FASHIONDISTILL: GENERATIVE DESIGN SYNTHESIS VIA BESTSELLER PATTERN EXTRACTION FROM THE H&M DATASET

MING-LUN HSIEH, HSIANG HSIEH, YAN-HONG CHEN, PIN-JOU LIN, MU-EN CHIU, LI-HONG GUO

National Taiwan University

ABSTRACT

Fashion design generation aims to assist designers by providing visual inspiration that reflects diverse trend elements under extremely short product cycles. It also helps translate abstract textual concepts into vivid visual prototypes. Recently, generative AI models have gained widespread attention for their ability to synthesize fashion items from prompts. However, existing models often suffer from limited diversity and rely heavily on supervised learning paradigms, which restrict their adaptability to fast-changing market dynamics. To address these limitations, we propose an innovative framework, **FashionDistill**, which fine-tunes fashion design generation models via Direct Preference Optimization (DPO). Our framework offers a generalizable fine-tuning strategy that enables the model to better align with real-world fashion preferences. To ensure comprehensive and objective feedback, we introduce a multi-feedback module that incorporates text-to-text, image-to-image and image-to-text evaluations. This multi-check mechanism enables cross-modal validation of generation quality and stylistic suitability. Experiments conducted on real-world transaction datasets from H&M demonstrate the effectiveness of our approach in enhancing the model's ability to reflect evolving market trends while preserving feasible and expressive visual design principles. Source code is publicly available at: this https URL.

***Index Terms*—** Generative AI, Text-to-image synthesis, Trend forecasting, Direct Preference Optimization, Latent diffusion

1. INTRODUCTION

In the fast fashion industry, anticipating consumer preferences is both a design challenge and a commercial imperative. Unlike traditional apparel production cycles, fast fashion brands typically refresh their collections every week [1]. This compressed timeline forces designers and retailers to make rapid decisions with limited information. In addition, recent fast fashion trends reveal a notable resurgence of retro styles. To appeal to consumers, new designs are often expected to strike a balance between novelty and familiarity, introducing fresh visual ideas while subtly incorporating elements that evoke past popular aesthetics. Therefore, designers must not only be creative, but also deeply aware of historical style references that continue to influence current demand. Designing in this context requires a deeper awareness of evolving market cues, stylistic references, and temporal relevance.

At the same time, the rise of large-scale transaction-level data from global retailers and the advancement of generative AI models capable of producing high-quality, photorealistic clothing images have opened up new possibilities in fashion design support. While recent works in text-to-image diffusion models have achieved remarkable progress in producing visually compelling results, aligning such generative outputs with real consumer demand remains a largely unsolved problem.

In this work, we introduce a market-aware fashion generation framework that explicitly incorporates sales feedback from real bestseller items. Our method is designed to not only synthesize realistic and diverse fashion images, but also to learn from what has actually sold well in the past, and what is likely to succeed in the near future. Specifically, our proposed framework consists of three key components:

1. A large language model is trained to generate pseudo textual descriptions for anticipated bestsellers, conditioned on prior weeks' top-selling item multimodal information. This step models the temporal evolution of consumer preferences in a descriptive form.

2. A generative model takes as input both the pseudo text and structured item visual inputs (e.g. product image, segmentations) to synthesize candidate fashion images.

3. The generated image is evaluated through a multi-feedback mechanism comprising three components: (1) visual similarity to real bestseller product images from the following week, (2) semantic alignment with the corresponding textual description, and (3) consistency between the pseudo textual prompt and the actual future bestseller description.

To validate our approach, we conduct experiments using transaction data from H&M, covering 6 product categories across 106 weeks, including accessory, underwear, shoes, full body, lower body, and upper body. In this dataset, it shows that most fast fashion items reach peak sales within their first

week of launch, underscoring the importance of early popularity and fast response to consumer taste. These findings motivate our design of a week-over-week learning loop, where each generation cycle attempts to anticipate the next successful item from actual historical patterns.

Ultimately, our system aims to support designers in identifying the stylistic features that contribute to market success. By grounding the generative process in real-world outcomes and combining both image- and text-level feedback, we close the gap between consumer behavior and creative generation. Our approach offers a pathway toward data-informed design iteration that balances novelty with familiarity, enabling designers to create collections that are both expressive and commercially relevant.

2. RELATED WORK

2.1. Fashion Image Generation

Fashion image generation refers to the task of synthesizing fashion-related visuals using deep learning models. It has supported a wide range of applications such as virtual try-on systems, automated clothing design, and personalized fashion recommendation. In this context, computer vision can thus be used to improve the fashion design process. Generative models in this space have the potential to accelerate ideation and enhance design efficiency.

Previous works, such as MGD [2], focus on generating human-centric fashion images conditioned on multimodal prompts, including text, human body poses, and garment sketches. Built on latent diffusion models, MGD enhanced by the use of pose map conditioning and incorporation of sketches to enrich textual input with spatial details. Compared with baseline models like Stable Diffusion and SDEdit, MGD achieves significantly better performance across multiple evaluation metrics, including FID, KID, CLIP-S, Pose Distance, and Sketch Distance, demonstrating stronger visual realism and cross-modal coherence.

Given its ability to support controllable and multimodally guided fashion generation, MGD provides a strong foundation for downstream tasks involving trend-driven design synthesis and temporal comparison with real-world fashion data.

2.2. Direct Preference Optimization

Direct Preference Optimization (DPO) is a recent fine-tuning approach that leverages preference comparisons rather than explicit reward signals to align generative models with human intent. Originally proposed to reduce the complexity of reinforcement learning in language models, DPO has since been adapted to image generation tasks, including diffusion-based models.

In the context of fashion generation, FashionDPO [3] applies DPO to fine-tune a diffusion-based outfit generator using automatically constructed positive–negative image pairs. These pairs are generated from a multi-feedback module composed of expert models that assess each sample along three aspects: image quality, clothing compatibility, and personalization. The feedback scores are used to select better-performing generations, forming preference pairs without the need for human annotation or explicit reward design. The model is then fine-tuned with DPO to prefer the positively scored samples. This strategy enhances the diversity and quality of generated fashion images and achieves strong results on public benchmarks such as iFashion and Polyvore-U across metrics including Intra LPIPS, Preference Accuracy, and Win Rate.

Despite its success, this approach relies heavily on synthetic feedback from internal evaluators, which may fail to reflect real-world consumer behavior or rapid changes in market trends. To overcome this limitation, our method incorporates actual sales performance as an implicit form of preference signal. By grounding feedback in weekly top-selling items, we enable the model to adapt to real-time market preferences and generate designs that are both trend-aware and commercially viable.

3. METHODOLOGY

We propose **FashionDistill**, a trend-aware fashion generation framework that synthesizes fashion designs conditioned on multimodal inputs and refined through preference-based fine-tuning. The overall pipeline is illustrated in Figure 1.

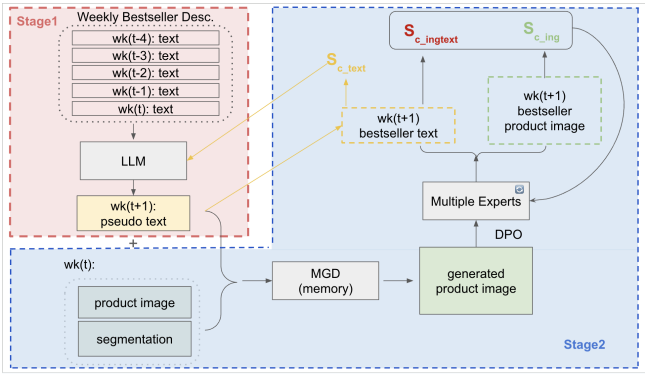


Fig. 1: The Overview of FashionDistill, which consists of two key stages: 1) Predicted Bestseller Description, 2) Fashion Design Generation

3.1. Weekly Bestseller Selection

We begin by identifying weekly top-selling items from historical transaction records. For each product category, we select the top-5 selling items within the current week to construct a dynamic trend reference. This set captures both established and emerging popularity signals, forming the basis for downstream prompt construction.

3.2. Prompt Construction via LLM

To create descriptive and trend-relevant prompts, we employ a two-stage prompt generation strategy. First, we use a large vision-language model, **LLaVA** [4], to generate detailed natural language descriptions for each product image, capturing visual attributes such as texture, color, and style. We collect these descriptions from the top-selling items over the past 4 weeks. Then, we encode this sequence of textual data as a time-series input into a pretrained Transformer-based language model, **Vicuña** [5]. Vicuña generates *pseudo prompts* that capture the predicted key attributes of next-week bestsellers by inferring trends from recent weeks. These synthetic textual prompts are used to guide the image generation process.

3.3. Image Generation with MGD

We adopt and extend the **MGD**, a latent diffusion model conditioned on multiple visual and textual modalities. To provide structured visual cues, we use **U2Net** [6] for image segmentation, labeling different class regions such as shirts, skirts and socks. Due to the architectural constraints of our backbone model **inpaint-v2** [7], which currently does not support multiple segmentation masks simultaneously, we adopt a simplified segmentation strategy that separates the foreground garment from the background. This approach is sufficient for producing coherent fashion layouts under current settings. However, finer-grained segmentation may become necessary in future work, especially for scenarios involving multiple color blocks or layered apparel components. The resulting segmentation masks, when combined with pseudo textual prompts, serve as conditioning signals for MGD, enabling the generation of plausible and stylistically aligned fashion images that reflect the predicted trend.

3.4. Preference-Based Fine-Tuning via DPO

To improve alignment with real-world fashion dynamics, we incorporate feedback from actual future bestseller data. We construct preference pairs using two feedback signals:

Based on these signals, we construct ranked image pairs (positive vs. negative) and fine-tune the generative model using **DPO**. This allows the model to prefer samples that are stylistically and semantically closer to real bestsellers, promoting market-aware generation without the need for manual annotation or reward engineering.

We use the following loss function to fine-tune the model:

$$L_{\text{total}} = L_{\text{diffusion}} + \alpha \cdot L_{\text{temporal}} + \beta \cdot L_{\text{DPO}}$$

$$L_{\text{diffusion}} = \mathbb{E}_{\mathcal{E}(I),Y,\epsilon\sim\mathcal{N}(0,1),t,\mathcal{E}(I_M),m,p,s} \left\|\epsilon - \epsilon_\theta(\gamma,\psi)\right\|_2^2$$

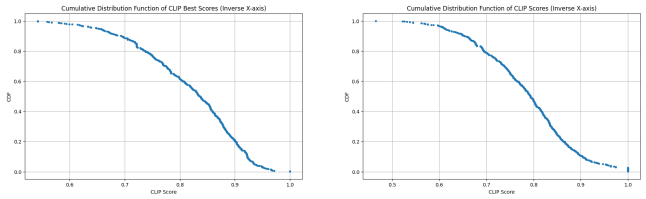
$$L_{\text{temporal}} = \lambda \cdot \frac{1}{d} \left\| f_t - \sum_{i=1}^{n} \gamma^i f_{i-1} \right\|^2$$

$$\nabla_\theta L_{\text{DPO}}(\pi_\theta;\pi_{\text{ref}}) = -\beta\, \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \Bigg[\underbrace{\sigma\left(\hat{r}_\theta(x,y_l) - \hat{r}_\theta(x,y_w)\right)}_{\text{reward estimate wrong}} \times \Big(\underbrace{\nabla_\theta \log \pi(y_w \mid x)}_{\text{increase } y_w} - \underbrace{\nabla_\theta \log \pi(y_l \mid x)}_{\text{decrease } y_l} \Big) \Bigg]$$

4. EVALUATION & RESULTS

4.1. Text Prompt

To assess the quality of the generated pseudo prompts during training, we employ a text-to-text semantic similarity evaluation using the CLIP score. Specifically, we measure how well each predicted description aligns with the ground-truth bestseller text. As illustrated in Figure 2, the resulting cumulative distribution functions (CDFs) indicate that our method achieves a higher average CLIP score (0.8181) compared to a baseline model without fine-tuning (0.7840), reflecting improved prompt construction and trend alignment. This evaluation validates that the learned textual representations better capture relevant attributes present in real-world bestseller descriptions.



(b) w/ Fine-tuning (c) w/o Fine-tuning

Fig. 2: Inverse CDFs of Clip Score

4.2. Image Generation

First, we can compare the performance of our model to the original model without fine-tuning. To better understand the impact of each component in our proposed framework, we also conduct an **ablation study** focusing on the effect of DPO. From Table 1, we can find that our model can apply the text description better.

Besides observing some examples, we perform quantitative evaluation to assess how well the generated fashion images align with expected trends. Since our goal is to generate designs that resemble actual future bestsellers and match their descriptive attributes, we employ CLIP-based metrics for cross-modal similarity. CLIP offers robust embeddings for both image and text modalities, making it suitable for evaluating visual-semantic alignment in our setting. We adopt the following two metrics:

Prompt	Stable-Diffusion-2-inpainting	Ours w/o DPO	Ours w/ DPO
Stylish, earthy tote with brown leather trim.. Tan leather, woven **straw body**.. Purpose: Carry items.			
Classic, chic, black blazer with belt and cuff details.. Black, satin, peplum jacket with belt.. Stylish, smart, versatile business blazer for women.			
Smart-casual pants with ribbed waistband, side pockets, and a slightly flared leg.. Black, stretchy fabric.			
Blue collared shirt with white buttons, short sleeves, and a pointed collar.. Blue cotton polo shirt with buttoned collar.			
Tall, skinny, black ankle boot with high heel.. Black, textured leather.. Designed for fashion, style, and sophistication.			

Table 1: Comparison of three models on five prompts

1. **Image similarity** (*CLIP-I*): The cosine similarity between the generated image and the real next-week bestseller image using CLIP embeddings;

2. **Text-image alignment** (*CLIP-T*): The alignment between the generated image and the real bestseller description using CLIP's cross-modal scoring;

The result is showed in Table 2. We can find that our model with DPO has a higher CLIP score.

Method	CLIP-I	CLIP-T
Stable-Diffusion-2-inpainting	0.741	0.24
Ours w/o DPO	0.785	0.25
Ours w/ DPO	**0.791**	**0.26**

Table 2: Performance Comparison on CLIP-I and CLIP-T

5. DISCUSSION

While our proposed framework demonstrates promising results in aligning generated fashion images with real-world bestseller trends, several challenges and limitations remain.

First, the model's capacity to generate entirely novel styles is constrained. Because the generation process relies heavily on historical bestseller patterns and pseudo prompts distilled from existing products, the system tends to produce variations or recombinations of previously observed elements rather than introducing radically innovative designs. This reflects a fundamental limitation of training on empirical sales data that while it reinforces what has worked in the past, it may hinder exploration into untested, forward-looking fashion concepts. As a result, the generative model may be more effective in trend-following rather than trend-initiating contexts.

Second, seasonality introduces an additional layer of complexity. Consumer preferences in fashion might be sensitive to seasonal factors, such as temperature, holidays, and cultural events, that influence both product design and purchasing behavior. Although our framework leverages week-by-week sales data, it does not explicitly model seasonality as a distinct variable. Consequently, designs generated during transitional or atypical periods (e.g., between summer and fall) may not capture temporal signals accurately. Integrating seasonal indicators or external calendar-based features may enhance the temporal robustness of future models.

Despite these limitations, our framework represents a meaningful step toward integrating real-world consumer behavior into generative fashion design. Future work can explore mechanisms for introducing novelty beyond past data patterns, as well as incorporating external temporal signals to improve robustness across seasons. Addressing these challenges will be essential for developing more adaptive and forward-looking fashion generation systems.

6. WORK DISTRIBUTION

- MING-LUN HSIEH: paper survey, dataset preprocessing, stage 2 training & evaluation, code organization
- HSIANG HSIEH: paper survey, dataset preprocessing, stage 1 training & evaluation, oral presentation
- YAN-HONG CHEN: paper survey, baseline model, oral presentation
- PIN-JOU LIN: paper survey, slides, written report
- MU-EN CHIU: paper survey, slides, written report
- LI-HONG GUO: paper survey, functions testing

7. REFERENCES

[1] N. K. Y. Diantari, "Trend cycle analysis on fast fashion products," *Journal of Aesthetics, Design, and Art Management*, 2021.

[2] A. Baldrati, D. Morelli, G. Cartella, M. Cornia, M. Bertini, and R. Cucchiara, "Multimodal garment designer: Human-centric latent diffusion models for fashion image editing," 2023.

[3] M. Yu, Y. Ma, L. Wu, C. Wang, X. Li, and L. Meng, "Fashiondpo:fine-tune fashion outfit generation model using direct preference optimization," 2025.

[4] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," 2024.

[5] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," 2023.

[6] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern Recognition*, vol. 106, p. 107404, Oct. 2020.

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.