Supplemental Material: 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

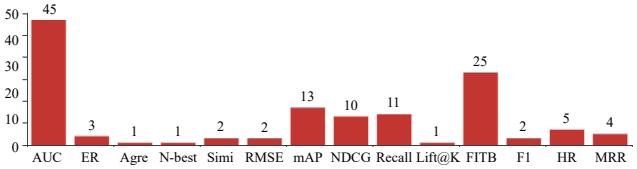

Figure 1. Evaluation indicators had been used in the previous fashion compatibility learning papers. Noted that if one paper uses both AUC [20] and FITB accuracy [5], they will be counted respectively. For clear demonstration, we use shortening of the words in the Figure. Agre refers to Agreeable [8]; N-best refers to N-best Accuracy [7]; Simi refers to Similarity [7].

1. Existing Quantitative Indicators

This section is mainly to serve as the supplementary of Section 2.1 in the main paper, which aims to introduce the current situation in objective evaluations of fashion compatibility models.

We searched for papers focused on fashion compatibility learning. The earliest paper to be reviewed was published on 28 January 2007 [16]; the most recent paper was published on 9 June 2021 [21]. The statistical results of the evaluation indicators that had been used in these papers are shown in Figure 1. 78 out of 103 papers presented the quantitative results. There are total of 14 kinds of evaluation indicators were adopted in previous research. There are AUC [20], Error Rate (ER) [13], Agreeable [8], N-best accuracy [7], Similarity [7], Root Mean Squared Error (RMSE) [3], mAP [9], Normalized Discounted Cumulative Gain (NDCG) [6], Recall [4], FITB Accuracy [5], F1 score [22], Hit Ratio (HR) [2], Mean Reciprocal Rank (MRR) [10], and Lift@K [15]. AUC is the most popular indicators among those 14 which is widely used to evaluate the item-item recommendation based on the compatibility score. Specifically, for each testing positive pair $(h_i, t_{ig}) \in \mathcal{P}_t$, the tail item is replaced with N negative items $\{t_{in}\}_{n=1}^{N}$ which do not co-occur with h_i in the same outfit but are from complementary categories with h_i. Then, the AUC can be calculated as:

$$\text{AUC} = \frac{1}{N\,|\mathcal{P}_t|} \sum_i \sum_n \delta(s(h_i, t_{ig}) > s(h_i, t_{in})) \quad (1)$$

where $\delta(a)$ is an indicator function that returns 1 if the argument a is true, otherwise 0. $s(h_i, t_{ig}) = P((h_i, t_{ig}) \in \mathcal{P}_t)$ denotes the compatibility score. $|\mathcal{P}_t|$ denotes the total number of testing positive pairs. Over 54% fashion compatibility learning related papers adopted AUC to show the overall performance of their approaches. However, the standard of positive and negative outfits is defined by the training set, which means AUC reflects how similar a model's aesthetic taste is to which of the training data. Like AUC, other indicators are widely used in recommending tasks, such as MAP, NDCG, Recall, etc. The former two are evaluation indicators of ranked retrieval, while the latter reflects items not in order. In addition, F1 score, MRR, Lift@K also belongs to the type of indicators that reflect the ranking performance of the model. For brevity, we introduce the indictors which are not so well-known. Lift@K is defined as:

$$\text{Lift@K} = \frac{AP@K\,(model)}{AP@K\,(random)} \quad (2)$$

ER [13] which was adopted to predict "also-bought" relationship on Amazon dataset [13]. It involves extra information, i.e. "also-bought" relationship. Agreeable is to measure how agreeable the recommendation algorithm's results are across solid and patterned queries. The N-best accuracy represents the rate of recommending the correct top (bottom) with N recommendations given a test bottom (top) set. The Similarity evaluation measurement is the average similarity between the recommended clothing and the held-out paired clothing. These two indicators are designed for an outfit that only has a top and bottom. These four indicators were adopted in early papers i.e. 2011 [7], 2014 [8], 2015 [13], which are seldom used in later research due to the increasingly complicated tasks. FITB: given an outfit with one missing item, recommend an item that matches

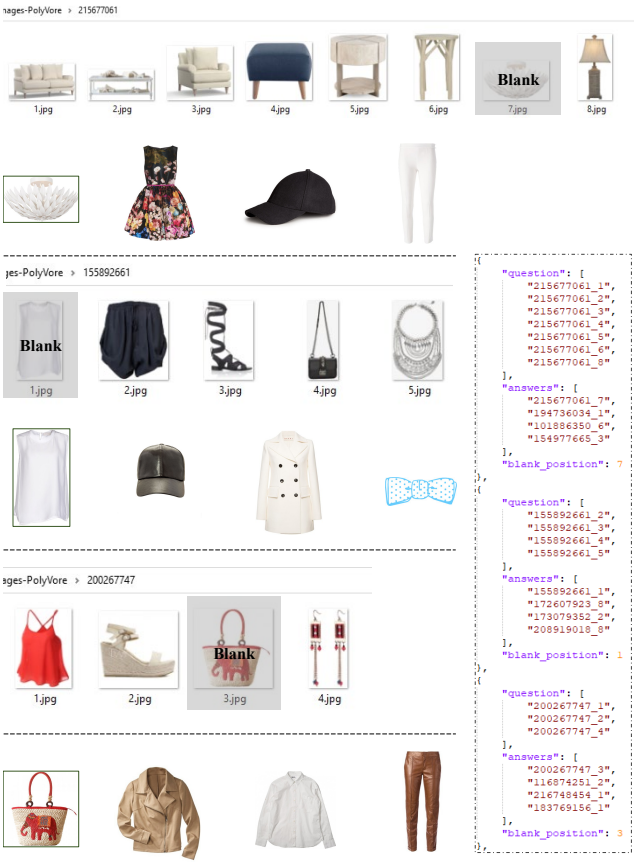

Figure 2. Example questions in the Maryland FITB test [5].The green box indicates the correct answer.

well with the existing set [5]. FITB has been designed to evaluate the performance of tasks that needs to bridge visual and semantic information, such as evaluating video understanding model via FITB question answering [12], FITB Description Generation, and Question Answering [23]. All in all, we found that all of the existing quantitative indicators are not effective enough to reveal the aesthetic ability of the fashion compatibility model. None existing research focused on objective evaluations in fashion compatibility task [14,24]. Indeed, it is hard to define objective metrics to reflect many notions in fashion like compatibility, novelty, beauty, etc. However, most research still adopts subjective assessments such as user study, which can be inaccurate and biased [14].

2. Existing FITB tests

This section is mainly to serve as the supplementary of Section 2.2 in the main paper, which aims to demonstrate further the limitations of current FITB tests for aesthetic ability evaluation with visualized examples.

The FITB test was first introduced to evaluate the performance of the fashion compatibility model by Han *et al.* in 2017. Then, it quickly becomes a mainstream evaluation way in this task. We randomly visualize example questions in the Maryland FITB test in Figure 2. It can be seen that: 1. This FITB test is not clean enough with a certain number of unrelated images. For example, the items in the first question belong to the furniture instead of the fashion item. 2. The incorrect answers are quickly excluded from the choice set according to the principle of a complete outfit. Specifically, as shown in the second case of Figure 2, the "black hat" can not form a complete outfit together with the rest items in the question. 3. The original outfit to create that question is not valid. As shown in the third case in Figure 2, we can see that the bottom part is missing. If we do not consider factors related to any aesthetic, the correct choice among those four candidates should be the "brown pants" instead of the bag.

Towards those problems above, Vasileva *et al.* [19] introduced the Type-aware dataset with fine-grained item type. Compared with 3,076 questions in the Maryland FITB test, there is a total of 10,000 questions in the Type-aware FITB test. In addition, unlike the previous way to create the choice set, the incorrect choices in each question of the Type-aware test are sampled from items with the same category as the correct choice. We visualize some examples in the Type-aware FITB test in Figure 3. The examined aspect for each question is mixed. Here we take the first question in Figure 3 as an example. After detailed analysis, we thus make the following observations: 1. We could exclude the second choice for its **color** is not compatible with the questions. Meanwhile, the **print** of both the first and the third one is not compatible with the questions as well. 2. The randomly formed choice set is questionable. As shown in the second case and third case in Figure 3, it may have another option. For example, the "black bag" is also compatible, while the silhouette of this bucket bag has a more similar style with the question than the pink flap bag. 3. It also has unrelated images as well as invalid questions. The original Type-aware dataset has a total of 68,306 outfits and 365,054 items. After cleaning, the number of remaining items is 206,656. In addition, we investigate the Type-aware FITB test with 10,000 outfits. The number of invalid outfits is 1,875.

Apart from the most mainstream FITB tests introduced above, there have some FITB tests in other papers. Similar to those two tests, they are also split from the newly introduced outfits dataset. For example, iFashion [1] is an outfit dataset collected from Taobao.com. The iFashion FITB test is followed some strategy along with the Maryland test. Specifically, it split 10% data as the test set. Then, for each masked item, they randomly select three items from other outfits along with the ground truth item to obtain a multiple-choice set. FashionVC [17] test only has top and bottom

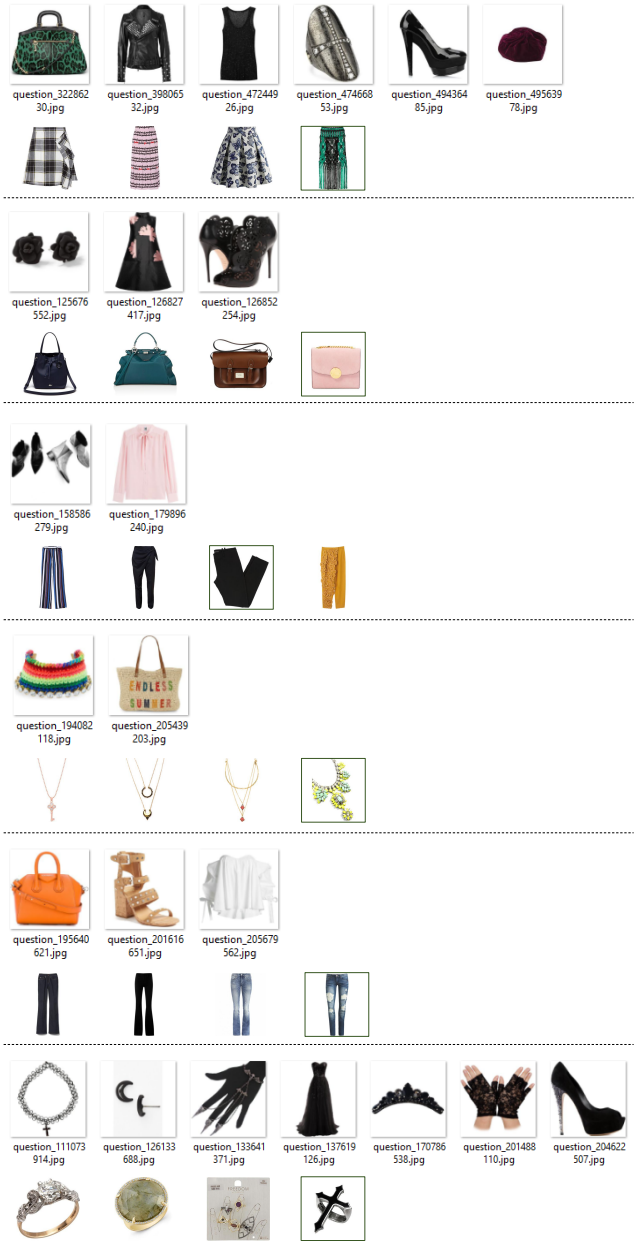

Figure 3. Example questions in the Type-aware FITB test [19].The green box indicates the correct answer.

images while Polyvore-U [11] dataset only has top, bottom, and shoe images. After careful investigation, we concluded that: 1. The aesthetic standard in all of the existing FITB tests is highly-mixed and has less collective consensus; This is because different online users create all of the outfits used to create the questions. 2. The way to create the choice set is questionable; There is a big chance that the masked item is not most compatible with the rest of the choice sets. 3. None of them systematically reflect the fashion aesthetic standard.

Table 1. The details of defined dimensions to assess the aesthetic ability of fashion compatibility model.

Dimension	Sub-dimension	Question Number
Color	Same Color	1 - 5
	Warm Tone	6 - 10
	Cool Tone	11 - 15
	Contrast Color	16 - 20
Style	Street Wear	21 - 24
	Modern	25 - 28
	Vintage	29 - 32
	Sweet	33 - 36
	Sporty	37 - 40
	Classic	41 - 44
	Gender Neutral	45 - 48
	Mash-up	49 - 52
Occasion	Formal	53 - 55
	Cocktail	56 - 58
	Smart Casual	59 - 61
	Casual	62 - 64
	Holiday	65 - 67
Season	Spring	68 - 70
	Summer	71 - 73
	Autumn	74 - 76
	Winter	77 - 79
Material	Element	81 - 83
	Pattern	84 - 87
	Texture	88 - 91
Balance	Silhouette	92 - 94
	Simple & Complicated	95 - 97
	Proportion	98 - 100

3. Details of the AAT

This section is mainly to serve as the supplementary of Section 3.3 in the main paper, which aims to provide the defined dimensions and their sub-dimensions to examine the model's aesthetic ability.

As shown in Table 1, it can be found that the factors that will affect outfit compatibility, e.g., Color, Material, Silhouette, etc., are summarized into a two-layer tree structure. The six dimensions for evaluating the aesthetic ability of fashion compatibility models are Color, Style, Occasion, Season, Material, Balance. Each dimension has the sub-dimensions to evaluate the model in more fine-grained aspects further. Specifically, Color can be further divided into the Same Color, Warm Tone, Cool Tone, Contrast Color. Style can be further divided into Streetwear, Moder, Vintage, Sweet, Sporty, Classic, Gender Neutral, Mash-up. Occasion including Formal, Cocktail, Smart Casual, Holiday. Season includes Spring, Summer, Autumn, Winter. Material including Element, Pattern, Texture. Balance including silhouette, Simple & Complicated, Proportion.

In addition, we emphasize that all the 100 questions in AAT belong to only one of the sub-dimension. The Ques-

Figure 4. Example questions in the AAT.

tion No. is pre-defined as well. This strategy enables A100 to identify the detailed information. We provide more visualized questions of the **AAT** in Figure 4.

Table 2. Results of Bi-LSTMs [5], FHN [11], SCE-Net [18], and CSN [19] evaluated on the AAT. The numbers in this Table refer to how many correct questions among the dimensions of Season.

	Bi-LSTMs [5]	**FHN [11]**	**SCE-Net [18]**	**CSN [19]**
Spring	2	1	1	**3**
Summer	1	**2**	1	1
Autumn	**2**	1	1	1
Winter	1	1	1	**2**
Total	6	5	4	**7**

4. Qualitative Results of Different Dimensions

This section is mainly to serve as the supplementary of Section 4.2 in the main paper, which aims to provide more analysis to demonstrate the Explainability of A100.

As shown in Table 2, we present the detailed results of these four models, including Bi-LSTMs [5], FHN [11], SCE-Net [18], and CSN [19], evaluated on the dimension of Season. There are four sub-dimensions defined in the group of Season, which are Spring, Summer, Autumn, Winter. It can be seen that CSN achieves the best performance among those four while SCE-Net seems not to perform so well in the dimension of Season. Specifically, Bi-LSTMs has good performance on the group of Autumn. FHN obtains highest scores on the group of Summer while CSN shows relatively strong ability on the group of Spring and Winter. We further present the visualized examples in Figure 5.

Additionally, we present the detailed results of these four models evaluated on the dimension of Balance. Specifically, three sub-dimensions are included in the group Balance: Silhouette, Simple & Complicated, and Proportion. Bi-LSTMs obtain 0 points on both groups of Silhouette and Proportion, while achieving 1 point on the Simple & Complicated group. FHN also achieves 0 points on the group of Silhouette, while SCE-Net gets 0 points on the group of Proportion. The Balance dimension is the most challenging part among all dimensions in AAT. It is more related to the harmony of fashion items in shape, complexity, and overall ratio. For example, the silhouette means the shape of a fashion item. The general silhouettes include A-line, H-line, Y-line, etc. Indeed, many fine-grained cases exist. The minor difference will significantly affect the visual perception, which increases the difficulty of this sub-dimension. Additionally, Simple & Complicated refers to whether the design elements are together reaching the visual balance. Involving a large number of design elements in an outfit will not always be terrible. It conversely will be attractive and bring fresh-new feelings if all the elements reach the fantastic visual balance. This also profoundly relies on the aesthetic ability of a designer. Proportion as well. We also put some

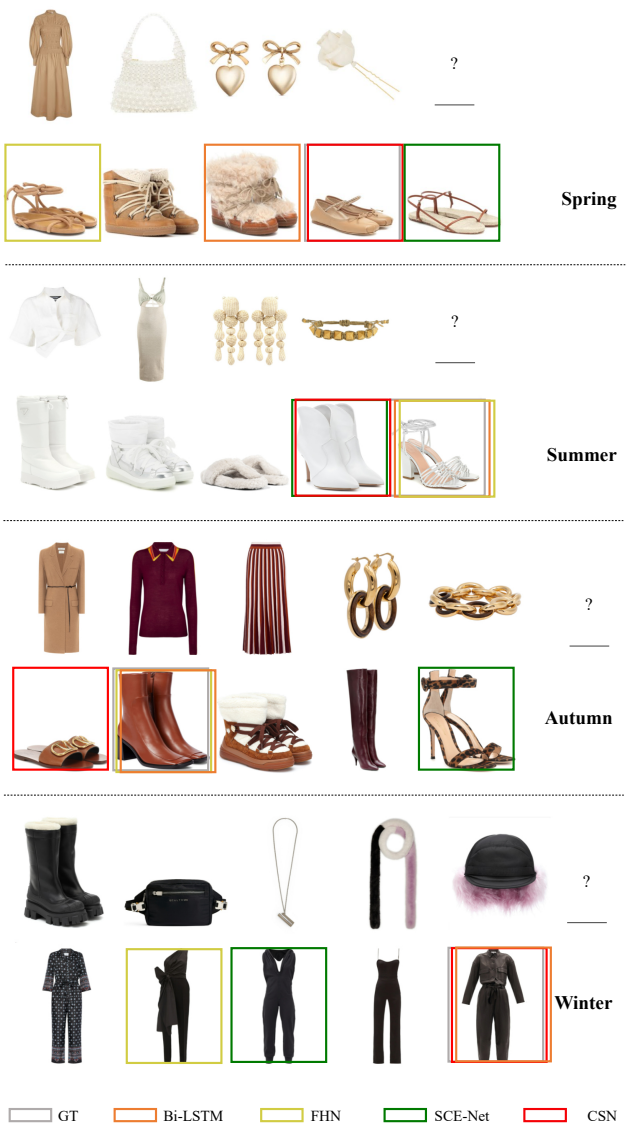

Figure 5. Examples of results reflect the models' performance of Season in AAT.

visualized examples as shown in Figure 6 for better demonstration. It can be seen that, for the first question related to Silhouette, excepting the third black dress, all rest four dresses are not compatible with the blazer in the questions for the problem of Silhouette. Specifically, all the sleeve types of these black dresses can not be well fitted with the outwear. The second question is designed to examine the sub-dimension of Simple & Complicated. As shown in Figure 6, the second Gucci tote bag is most compatible with the items in the question, for its design elements are well echo with the whole outfit. The third question is mainly about Proportion. The second boot is most compatible with the length of the skirt in the question than rest four candidates.

Figure 6. Examples of results reflect the models' performance of Balance in AAT.

References

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

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

[3] Ernani Viriato De Melo, Emilia Alves Nogueira, and Denise Guliato. Content-based filtering enhanced by human visual attention applied to clothing recommendation. In *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 644–651. IEEE, 2015. 1

[4] 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

[5] 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, 4

[6] 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. 1

[7] 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. 1

[8] Vignesh Jagadeesh, Robinson Piramuthu, Anurag Bhardwaj, Wei Di, and Neel Sundaresan. Large scale visual recommendations from street fashion images. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1925–1934, 2014. 1

[9] 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

[10] 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

[11] 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. 3, 4

[12] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6884–6893, 2017. 2

[13] 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. 1

[14] Seyed Omid Mohammadi and Ahmad Kalhor. Smart fashion: A review of ai applications in the fashion & apparel industry. *arXiv preprint arXiv:2111.00905*, 2021. 2

[15] 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. 1

[16] Edward Shen, Henry Lieberman, and Francis Lam. What am i gonna wear? scenario-oriented recommendation. In *Proceedings of the 12th international conference on Intelligent user interfaces*, pages 365–368, 2007. 1

[17] 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

[18] 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. 4

[19] 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, 4

[20] 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

[21] Jianfeng Wang, Xiaochun Cheng, Ruomei Wang, and Shaohui Liu. Learning outfit compatibility with graph attention network and visual-semantic embedding. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 1

[22] 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. 1

[23] Licheng Yu, Eunbyung Park, Alexander C Berg, and Tamara L Berg. Visual madlibs: Fill in the blank description generation and question answering. In *Proceedings of the ieee international conference on computer vision*, pages 2461–2469, 2015. 2

[24] Xingxing Zou and Waikeung Wong. fashion after fashion: A report of ai in fashion. *arXiv preprint arXiv:2105.03050*, 2021. 2