Can Stable Diffusion Recommend Outfits?: Outfit Recommendation from Fashion Item Images via Generative AI

Yuma Oe[1] and Yoshiyuki Shoji[1][0000−0002−7405−9270]

Shizuoka University, Hamamatsu, Shizuoka 432–8011, Japan
oe.yuma.21@shizuoka.ac.jp, shojiy@inf.shizuoka.ac.jp

Abstract. This paper proposes a method for generating and recommending fashionable outfit images based on a given image of a fashion item. The system uses image generation AI, specifically Stable Diffusion, to produce images of a person wearing the input item, leveraging inpainting techniques to complete the surrounding area. Two models were prepared: a fashionable model fine-tuned on highly rated outfit images from social media, and a normal model without fine-tuning. Both models generated multiple images featuring the input item, and object detection techniques (YOLO and CLIP) were used to identify and count frequently appearing items. Items that appeared more often in the outputs of the fashionable model were prioritized, and the corresponding images were ranked and presented as outfit recommendations. A subject experiment was conducted to evaluate the system, demonstrating that the proposed method can recommend stylish outfits and reflect query items more effectively than metadata-based recommendations.

Keywords: Outfit Recommendation · Fashion · Stable Diffusion.

1 Introduction

Most individuals likely face the challenge of selecting an appropriate clothing combination at least once daily; they transition from casual home attire to outfits suited for going out when leaving their homes each morning. For instance, they must consider questions such as, "Which pants should be paired with this shirt? If so, what should be worn as an outer layer?" In recent years, there has been a greater emphasis on coordinating clothing, or fashion sense. On widely adopted social media platforms, influencers frequently showcase stylish outfits, sharing their clothing combinations publicly. Moreover, this demand for fashion sense extends beyond influencers, as it has become an expectation for ordinary people to maintain a sense of style in their everyday attire. For example, platforms like BeReal[1] encourage users to post images of their outfits as they are in daily life.

[1] BeReal
https://bereal.com/

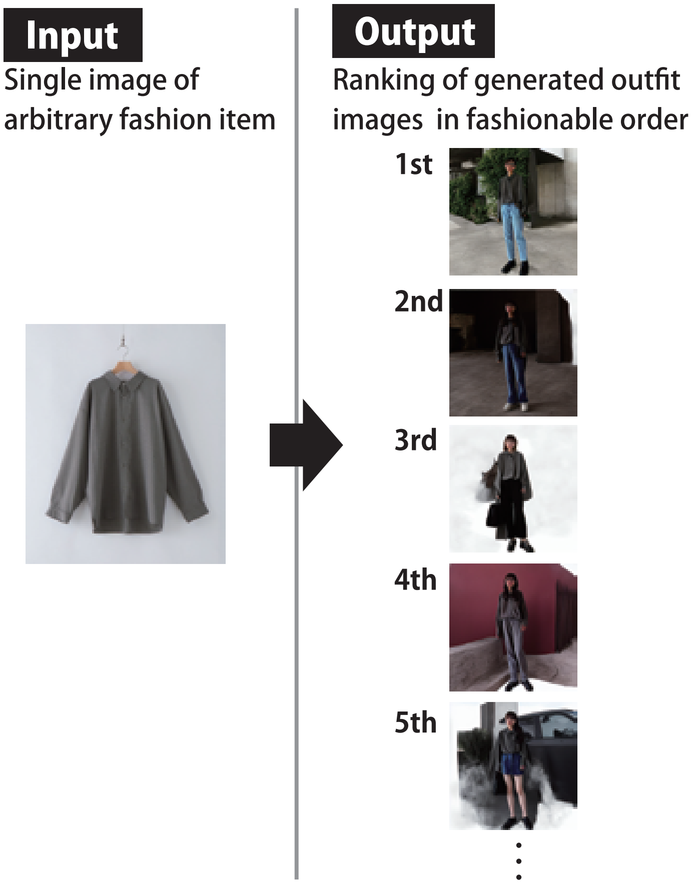
Input
Single image of arbitrary fashion item
Output
Ranking of generated outfit images in fashionable order
1st
2nd
3rd
4th
5th

Fig. 1. Sample input and output.

The importance of wearing well-coordinated, fashion-forward outfits on a daily basis has been increasing.

Some fashion social media platforms and apparel retail websites provide functionality for searching for outfits. However, searching for items to match a specific fashion item and complete a total outfit is often challenging. For instance, suppose a user wishes to find a coordinated outfit for a specific item (*e.g.*, “shirt with narrow vertical blue and green stripes and a wide collar, with fluffy silhouette”). Describing the visual characteristics of an item with search keywords is difficult, and it is not guaranteed that such fine-grained metadata has been assigned to the search target. This paper proposes an outfit generation algorithm that takes an image of any fashion item as input, and outputs an outfit image featuring the input item and the items appearing in the generated image, as shown in Figure 1. The algorithm utilizes an image generation AI (*i.e.*, Diffusion

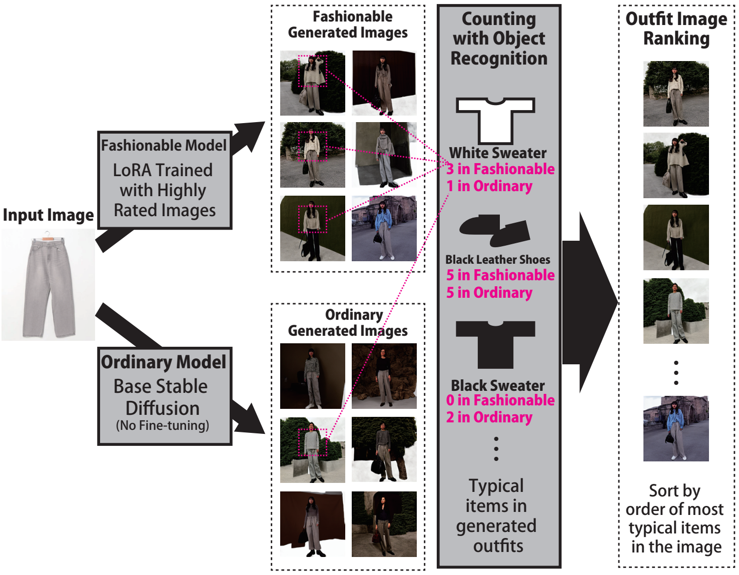

Fig. 2. Algorithm overview.

Model) when generating outfits from fashion item images. Image generation AIs excel at inferring and filling in missing portions of an image (*i.e.*, outpainting). Additionally, these models can be fine-tuned with specific image sets to embed distinctive characteristics into the model. By combining these capabilities, the model is fine-tuned using highly rated images from SNS platforms, enabling the generation of multiple images featuring the input fashion item. Object recognition is then applied to the generated images to identify frequently appearing items, allowing for the ranking of both images and items (see Figure 2).

Outfit recommendation based on our generation-based method offers several advantages. First, it allows a high degree of input flexibility. Image generation AI creates images based on visual norms, inferring what items typically appear around a given appearance. Consequently, it can generate outfits even for items that are difficult to describe in words. Second, the method could be more adaptable to non-existent items. It can discover items that pair well with new products that have not yet been posted on fashion social media, as well as with unknown handmade items.

To implement the recommendation algorithm, we developed a prototype. A dataset of highly rated outfit images was collected from social media, and Stable Diffusion [18] was fine-tuned on this set to generate images that incorporate the input fashion item. The system employed YOLO [17] and CLIP [16] for object detection, counting item occurrences, and assigning scores that prioritized frequent appearances, especially in images generated by the high-rated model. The outfit images were ranked accordingly, and user testing confirmed that the system achieved accuracy comparable to metadata-based recommendations.

The main contributions of this paper are as follows:

1. enabling the generation of fashionable outfit images by using highly rated images in social media to train an image generation AI,
2. allowing fashion item images to be used as input for outfit recommendation applications, and
3. making it possible to recommend outfits through outfit images generated by AI.

While a prior study [23] has applied image generation AI to outfit recommendation tasks, they have been limited to generating images of individual recommended items, without providing any visual indication of how those items would look when worn. Through experiments with the prototype, several areas for improvement were identified.

The structure of this paper is as follows: Section 2 reviews related work and positions our study within the relevant research landscape. The proposed method is described in Section 3. Section 4 presents the user study conducted to evaluate the effectiveness of our method. The experimental results are discussed in Section 5. Finally, Section 6 provides concluding remarks and discusses directions for future work.

2 Related Work

This study examines the generation of outfit images from a single fashion item image to recommend coordinated outfits, drawing on research in outfit recommendations and Virtual Try-on.

2.1 Outfit Recommendation

Research on fashion outfit recommendations has been actively conducted in recent years. Outfit recommendations must consider various factors, including the compatibility between fashion items and individual user preferences. Ding *et al.* [4] and Han *et al.* [6] focused on item compatibility using metadata. Chen *et al.* [1], Dong *et al.* [5], Sarkar *et al.* [19] and Jung *et al.* [11] incorporated user preferences with Transformer models, while Liu *et al.* [14] and Ye *et al.* [25] used fashion snap images for Situation-based recommendations. These approaches primarily utilize text-based information or image features as their items.

Unlike conventional outfit recommendation methods, this study introduces a novel approach by utilizing image generation to accept image input, and directly create and visually present outfits.

Moosaei *et al.* [15] and Xu *et al.* [23] have proposed methods that utilize image generation AI for outfit recommendation. However, the outfits recommended by these methods consist only of a set of fashion items, making it difficult to visualize how they would look when actually worn. The proposed method in this study recommends coordinated outfits as wearing images, where the input fashion items are worn by a subject.

In many previous studies, the Polyvore Dataset [6] and the iFashion Alibaba Dataset [1] have been commonly used as datasets. However, the term "coordination" in these datasets refers to a set of fashion item images, and thus, they cannot be used as outfit-wearing images. Moreover, these datasets do not contain labels related to "fashionableness." Therefore, in this study, we constructed a dataset from coordination images posted on social media, and defined highly-rated coordination images—those that received positive user feedback—as "fashionable coordinations."

2.2 Virtual Try-On

In recent years, there has been active research [2, 7, 8, 22] on Virtual Try-on (VTON) technologies that utilize algorithmic image editing techniques. VTON refers to methods that replace the clothing in a model's image with different garments. For instance, Shizhan *et al.* [28] proposed a VTON method using the sentence description of a new desired outfit.

In proposed VTON methods, it is possible to generate images in which a model is actually wearing the clothing item, allowing users to visualize how the item might look when worn. However, since the coordination with other items is typically determined by individual preferences, VTON does not provide information about complete outfit combinations based on a given fashion item. Furthermore, generating such try-on images typically requires not only the fashion item image but also a reference image of the person to be dressed.

This study proposes a method for generating try-on images from a single image of a fashion item. Specifically, we employ a diffusion model [9] to generate the outfit images. While diffusion models have recently been used in VTON research [3, 12, 24, 27], this study is the first to utilize them for the purpose of outfit recommendation by generating outfit try-on images.

3 Outfit Recommendation by Image Generation AI

This section describes the proposed method for recommending coordinated outfits using image generation AI. The algorithm takes an image of a single fashion item as its input, as shown in Figure 2. After preprocessing, the generated images are fed into two image generation models: the Fashionable Model and the Ordinary Model. The Fashionable Model is fine-tuned on highly rated social media outfit photos, whereas the Ordinary Model is a normal image generation AI without fine-tuning. Each image generation model produces images featuring the input item and other fashion items in a coordinated outfit. The system then uses object detection to estimate the categories of fashion items (*e.g.*, "brown leather shoes" or "gray denim pants") present in the generated images. The frequency of each item's appearance in the generated images is counted. Items that frequently appear, especially in the outputs of the Fashionable Model, are considered recommended items. The generated coordinated images are ranked as the output in order of appearance for these recommended items.

3.1 Fine-Tuning the Image Generation Model

First, an image generation model is trained using stylish fashion images. The commonly used Stable Diffusion model can be fine-tuned with a relatively small dataset to create a LoRA (Low-Rank Adaptation) [10] model that specializes in specific characteristics of the generated images. In this study, the LoRA model is trained on fashionably styled images to conduct a ranking focused on "fashionability." To achieve this, we constructed the dataset for fine-tuning.

The outfit images were collected from WEAR [2], a well-known and widely used fashion social media platform in Japan. This platform assigns ratings to numerous fashion images based on user votes, enabling a ranked display. Furthermore, some of the top-ranked users are recognized as verified users. Similar to previous studies that utilized social media data [13, 20, 21], this study regards outfits from highly-rated user posts as "fashionable."

In the experiment, we collected images of women's winter outfits posted by verified users. The images were then preprocessed by removing the background using Rembg[3] in Python. As a result, 500 images from verified users were prepared as the training dataset. Next, we fine-tuned Stable Diffusion using this dataset. The model trained on images from verified users is referred to as the "Fashionable Model," while the model without fine-tuning is referred to as the "Ordinary Model."

3.2 Generating Outfit Image Sets from the Input Image

Next, many images, including the user-specified query image, are generated using the two trained models. The query image contains a standalone fashion item. As a preprocessing step, extraneous parts of the item image were removed using Rembg. CLIP was then used to determine the category of the item. Depending on the item category, the item was repositioned to a predetermined location in the image, such as placing tops in the upper section, bottoms in the lower section, and hats at the very top.

Once the images were preprocessed in this manner, Stable Diffusion was used to fill in the blank areas. We used the ControlNet [26] extension of Stable Diffusion, specifically the OpenPose and Reference Only features. OpenPose can generate a person aligned with specified joint positions when the joint coordinates are provided along with the input image. This functionality ensured that limbs and the head were correctly positioned around the item to be worn. The Reference Only feature was applied when the query image was input. This feature allows Stable Diffusion to transform and generate the image while maintaining the characteristics of the input item, instead of merely filling in the external areas (out-painting). By using these techniques, outfit images were generated

[2] WEAR
https://wear.jp/

[3] GitHub: "danielgatis/rembg"
https://github.com/danielgatis/rembg

in which a standing person appears to be wearing the input fashion item, preserving the visual features of the original query image. The text prompt in the generating process is as

prompt:
1woman wearing “item category”,
full body,
fashionapp,
masterpiece,
best quality,
< LoRA:1 >
negative prompt:
(worst quality:1.2),
(low quality:1.2).

The term **item category** represents a coarse category of the input fashion item. This is used to prevent the loss of visual characteristics from the input image by avoiding overly detailed textual descriptions in the prompt. The keyword **fashionapp** serves as a trigger word to invoke the trained LoRA model. The Ordinary Model does not use the trigger word, and the LoRA Model. Although it is possible to generate more natural images by further refining the textual prompts, our proposed method deliberately employs minimal prompts to keep the system simple and focused.

3.3 Ranking the Generated Images

Finally, the outfits to be recommended were determined and ranked in order of their frequency of appearance in the outfit images. The two image generation models used in this study were the Fashionable Model, which was trained on images from fashionable verified users, and the Ordinary Model, which was not fine-tuned. In this study, we assume that items that frequently appear in the Fashionable Model’s generated results are both well-matched with the input item and fashionable. Based on this assumption, an image of a subject wearing multiple fashionable items is considered a fashionable outfit image. To determine whether an image is fashionable, image ranking is conducted by comparing the generated results of the Fashionable Model and the Ordinary Model. To rank the images, object recognition technology was first employed to identify items at a coarse granularity. YOLO was used to detect objects in each image, determining the location and type of item present (*e.g.*, pants, shirts, shoes). Next, the identified portions were cropped, and CLIP was used to generate detailed descriptions. This process classified the items in the images into categories with finer granularity, such as “brown leather shoes,” based on color and item type. As a data cleansing step, only items containing specific color names and category names from a predefined dictionary were retained. Items that appeared only a few times were removed as noise, which was considered an error from object detection or image generation.

The fashionableness of a given item can be defined as the ratio between the number of times the item appears in images generated by a vanilla image generation AI and the number of times it appears in images generated by a fine-tuned model. Let $p \in_{P_{gen}} (m_f, i_q)$ be an outfit image generated using the input fashion item i_q by the fashionable model m_f. The number of times i_q appears in the images generated by m_f, denoted as $n_{m_f}(i)$, is defined as:

$$n_{m_f}(i) = \sum_{p \in_{P_{gen}} (m_f, i_q)} |I_{inpict}(p) \cap \{i\}| \tag{1}$$

Similarly, the number of times item i appears in outfit images generated by the vanilla Stable Diffusion model m_u, denoted as $n_{m_u}(i)$, is defined as:

$$n_{m_u}(i) = \sum_{p \in_{P_{gen}} (m_u, i_q)} |I_{inpict}(p) \cap \{i\}|, \tag{2}$$

where $I_{inpict}(p)$ denotes the set of items detected in image p.

Based on the frequency of appearance of each item in the generated images from both models, we compute the fashionableness score for each item. By calculating the fashionableness score for each fashion item that appears in a generated outfit image and summing them, we can assign a fashionableness score to each individual generated image. The fashionableness score for a generated outfit image p produced from input image i_q is defined as:

$$deg(p) = \sum_{i \in I_{inpict}(p)} \frac{n_{m_f}(i) - n_{m_u}(i)}{(n_{m_f}(i) + n_{m_u}(i))^{\alpha}} \tag{3}$$

Here, α is a manually set weight that adjusts the influence of how frequently the item appears in both the fashionable model and the vanilla Stable Diffusion model.

Items with low frequency in the generated images were excluded from the fashionableness computation. Specifically, items that appeared fewer than twice in the images generated by the fashionable model were often due to detection errors or generation noise and were thus ignored.

Finally, we rank the generated outfit images from the fashionable model in descending order of fashionableness score.

4 Evaluation

A subject experiment was conducted to verify whether the proposed algorithm can effectively recommend outfits using our implemented prototype. The experiment involved two participants, staff members from a major apparel store.

Table 1. Questionnaire items for the user study. For each of the 450 images, participants answered five questions.

Evaluation Aspect	Question Items
Outfit	Is the outfit of the person in the image fashionable? How many fashionable items are shown in the image?
Image	Do the fashion items appear natural in the image? Does the subject appear natural in the image? To what extent is the input fashion item represented in the image?

4.1 Evaluation Task

Fashion coordination is a subjective task that requires expert judgment rather than crowd opinions. Thus, large-scale experiments via crowdsourcing are impractical. Instead, we conducted evaluations with a small number of fashion professionals to ensure high-quality assessments. We recruited two professional shop staff from UNIQLO[4], one of the world's leading apparel retailers. The actual questions are listed in Table 1. They labeled the top-ranked outfit images recommended by the five methods, along with the individual items within those images. The participant labeled the top five results for each of the five methods. There were 18 queries, covering three items each from various clothing categories: blouson, down jackets, pants, skirts, sweaters, and sweatshirts.

4.2 Dataset

The first dataset is used for training the LoRA model. It consists of 500 diverse outfit images used to train the Fashionable Model. All outfit images were collected from WEAR. For the experiment, we limited the collected images to those posted by verified female users during the winter season. To enable the image generation AI to produce fashionable images, we collected images from verified users. Additionally, we focused on winter outfits because such posts typically feature a wide variety and number of fashion items within a single outfit.

The second dataset consists of fashion item images used as input and contains 18 images. There are six fashion item categories, with three fashion items included in each category. All fashion item images were collected from ZOZOTOWN [5], the most famous fashion EC site in Japan. Only images in which the fashion items appear alone were included in the dataset.

The third dataset is used for training a model to detect fashion items in generated outfit images. It consists of 2,045 images and corresponding caption files that record the category and coordinates of each fashion item appearing

[4] UNIQLO
https://www.uniqlo.com/jp/ja/

[5] ZOZOTOWN
https://zozo.jp/

in the image. A dictionary of fashion item categories was manually created, and 120 posts were collected from WEAR for each category. From the collected images, we manually removed those deemed unsuitable as training data, such as images where the target item was not visible or only partially shown. As a result, 2,045 images were retained for use. Manual annotation was performed to prepare the dataset for training the object detection model. We used the annotation tool labelImg[6]. After annotation using labelImg, caption files were automatically generated. Out of the 2,045 image-caption pairs, 1,432 were used for training and 613 for evaluation.

The fourth dataset is used for a baseline method, which recommends outfit coordinations based on metadata co-occurrence. It consists of the outfit image URLs with their corresponding metadata. The metadata refers to the categories and colors of fashion items contained in outfit images in WEAR. Each coordination image is associated with multiple metadata entries, such as “skirt (white).”

4.3 Implementation

To evaluate the proposed method in practice, a working system is implemented. Stable Diffusion Web UI [7] is adopted as the image generation environment. The base model used in this environment was Stable Diffusion v1-5 [8].

It is worth noting that many newer models, such as SDXL [9], can generate higher-quality images with greater fidelity. However, these models tend to have a significantly larger number of parameters, making them less suitable for a wide range of tasks. Therefore, relatively lightweight and highly versatile Stable Diffusion v1-5 is chosen. The LoRA model was trained using Kohya’s GUI [10].

We set the weight α in Equation 3, which determines the influence of the fine-tuned model, to 0.9. This value was empirically determined through preliminary experiments. For conditioning during image generation, we used ControlNet, an extension of Stable Diffusion that enables additional control over outputs.

4.4 Comparative Methods

The proposed method in this paper incorporates three key innovations:

1. trained a LoRA model using posts from verified social media users,
2. compared the outputs of different image generation AIs, and

[6] GitHub: “HumanSignal/labelImg”
https://github.com/HumanSignal/labelImg

[7] GitHub: “AUTOMATIC1111 / stable-diffusion-webui”
https://github.com/AUTOMATIC1111/stable-diffusion-webui

[8] Hugging Face: “stable-diffusion-v1-5/stable-diffusion-v1-5”
https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5

[9] Hugging Face: “stabilityai/stable-diffusion-xl-base-1.0”
https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0

[10] GitHub: “bmaltais/kohya_ss”
https://github.com/bmaltais/kohya_ss

Table 2. Results of Outfit Evaluation by one professional (18 items, 4-point scale, * $p < 0.05$, ** $p < 0.01$ from Ordinary in Student's t-test)

	Proposed	Fashionable	Ordinary	Random	Metadata
Fashionability of Coordination	${}_{**}$3.05	${}_{**}$3.07	2.84	${}_{**}$3.04	${}_{**}$**3.36**
Fashionability of Items	${}_{**}$3.13	${}_{**}$3.04	2.69	${}_{**}$3.12	${}_{**}$**3.63**
Item Naturalness	3.50	3.46	3.39	3.44	${}_{**}$**3.92**
Person Naturalness	${}_{**}$3.29	${}_{**}$3.39	3.19	${}_{**}$3.29	${}_{**}$**3.94**
Query Reflectivity	${}_{**}$2.89	${}_{**}$**2.99**	2.57	${}_{**}$**2.99**	2.27

3. enabled fashion items to be input as images.

To evaluate the effectiveness of these innovations, we constructed four comparative methods. The methods compared were as follows:

- **Proposed**: the algorithm described in Section 3,
- **Fashionable**: using images generated by the Fashionable Model and ranking based solely on the frequency of items in them,
- **Ordinary**: using images generated by Ordinary Models and ranking based solely on the frequency of items in them,
- **Random**: selecting generated images randomly from the Fashionable Model, and
- **Metadata**: based on the co-occurrence of metadata.

The **Metadata** method collected co-occurrence data for simple tags (*e.g.*, "blue shirt" or "red trousers") assigned to each item on the fashion social media. It measured how often these tags appeared with the query item. Fashionability $deg_{metadata}$ of each outfit image is calculated by,

$$deg_{metadata}(p) = \sum_{d \in_D (i_q)} (\sum_{p \in_P (i_q)} |D(i_q) \cap \{d\}|), \tag{4}$$

where i_q is an inputed item, $P_{posted}(i_q)$ is a set of outfit images, and $D(i_q)$ is a set of inputed metadata d. The $deg_{metadata}$ tends to be higher for outfit images that contain metadata elements frequently observed across the entire set of outfit images. We compute the fashionableness score for each outfit image and rank them in descending order. As outfit recommendations, we present the top five ranked images. In the metadata-based method, real-world images from social media are used for the recommendations.

4.5 Experimental Result

Table 2 presents the experimental results. The responses to each question represent the average ratings given by two subjects for the coordination of all 18 types of fashion items.

Regarding the fashionability of the coordination, the proposed method received a significantly higher evaluation compared to the Ordinary Model. However, it did not receive higher ratings compared to the Fashionable Model or

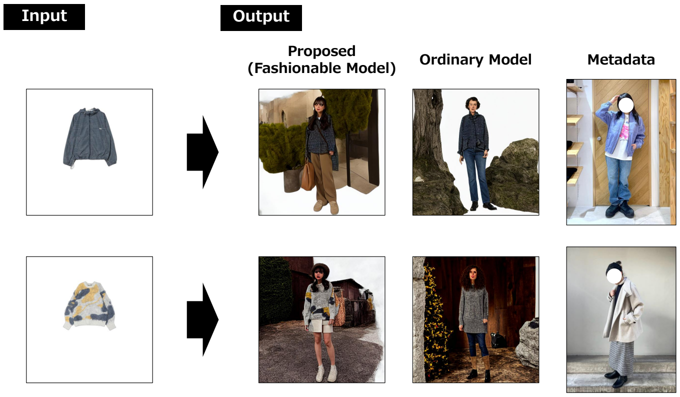

Fig. 3. Comparison of generation results by Fashionable Model in our proposed method, Ordinary Model, and Metadata-based method.

the Random model. For the fashionability of the outfit and the naturalness of fashion items and subjects, the metadata-based method received the highest evaluation. Concerning the query reflectivity, the methods using image generation AI (**Proposed**, **Fashionable**, **Ordinary**, and **Random**) generally received higher ratings than **Metadata**.

Figure 3 shows an example of an input fashion item image along with the recommended outfit images generated by the Fashionable Model, the Ordinary Model, and the metadata-based method. Compared to the Ordinary Model, which corresponds to the vanilla Stable Diffusion, and the metadata-based method, which does not allow image input, the Fashionable Model used in our proposed method generates recommended outfit images that not only reflect the characteristics of the input fashion item but also appear relatively fashionable.

5 Discussion

Based on the experimental results, we discuss the effectiveness and limitations of our proposed method. First, the proposed method successfully generated a large number of coordinated outfit images that not only incorporated many fashionable items but also accurately reflected the query item. This is evident from the high Fashionability of Items and Query Reflectivity scores.

Fine-tuning the model using fashion social media images enhanced the presence of appropriate items in the generated images. Notably, when comparing the **Proposed** and **Fashionable** models against the **Ordinary** model, significant improvements were observed in the Fashionability of Coordination score and the Fashionability of Items score, respectively. The results suggest that the fine-tuned image caption AI can feature images with numerous fashionable items.

However, when comparing the **Proposed** method with the **Fashionable** model, we observed a trend where the **Proposed** method outperformed in Fashionability of Items but slightly underperformed in Fashionability of Coordination. This result challenges our initial hypothesis. Our ranking strategy was implicitly based on two assumptions:

1. Items frequently appearing in the outputs of a model trained on fashion social media images are likely to be fashionable,
2. Outfit coordinations that include a large number of fashionable items are inherently fashionable.

In practice, while the first assumption proved correct, the second assumption did not hold. The items frequently generated by the Fashionable model were relevant to the query and fashionable in nature. However, simply incorporating fashionable items did not necessarily result in a well-coordinated outfit, as compatibility issues arose. Instead, excessively fashionable elements might have led to an unbalanced or visually unappealing ensemble.

This finding has important implications for the practical application of our proposed method. For recommending individual fashion items, a helpful approach would be to rank and present items that appear significantly more frequently in images generated by the fine-tuned model than in those generated by the Ordinary model. On the other hand, if the goal is to recommend complete outfit images, simply ranking the fine-tuned generated images based on their typicality may suffice.

Next, we compare the proposed method against the Metadata method, which does not rely on generative AI. The **Metadata**-based method outputs images with higher quality regarding the Naturalness of Item and Person score and overall coordination. This outcome is expected, as these images were fashion social media posts that received high ratings. However, while these images were visually well-received, they exhibited low Query Reflectivity. In other words, even if the outfit coordination was highly rated, the top-ranked images often had little relevance to the query item, making them unsuitable for item-based outfit recommendations. Ultimately, the proposed method successfully generated outfit images that included the query item while maintaining recommendation accuracy comparable to metadata-based approaches.

6 Conclusion

This paper demonstrated the feasibility of using image generation AI for outfit recommendation. Fine-tuned Stable Diffusion can recommend fashionable coordinates in response to an image query of a fashion item. Additionally, the proposed method more effectively reflects query items than metadata-based recommendations. On the other hand, the proposed method has some drawbacks: the low quality of the generated outfit images and the ineffectiveness of item-based ranking for fashionable coordinates. Future work will focus on addressing these limitations to create a more accurate and reliable recommendation system.

7 Acknowledgments

This work was partially supported by JSPS KAKENHI under Grant Numbers 25K03229, 25K03228, and 24K03228.

References

1. Chen, W., Huang, P., Xu, J., Guo, X., Guo, C., Sun, F., Li, C., Pfadler, A., Zhao, H., Zhao, B.: Pog: Personalized outfit generation for fashion recommendation at alibaba ifashion. In: Proc. of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data mining. pp. 2662–2670 (2019)
2. Choi, S., Park, S., Lee, M., Choo, J.: Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14131–14140 (2021)
3. Cui, A., Mahajan, J., Shah, V., Gomathinayagam, P., Liu, C., Lazebnik, S.: Street tryon: Learning in-the-wild virtual try-on from unpaired person images. In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8235–8239 (2024)
4. Ding, Y., Mok, P., Ma, Y., Bin, Y.: Personalized fashion outfit generation with user coordination preference learning. Information Processing & Management **60**(5), 103434 (2023)
5. Dong, X., Song, X., Feng, F., Jing, P., Xu, X.S., Nie, L.: Personalized capsule wardrobe creation with garment and user modeling. In: Proc. of the 27th ACM International Conference on Multimedia. pp. 302–310 (2019)
6. Han, X., Wu, Z., Jiang, Y.G., Davis, L.S.: Learning fashion compatibility with bidirectional LSTMs. In: Proc. of the 25th ACM International Conference on Multimedia. pp. 1078–1086 (2017)
7. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7543–7552 (2018)
8. He, S., Song, Y.Z., Xiang, T.: Style-based global appearance flow for virtual try-on. In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3470–3479 (2022)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
10. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: Proc. of International Conference on Learning Representations (2022)
11. Jung, M.C., Monteil, J., Schulz, P., Vaskovych, V.: Personalised outfit recommendation via history-aware transformers. In: Proc. of the Eighteenth ACM International Conference on Web Search and Data Mining. pp. 633–641 (2025)
12. Kim, J., Gu, G., Park, M., Park, S., Choo, J.: Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8176–8185 (2024)
13. Li, Y., Cao, L., Zhu, J., Luo, J.: Mining fashion outfit composition using an end-to-end deep learning approach on set data. IEEE Transactions on Multimedia **19**(8), 1946–1955 (2017)

14. Liu, S., Feng, J., Song, Z., Zhang, T., Lu, H., Xu, C., Yan, S.: Hi, magic closet, tell me what to wear! In: Proc. of the 20th ACM International Conference on Multimedia. pp. 619–628 (2012)
15. Moosaei, M., Lin, Y., Akhazhanov, A., Chen, H., Wang, F., Yang, H.: Outfitgan: Learning compatible items for generative fashion outfits. In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2273–2277 (2022)
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
17. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
18. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
19. Sarkar, R., Bodla, N., Vasileva, M.I., Lin, Y.L., Beniwal, A., Lu, A., Medioni, G.: Outfittransformer: Learning outfit representations for fashion recommendation. In: Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3601–3609 (2023)
20. Simo-Serra, E., Fidler, S., Moreno-Noguer, F., Urtasun, R.: Neuroaesthetics in fashion: Modeling the perception of fashionability. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 869–877 (2015)
21. Tangseng, P., Yamaguchi, K., Okatani, T.: Recommending outfits from personal closet. In: Proc. of the IEEE International Conference on Computer Vision Workshops. pp. 2275–2279 (2017)
22. Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., Yang, M.: Toward characteristic-preserving image-based virtual try-on network. In: Proc. of the European Conference on Computer Vision (ECCV). pp. 589–604 (2018)
23. Xu, Y., Wang, W., Feng, F., Ma, Y., Zhang, J., He, X.: Diffusion models for generative outfit recommendation. In: Proc. of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1350–1359 (2024)
24. Xu, Y., Gu, T., Chen, W., Chen, C.: Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. arXiv preprint arXiv:2403.01779 (2024)
25. Ye, T., Hu, L., Zhang, Q., Lai, Z.Y., Naseem, U., Liu, D.D.: Show me the best outfit for a certain scene: A scene-aware fashion recommender system. In: Proc. of the ACM Web Conference 2023. pp. 1172–1180 (2023)
26. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proc. of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
27. Zhu, L., Yang, D., Zhu, T., Reda, F., Chan, W., Saharia, C., Norouzi, M., Kemelmacher-Shlizerman, I.: Tryondiffusion: A tale of two unets. In: Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4606–4615 (2023)
28. Zhu, S., Urtasun, R., Fidler, S., Lin, D., Change Loy, C.: Be your own prada: Fashion synthesis with structural coherence. In: Proc. of the IEEE International Conference on Computer Vision. pp. 1680–1688 (2017)