. Situations like layer (Figure 2(c)) and particular way to do mix and match (Figure 2(d)) are not taken into consideration because it lacks universality. Each outfit can only contain one item from each clothing-sub category. Bags and accessories are not necessary for a complete look. There is only one bag in an outfit (without considering the particular situation like Figure 2(e)). Similarly, only one of each accessories sub-item can exist in each outfit. Meanwhile, since the length of an outfit will not have more than 8 in practice [16], we set the number of fashion items in each question within the range of $[1, 7]$. Unlike the previous FITB tests, we have five items in each choice set instead of four. We find that five choices balance the test complexity and the workload to create these questions.

Table 1. Number of images in the 20 categories in the cleaned Maryland [9] (Cleaned-M), cleaned Type-aware [25] (Cleaned-T), cleaned FashionVC Dataset [23] (Cleaned-F), and newly collected Mytheresa dataset.

Category	Cleaned-M	Cleaned-T	Cleaned-F	Mytheresa
Tops	19,397	26,528	9,537	1,405
Skirts	5,307	8,592	4,102	527
Pants	8,957	12,653	4,703	833
Outwear	10,169	14,172	2,368	961
Dresses	7,480	12,649	2,607	1,922
Jumpsuits	296	820	5	154
Shoes	20,135	38,961	0	687
Bags	21,268	34,882	6	719
Earrings	5,508	12,450	0	123
Necklaces	4,664	7,781	0	352
Rings	3,227	6,265	0	212
Bracelets	5,189	7,522	0	207
Watches	2,290	3,505	0	28
Hats	2,913	5,550	0	196
Eyewear	6,685	8,990	1	156
Gloves	386	723	0	86
Legwear	202	507	0	155
Neckwear	1,189	2,778	0	189
Hair wear	962	1,048	0	52
Brooch	995	280	0	8

3.2. Liberalism Aesthetic Test (LAT)

Based on all the above insights, we build the Liberalism Aesthetic Test as follows.

Step 1: Obtaining source images. Firstly, we clean three widely-adopted outfit datasets including Maryland Dataset [9], Type-aware Dataset [25], and FashionVC Dataset [23]. (Noted that we ensure all images used in A100 have contained a single fashion product with a clean background, which is consistent with all general outfit datasets. It considers the possible domain shift among different types of images.) These three datasets are directly crawled from the Internet without manual filtering. We first delete the decoration images such as lipsticks, newspapers, flowers, etc. Meanwhile, the images with clutter backgrounds, multiple items, folded clothes, and partial clothing visibility are deleted. Then, we re-organize these fashion items into the 20 categories, which are summarized in Table 1. After careful labeling, we obtained an image pool of fashion items associated with the category labels. Additionally, since those datasets were published a few years ago, to catch up with the trend, we newly collected 8,972 fashion items from Mytheresa[1]. Adding up all datasets, we obtained an image pool with 366,176 fashion items. Then, we target generating a large number of valid outfits.

Step 2: Generating seed outfits. To reduce the influence of personal taste resulting in the data, we propose the Outfit

[1]https://www.mytheresa.com/

(a) one piece (or top and bottom) with a pair of shoes

(b) basic fashion outfit show in (a) added optional fashion items

(c) using multiple clothes to create layer in mix and match

(d) dress with pant (e) multiple bags

Figure 2. Examples of complete outfits.

Generation Principle to generate the outfits automatically. Following the definition of a valid outfit described in Section 3.1, we firstly select shoes, then clothing that can cover the whole body, and finally the items from optional categories. Since we do not consider the exceptional cases, e.g., dress with pants or skirt with pants, the Dress and Jumpsuit can be collectively called as One-piece, the Skirt and Pants can be collectively called as Lower-body. The details of the Outfit Generation Principle can be found in the Algorithm 1.

Step 3: Verifying seed outfits. The obtained seed outfits are valid but not compatible enough. We thus manually verify it with the help of experts. We follow with several coarse to fine steps to avoid bringing bias into the data as much as possible. Specifically, we randomly generated 50,000 compositions as the seed outfits and then invited five experts majoring in fashion to delete incompatible outfits (10,000 for each of them). A total of 12,096 outfits remained. Next, they were asked to rate every remaining outfit with an aesthetic score from 1 to 10, and we chose the top 2,000 outfits.

Step 4: Creating multi-choice questions. To ensure the objectivity of the created question, for those 2,000 outfits, we randomly blanked one item from each outfit to obtain the initial FITB questions. The incorrect choices in each question were sampled from the rest of the same category as the masked item. Then, considering the effectiveness of the FITB test, we ensure there is one and only correct answer in each question. Specifically, we released the 2,000 questions to 10 team members (including those five members) and asked them to do the test. The newly joined members reduce the possible bias in the test brought by those previous five members who have seen the questions. We rank

Algorithm 1: Generating valid outfits

Data: Shoes $\mathcal{S}$, Bag $\mathcal{B}$, Accessories $\mathcal{A}$, Clothing $\mathcal{C}$, Tops $\mathcal{C}_t$, Lower-body $\mathcal{C}_{sp}$, Outwear $\mathcal{C}_o$, One-piece $\mathcal{C}_{dj}$
Result: n complete outfits $\mathcal{O}$

$i = 1$;
while $i \leq n$ **do**
  α = random.randint(2, 8) ▷ Length of the outfit;
  $\mathcal{O}_i \leftarrow \mathcal{O}_i \cup \mathcal{S}_{rand(1)}$ ▷ The subscript $rand(1)$ denotes random selecting 1 elements from the set;
  $\alpha = \alpha - 1$;
  switch α **do**
    case 1 **do**
      $\mathcal{O}_i \leftarrow \mathcal{O}_i \cup (\mathcal{C}_{dj})_{rand(1)}$ ▷ One-piece;
    end
    case 2 **do**
      $c = (\mathcal{C} - \mathcal{C}_o)_{rand(1)}$;
      $\mathcal{O}_i \leftarrow \mathcal{O}_i \cup c$ ▷ Clothing excepting Outwear;
      if $c \in \mathcal{C}_t$ **then**
        $\mathcal{O}_i \leftarrow \mathcal{O}_i \cup (\mathcal{C}_{sp} \cup \mathcal{C}_{dj})_{rand(1)}$;
      else if $c \in \mathcal{C}_{sp}$ **then**
        $\mathcal{O}_i \leftarrow (\mathcal{C}_t)_{rand(1)}$;
      else
        $\mathcal{O}_i \leftarrow (\mathcal{C}_t \cup \mathcal{C}_o \cup \mathcal{B} \cup \mathcal{A})_{rand(1)}$;
      end
    end
    otherwise do
      $c = \mathcal{C}_{rand(1)}$ ▷ Select one clothing;
      $\mathcal{O}_i \leftarrow \mathcal{O}_i \cup c$;
      if $c \in \mathcal{C}_t$ **then**
        $\mathcal{O}_i \leftarrow \mathcal{O}_i \cup (\mathcal{C}_{sp} \cup \mathcal{C}_{dj})_{rand(1)}$;
        $\mathcal{O}_i \leftarrow \mathcal{O}_i \cup (\mathcal{C}_o \cup \mathcal{B} \cup \mathcal{A})_{randOP(\alpha-2)}$ ▷ The subscript $randOP(n)$ denotes random selecting 1 elements from each category (excepting the $\mathcal{A}$) in the set;
      end
    else if $\mathcal{C}_{rand(1)} \in \mathcal{C}_{sp}$ **then**
      $\mathcal{O}_i \leftarrow \mathcal{O}_i \cup (\mathcal{C}_t)_{rand(1)}$;
      $\mathcal{O}_i \leftarrow \mathcal{O}_i \cup (\mathcal{C}_o \cup \mathcal{B} \cup \mathcal{A})_{randOP(\alpha-2)}$;
      else
        $\mathcal{O}_i \leftarrow \mathcal{O}_i \cup (\mathcal{C}_t \cup \mathcal{C}_o \cup \mathcal{B} \cup \mathcal{A})_{randOP(\alpha-1)}$;
      end
  end
  $i = i + 1$
end

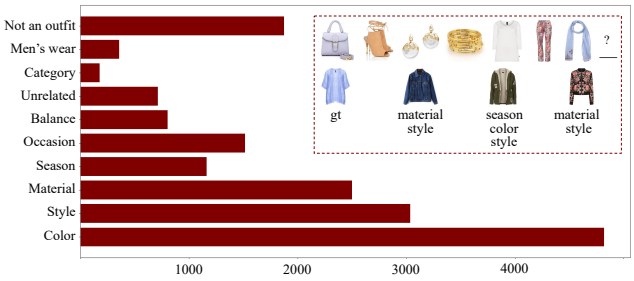


Figure 3. Statistical results of examined dimensions in the Type-aware test set. The number indicates how many times the factor has been examined.

the questions according to the consistency of the obtained answers. The top 500 questions are kept. We manually changed the interfered choices. Next, these ten members were asked to answer those questions every three days, and this practice was repeated three times. Finally, we selected 100 questions with 100% answer consistency as the LAT questionnaires. Note that the questions we selected consider the balance of different categories (Top : Bottom : One-piece : Outwear : Shoes : Bags : Accessories = 1 : 1 : 2 : 1 : 2 : 2 : 1.).

Step 5: Obtaining answers from the crowd. Finally, we built a questionnaire website and released the LAT to the fashion community to obtain the ground truth of this test. Then, we defined the LAT score (LATs) as:

$$LATs = \sum_{n=1}^{100} \delta_{Anw_{model(n)} Max(Anw_{expert(n)})}/100, \quad (1)$$

where the δ_{ij} is the Kronecker delta function, the $Anw_{model(n)}$ is the answer given by a fashion compatibility model of the n-th question. The $Max(Anw_{expert(n)})$ denotes the answer that is most selected by the fashion specialists of the n-th question. Additionally, we proposed the mLATs to reflect the deviation of fashion aesthetics among different people. The mLATs can be calculated as:

$$mLATs = \sum_{n=1}^{100} COUNT(Anw_{model(n)})/x, \quad (2)$$

where the $COUNT(\cdot)$ is how many people select the corresponding answer. x refers to, in total, how many people participate in the test.

3.3. Academicism Aesthetic Test (AAT)

Academicism Aesthetic Test is designed in a totally different manner. To deal with the high professional requirements of developing such a test, we worked with 9 fashion designers to a profound extent to obtain the domain knowledge. This test represents the aesthetic standard of "Top-down". Details to create the AAT are shown as follows.

Step 1: Dimensions of assessment. Detailed investigations and several discussions are made to specify which aspects should be examined when evaluating the model's aesthetic ability. This work is complicated since there are no standard answers in textbooks. [5, 24, 25] indicates several factors, such as color, style, and material, etc., that will affect outfit compatibility. With these hints, we investigate questions in the Type-aware test set [25]. A detailed analysis of each question is made to conclude which factors resulted in the choice being correct or incorrect. For better demonstration, we visualize one question in Figure 3. We can see that two or more factors cause all incorrect answers. The army-green cotton outwear is not compatible with those above fashion items since 1. Season; this outwear is Winter wear, and the rest should be worn in the Spring. 2. Color; the army green is not compatible with the taro-purple bag. 3. Style; this outwear tends to be casual style while the rest in the question as a whole is of a more elegant style. Then, we summarize that this question relates to the factor of material, style, color, and season. The statistical result is shown in Figure 3. We ignore the factors in the long tail position. After several discussions, as shown in Table 2, we organize the remaining factors and put them into a tree structure with a total of six main dimensions, including Color, Style, Occasion, Season, Material, and Balance. More details can be found in the supplementary material Section 3.

Table 2. Defined dimensions to evaluate fashion aesthetic ability.

Dimensions	Sub-Dimensions (Number of sub-dimensions)
Color	Same Color, Warm Tone, Cool Tone ...(4)
Style	Street-wear, Modern, Vintage, Sweet ... (8)
Occasion	Formal, Cocktail, Smart Casual, Casual ...(5)
Season	Spring, Summer, Autumn, Winter (4)
Material	Element, Pattern, Texture (3)
Balance	Silhouette, Simple & Complicated, Proportion (3)

Figure 4. Example questions in the AAT. The candidates in the choice set are only different in the sub-dimension of the question designed to examine.

Step 2: Creating outfits with styling ideas. Next, those nine designers were searching for new styling ideas (the images are collected by designers from varied online websites, *e.g.*, SSENSE, respectively) and created a total of 450 outfits (50 per each). Then, the top 100 outfits are selected together with the voting mechanism.

Step 3: Designing choices set accordingly. According to the sub-dimensions defined in Table 2, the examined sub-

Table 3. Evaluation results of different methods that be evaluated on the Maryland FITB test set [9], Polyvore-630 FITB test set [19], Type-aware FITB test set [25], LAT, and AAT, respectively. Noted that 1. we directly use the released model of CSN; 2. We retrained Bi-LSTMs, FHN, and SCE-Net according to their released code.

Methods	Training data	Maryland FITB acc	Polyvore-630 FITB acc	Type-aware FITB acc	LATs	mLATs	AATs
Bi-LSTMs [9]	Maryland [9]	53.50%	41.68%	37.46%	36%	30.82%	35%
FHN [19]	Polyvore-630 [19]	46.20%	**53.13**%	45.80%	54%	41.62%	40%
SCE-Net [24]	UT-Zappos50k [29]	51.30%	42.92%	51.53%	72%	54.63%	42%
CSN [25]	Type-aware [25]	**54.97**%	47.07%	**57.69**%	**73**%	**56.17**%	**59**%

dimension of each question is planned in advance. For example, Q1 to Q20 is set to evaluate the model's performance from the dimension of Color. Among them, Q1 - Q5 is for Same Color, Q6 - Q10 is for Warm Tone, Q11 - Q15 is for Cool Tone, and Q16 - Q20 is for Contrast Color. This strategy enables to reveal of the characteristic performance of the compatibility model intuitively. The accuracy on the Color set is then defined as the Color Index. Similarly, the accuracy on the Style set is called the Style Index. Noted that the ratio of each dimension is taking both balance and importance into consideration. Additionally, when creating the incorrect answers, we ensure two things: 1. There is one and only correct answer to this question. 2. The incorrect answer is wrong only because of the pre-defined factor. Specifically, as shown in Figure 4, the correct answer for the above question is "the third one" since the color of this camisole is matched with the rest of this outfit composition. We can see that except for the color, which is different, the incorrect answers are totally the same as the correct answer in any dimension. Such a way also ensures that our evaluation can clearly indicate the shortcomings of the model. The accuracy of AAT is recorded as the AAT score (AATs).

4. Analysis

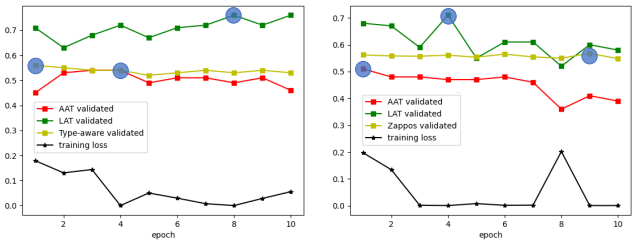
In this section, we demonstrate the characteristics of A100 via quantitative and qualitative results. Firstly, we elaborate the **Reliability** of A100 from two perspectives: Is the evaluation accurate? Furthermore, is this indicator objective? Secondly, we present detailed examples to show the **Explainability** of the A100.

4.1. Analysis of Reliability

In this subsection, we aim to elaborate on the Reliability of A100. We firstly examine the accuracy of A100 on evaluating the performance of models. Specifically, we compare the performance of four mainstream fashion compatibility approaches including Bi-LSTMs [9], FHN [19], SCE-Net [24], and CSN [25]. The quantitative results on three widely-used FITB test sets, i.e., Maryland FITB test set, Polyvore-630 FITB test set, Type-aware FITB test set, are reported. It is worth noting that the voting mechanism will be adopted for judgment when the performance of models reflected by these three test sets has conflicts. The results of this perceptual experiment are not quite a benchmark. We emphasize that all models use the default settings and parameters according to the original papers for two reasons: 1. fairness (regarding them as off-the-shelf models); 2. their input data requirements and training conditions are not the same. From Table 3, it can draw a conclusion that the order of those four methods (from high to low) will be CSN, SEC-Net, FHN, and Bi-LSTMs. A100 reflects the consistent evaluation results with this conclusion which indicates it can assess the performance of different compatibility models accurately.

Additionally, as indicated in the second column of Table 3, Bi-LSTMs, FHN, SCE-Net, and CSN are trained on different datasets. Therefore they suffer risks to overfit their training data. For example, Bi-LSTMs achieve the competitive performance among these methods on the Maryland test set, i.e., only lower than CSN. Testing on these dataset would be a severe problem of model generalization and transfer learning knowledge, *i.e*. Bi-LSTMs, FHN, and CSN perform better than other baselines when evaluating their testing domain. It leads to a biased or even false-positive judgment of performance comparison. We thus repeated the experiments on POG [1], a non-Polyvore dataset, and made the same conclusion (Bi-LSTMs: 47.21% $<$ FHN: 54.13% $<$ SCE-Net: 57.27% $<$ CSN: 66.65%). including NGNN, and CAS-Net, which also demonstrate the reliability of A100. We also test the model trained on the same dataset, *i.e*., NGNN [4] trained on Maryland such as Bi-LSTMs. The evaluation results of NGNN and Bi-LSTMs are Maryland FITB acc: 50.68% $<$ 53.50%; LATs: 33% $<$ 36%, mLATs: 29.88% $<$ 30.82%, AATs: 30% $<$ 35%. It can see that A100 is effective as an a standalone protocol.

To further verify the objectivity of A100, we conduct a comparative experiment. Specifically, we retrain the CSN model with suggested experimental settings on the same training data and observe the optimal training steps from 3 validation losses: 1. loss of original validation data, 2. LAT, and 3. AAT of A100 data. Thus, we have three checkpoints of the model (as marked in Figure 5) to compare in multiple widely-used testing datasets. Better choices should be superior in most testing tasks and suggest greater generalization. We present the training and validation loss per epoch in Figure 5 (a). The blue circles on the green, yellow, red curves



(a) training on Type-aware dataset (b) training on UT-Zappos50k dataset

Figure 5. Experiments for early stopping in CSN [25] and SCE [24] retrain, based on original validation set (yellow), LAT (green), AAT (red) to indicate the early stopping using three validation tests. The blue circles indicate the optimal stop points at which epoch.

in Figure 5 (a) indicates that the early stopping points of using these three validation tests are epoch 1, epoch 4, and epoch 8, respectively. (Noted that the strategy to select the stop points are the same and be consistent in original paper.) Models saved at these checkpoints are Type-aware validated models, LAT validated models, and AAT validated models. We evaluate these three models and report the quantitative results in Table 4. This table shows that, although the Type-aware validated model achieves the highest performance on Type-aware FITB accuracy, this model is consistently weak on all rest indicators. It shows that A100 can help to find a model with more generalized performance. We conducted the same experiment using SCE-Net, i.e., trained on UT-Zappos50k dataset and validated on UT-Zappos50k, LAT, and AAT. The results show the same patterns. As shown in Table 4, the results of two models validated by A100 show consistently higher performance on all indicators.

4.2. Analysis of Explainability

To demonstrate the Explainability of the A100, we further report the results on detailed indexes of Bi-LSTMs, FHN, SCE-Net, and CSN in Table 5. As introduced in Section 3.3, the design methodology of the AAT enables the A100 to reveal the aesthetic ability performance of a model on fine-grained aspects. Specifically, the 100 questions were divided into six groups which are Color, Style, Occasion, Season, Material, and Balance. There are 20, 32, 15, 12, 12, 9 questions in each group, respectively. This considers the balance of sub-dimensions. For the detailed sub-dimensions of each aspect, please check the supplementary material Section 3. From Table 5, we can find some interesting insights: 1. The main factor that the CSN achieves higher results is its outstanding performance on the dimension of Color, i.e., CSN has a score of **0.85**, while Bi-LSTMs and FHN only have 0.30 and 0.50 in this fine-grained aspect. 2. Even the overall performance is lower than the CSN, the results show that FHN has a better understanding of Occasion, i.e., its Occasion Index **0.60** is higher than that of the CSN **0.53** and the SCE-Net **0.33**. This is mainly because that the user's information is more related to the dimension of Occasion causes resulting in the model achieving better performance on this aspect. 3. Similarly, FHN has the lowest score (only has **0.22**) in Style Index among those three. It is possible since the involved varied users' information expands the influence of different personal tastes, which confuses to a certain extent of the model on understanding Style. In addition, we can see that, although the constrasts between FHN and SCE-Net are much narrowed on AATs, FHN enjoys significant advantages on Occasions and Season Indexes while SCE-Net performs well on Color Indexes and Style Indexes. The characteristic performance of A100 can provide a more comprehensive perspective for model's evaluation. 4. Bi-LSTMs achieves **0.11** in Balance Index. The questions in the Balance group are mainly focused on examining the Silhouette, Simple & Complicated, and Proportion. This result indicates that Bi-LSTMs are less sensible to the shape of fashion items.

In addition to the quantitative results, we present the qualitative results in Figure 6 for further demonstration. Color is adopted as an example for easy understanding. More detailed cases will be presented in the supplementary material Section 4. The Color dimension is divided into four sub-dimensions: Same Color, Warm Tone, Cool Tone, and Contrast Color. When we check the detailed results of those three models on the group of Color, we find that: 1. Bi-LSTMs shows a weak ability on color matching. As shown in Figure 6 (a), the green boots selected by Bi-LSTMs is unreasonable conditioned on the colors of these items included in the question. 2. On the contrary, CSN has a consistently good performance on the group of Color. All the wrong answers which were chosen among those 20 questions belong to the group of Contrast Color. Similarly, SCE-Net obtains 0.75 in Color Index, and the four wrong questions also belong to the Contrast Color. We show an example in Figure 6 (b). The army green boots is most suitable among the choices set since it creates an interesting color composition with the Bordeaux red in dress, bags, and earrings while being echo to the color of the sunglasses at the same time. 3. FHN has relatively mediocre performance on this dimension with scoring 0.5 in the Color Index. Specifically, it obtains 4 points on Warm Tone questions while 2 points on Same Color, Cool Tone, and Contrast Color, respectively.

The color performance of those methods on the LAT proves the obtained insights above. As shown in Figure 6 (c), we can see that Bi-LSTMs consistently show the bold taste on color matching while CSN still has good performance on the dimension of Color. In particular, when we observe how many people select each choice, it further proves the insights reflected by the AAT. In terms of the first question, Bi-LSTMs selects the answer with 4% participants agreeing with it while the agreeable ratio of the choice that CSN picked is over 75%. For the second question, Bi-

Table 4. The early stopping of CSN [25], and models retrained on UT-Zappos50k dataset are followed the method of SCE-Net [24].

training dataset	Models	Maryland FITB acc [9]	Polyvore-630 FITB acc [19]	Type-aware FITB acc [25]	LATs	mLATs	AATs
Type-aware [25]	Type-aware validated	56.17%	42.12%	**55.58%**	71%	54.81%	57%
	LAT validated	**57.41%**	**46.42%**	52.05%	**76%**	**57.65%**	58%
	AAT validated	56.93%	44.31%	53.69%	69%	52.76%	**64%**
UT-Zappos50k [24]	Zappos validated	51.21%	41.44%	49.57%	71%	54.17%	42%
	LAT validated	52.11%	42.64%	51.63%	75%	59.42%	49%
	AAT validated	**53.84%**	**43.87%**	**55.46%**	**79%**	**63.19%**	**52%**

Table 5. Results of Bi-LSTMs [9], FHN [19], SCE-Net [24], and CSN [25] evaluated on the AAT. Indexes refer to the performance on the specific dimensions in aesthetic ability evaluation. Q1 - Q20 is the group of Color, Q21 - Q52 is the group of Style, Q53 - Q67 is the group of Occasion, Q68 - Q79 is the group of the Season, Q80 - Q91 is the group of Material, and Q92 - Q100 is the group of Balance. The number is calculated as the correct questions divided by the total number of questions in this group. e.g. for **Color Index** computation, it will be the number of correct questions divided by **20**.

Indexes	**Bi-LSTMs [9]**	**FHN [19]**	**SCE-Net [24]**	**CSN [25]**
Color	0.30	0.50	0.75	**0.85**
Style	0.34	0.22	0.28	**0.50**
Occasion	0.40	**0.60**	0.33	0.53
Season	0.50	0.42	0.33	**0.58**
Material	0.42	0.50	0.50	0.50
Balance	0.11	0.33	0.33	**0.56**
AATs	0.35	0.40	0.42	**0.59**

Figure 6. Examples of results reflect the models' performance of Color. (a) An example of examining the Same Color in AAT. (b) An example of examining Contrast Color in AAT. (c) Examples in LAT. Noted that the number below each choice means the ratio of how many people select it.

LSTMs and FHN select the answer with a less agreeable percentage, i.e., 4% and 10%, respectively. On the contrary, CSN selects the same choice with over 72% participants. Similarly, apart from the Color, we can obtain the characteristic performance of the fashion compatibility model on other dimensions as well. Then, the model can be improved, specifically focusing on the insufficient aspects. For example, as indicated by A100, existing models have relatively weak performance on the dimension of Balance. A simple idea for model enhancement is to sample more data related to the Balance.

5. Limitations

Number of questions. The main challenge of scalability is that for LAT, the concentration of human beings is limited; for AAT, nine well-established designers need to spend lots of time working together under the voting mechanism to reduce personal preference. We will keep updating it under current principles.

Scope of the evaluations. The new evaluations only focus on the compatibility among a set of fashion items. In other words, the factors related to personal information such as body figure, skin color, or users' preference are not in the scope of this work. Meanwhile, as shown in Figure 2 (c) (d) (e), there are many ways to do mix and match, which causes styling to be an amusing and creative thing. To reduce the complexity, we did not consider particular ways. Additionally, the aesthetic criteria adopted in these evaluations follow the most common cognitive, which only assesses the basic level of aesthetic ability. The creativity and artistic sense of human is unmeasurable.

6. Conclusion

We introduce A100 for fashion compatibility model evaluation. It provides that fine-grained indexes can further reveal the insufficient aspects of the models. Incorporating these evaluations in the performance analysis can provide better insights for model improvement. The extensive analysis demonstrates the effectiveness of A100. We hope that the new aesthetic perception indicators can benefit the designation of the modern fashion intelligence system and inspire practical applications towards real fashion AI.

Acknowledgement: This work is supported by Laboratory for Artificial Intelligence in Design (Project Code: RP 3-1) under InnoHK Research Clusters, Hong Kong SAR Government. We thank Professor Zowie Broach, Xintong Han and all participants from the fashion community for their valuable support and constructive comments.

References

[1] Chen et al. Pog: personalized outfit generation for fashion recommendation at alibaba ifashion. In *SIGKDD*, 2019. 2, 6

[2] Xu Chen, Yongfeng Zhang, Hongteng Xu, Yixin Cao, Zheng Qin, and Hongyuan Zha. Visually explainable recommendation. *preprint arXiv:1801.10288*, 2018. 1

[3] Guillem Cucurull, Perouz Taslakian, and David Vazquez. Context-aware visual compatibility prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[4] Cui et al. Dressing as a whole: Outfit compatibility learning based on node-wise graph neural networks. *arXiv*, 2019. 6

[5] Molly Eckman and Janet Wagner. Aesthetic aspects of the consumption of fashion design: The conceptual and empirical challenge. *ACR North American Advances*, 1995. 5

[6] Molly Jean Eckman. *Consumers' aesthetic evaluation of clothing: The effect of age, sex, and fashion involvement*. PhD thesis, University of Maryland, College Park, 1992. 1

[7] Joanne Entwistle. *The aesthetic economy of fashion: Markets and value in clothing and modelling*. Berg, 2009. 1

[8] Sida Gu, Xiaoqiang Liu, Lizhi Cai, and Jie Shen. Fashion coordinates recommendation based on user behavior and visual clothing style. In *Proceedings of the 3rd International Conference on Communication and Information Processing*, pages 185–189, 2017. 1, 2

[9] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. Learning fashion compatibility with bidirectional lstms. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1078–1086. ACM, 2017. 1, 2, 3, 6, 8

[10] Ruth Estella Hawthorne. *Aspects of design preference in clothing: aesthetic, motivation, and knowledge*. The Ohio State University, 1967. 1

[11] Min Hou, Le Wu, Enhong Chen, Zhi Li, Vincent W Zheng, and Qi Liu. Explainable fashion recommendation: A semantic attribute region guided approach. *Twenty-Eight International Joint Conference on Artificial Intelligence*, 2019. 1

[12] Wei-Lin Hsiao and Kristen Grauman. Creating capsule wardrobes from fashion images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7161–7170, 2018. 1

[13] Yang Hu, Xi Yi, and Larry S Davis. Collaborative fashion recommendation: A functional tensor factorization approach. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 129–138. ACM, 2015. 2

[14] Tomoharu Iwata, Shinji Watanabe, and Hiroshi Sawada. Fashion coordinates recommender system using photographs from fashion magazines. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011. 2

[15] Yuncheng Li, Liangliang Cao, Jiang Zhu, and Jiebo Luo. Mining fashion outfit composition using an end-to-end deep learning approach on set data. *IEEE Transactions on Multimedia*, 19(8):1946–1955, 2017. 1, 2

[16] Zhi Li, Bo Wu, Qi Liu, Likang Wu, Hongke Zhao, and Tao Mei. Learning the compositional visual coherence for complementary recommendations. *arXiv preprint arXiv:2006.04380*, 2020. 3

[17] Yujie Lin, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jun Ma, and Maarten De Rijke. Explainable outfit recommendation with joint outfit matching and comment generation. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1502–1516, 2019. 1, 2

[18] Zhi Lu, Yang Hu, Yan Chen, and Bing Zeng. Personalized outfit recommendation with learnable anchors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12722–12731, 2021. 1

[19] Zhi Lu, Yang Hu, Yunchao Jiang, Yan Chen, and Bing Zeng. Learning binary code for personalized fashion recommendation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10562–10570, 2019. 2, 6, 8

[20] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52. ACM, 2015. 2

[21] Seyed Omid Mohammadi and Ahmad Kalhor. Smart fashion: A review of ai applications in the fashion & apparel industry. *arXiv preprint arXiv:2111.00905*, 2021. 1

[22] Luisa F Polanía and Satyajit Gupte. Learning fashion compatibility across apparel categories for outfit recommendation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4489–4493. IEEE, 2019. 2

[23] Xuemeng Song, Fuli Feng, Jinhuan Liu, Zekun Li, Liqiang Nie, and Jun Ma. Neurostylist: Neural compatibility modeling for clothing matching. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 753–761. ACM, 2017. 2, 3

[24] Reuben Tan, Mariya I Vasileva, Kate Saenko, and Bryan A Plummer. Learning similarity conditions without explicit supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10373–10382, 2019. 5, 6, 7, 8

[25] Mariya I Vasileva, Bryan A Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. Learning type-aware embeddings for fashion compatibility. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 390–405, 2018. 2, 3, 5, 6, 7, 8

[26] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4642–4650, 2015. 1, 2

[27] Agung Toto Wibowo, Advaith Siddharthan, Judith Masthoff, and Chenghua Lin. Incorporating constraints into matrix factorization for clothes package recommendation. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, pages 111–119, 2018. 2

[28] Xun Yang, Yunshan Ma, Lizi Liao, Meng Wang, and Tat-Seng Chua. Transnfcm: Translation-based neural fashion compatibility modeling. *arXiv preprint arXiv:1812.10021*, 2018. 1

[29] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE Con-*

ference on Computer Vision and Pattern Recognition, pages 192–199, 2014. 6

[30] Huijing Zhan, Jie Lin, Kenan Emir Ak, Boxin Shi, Ling-Yu Duan, and Alex C Kot. A3-fkg: Attentive attribute-aware fashion knowledge graph for outfit preference prediction. *IEEE Transactions on Multimedia*, 2021. 1

[31] Xingxing Zou, Zhizhong Li, Ke Bai, Dahua Lin, and Waikeung Wong. Regularizing reasons for outfit evaluation with gradient penalty. *arXiv preprint arXiv:2002.00460*, 2020. 1

[32] Xingxing Zou and Waikeung Wong. fashion after fashion: A report of ai in fashion. *arXiv preprint arXiv:2105.03050*, 2021. 1