StyleTailor: Towards Personalized Fashion Styling via Hierarchical Negative Feedback

Hongbo Ma[1], **Fei Shen**[2], **Hongbin Xu**[3], **Xiaoce Wang**[1], **Gang Xu**[4]
Jinkai Zheng[5], **Liangqiong Qu**[6], **Ming Li**[4*]
[1] Tsinghua University
[2] National University of Singapore
[3] Bytedance Seed
[4] Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
[5] Hangzhou Dianzi University
[6] The University of Hong Kong

Abstract

The advancement of intelligent agents has revolutionized problem-solving across diverse domains, yet solutions for personalized fashion styling remain underexplored, which holds immense promise for promoting shopping experiences. In this work, we present StyleTailor, the first collaborative agent framework that seamlessly unifies personalized apparel design, shopping recommendation, virtual try-on, and systematic evaluation into a cohesive workflow. To this end, StyleTailor pioneers an iterative visual refinement paradigm driven by *multi-level negative feedback*, enabling adaptive and precise user alignment. Specifically, our framework features two core agents, *i.e.*, *Designer* for personalized garment selection and *Consultant* for virtual try-on, whose outputs are progressively refined via hierarchical vision-language model feedback spanning individual items, complete outfits, and try-on efficacy. Counterexamples are aggregated into negative prompts, forming a closed-loop mechanism that enhances recommendation quality. To assess the performance, we introduce a *comprehensive evaluation suite* encompassing style consistency, visual quality, face similarity, and artistic appraisal. Extensive experiments demonstrate StyleTailor's superior performance in delivering personalized designs and recommendations, outperforming strong baselines without negative feedback and establishing a new benchmark for intelligent fashion systems.

Code — https://github.com/Ma-Hongbo/StyleTailor
Extended version — https://arxiv.org/abs/2508.06555

1. Introduction

Generative multimodal learning (Brown et al. 2020; Touvron et al. 2023; Durante et al. 2024; Qwen et al. 2025; Li et al. 2025; Zhao et al. 2025; Liu et al. 2025, 2024; Li et al. 2024; Miao et al. 2025; Shi et al. 2025; Xu et al. 2025b; Su et al. 2025) has achieved remarkable breakthroughs in recent years, enabling models to seamlessly integrate and process diverse data modalities, thereby transforming their role in practical applications. Intelligent agents, built upon these advancements, represent a powerful paradigm for creating automatic workflows for complicated and originally human-interactive tasks (Zi-Yi et al. 2023; Seo et al. 2025; Pang et al. 2025; Koh et al. 2025; Yue et al. 2025). For example, agents have significantly streamlined paperwork-related activities by enabling automatic chart generation (Koh et al. 2025), paper-to-code conversion (Seo et al. 2025), and the creation of scientific posters from manuscripts (Pang et al. 2025). Beyond administrative automation, agents are also transforming scientific disciplines that demand specialized expertise, such as computational fluid dynamics simulation (Zi-Yi et al. 2023; Yue et al. 2025). These applications underscore the versatility and impact of agent-based systems.

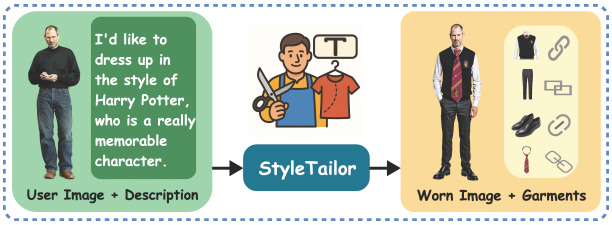

Figure 1: We present StyleTailor, the *first* agentic framework for personalized fashion styling that, given a full-body photo and dressing preferences, outputs virtual try-on results, curated garment images, and shopping links in a unified, closed-loop pipeline, advancing user-centric, interactive fashion recommendation.

Although substantial advancements have been made, the development of a unified agent framework for personalized fashion styling remains unaddressed. By tailoring recommendations to user preferences and appearance characteristics, personalized styling systems streamline decision-making, elevate user satisfaction, and are pivotal for increasing e-commerce traffic and operational efficiency (Zhu et al. 2023; Shen et al. 2025a,b).

However, constructing a collaborative agent framework for personalized fashion styling presents considerable technical difficulties owing to the intricate, fine-grained nature of the task. Despite significant progress in vision-language models (VLMs), existing systems continue to suffer from

*Corresponding Author

limited reasoning capabilities and persistent hallucination effects. Disparities in training data distributions and architectural differences among various VLMs further exacerbate inconsistencies, resulting in unreliable and highly domain-dependent outputs (Ji et al. 2023; Ling et al. 2024; Huang et al. 2025). Consequently, ensuring robust reliability, accuracy, and trustworthiness in VLM-based applications remains a substantial challenge.

Accordingly, the effective coordination of multiple agents to exploit the complementary advantages of heterogeneous VLMs constitutes a critical and unresolved research problem. Prevailing agentic frameworks predominantly rely on simplistic, repetitive selection or random output refinement strategies, which incur substantial computational cost yet often yield limited qualitative improvement (Qiao et al. 2024). Thus, establishing principled mechanisms for efficient feedback integration is an essential challenge for advancing collaborative agent systems in complex, user-centric scenarios.

To address these challenges, we introduce StyleTailor, the first collaborative agent framework with a *hierarchical negative feedback* mechanism to seamlessly integrate individualized garment design, shopping recommendation, virtual try-on, and systematic evaluation within a unified pipeline. The proposed framework is composed of two principal modules, *i.e.*, *Designer* and *Consultant*. As illustrated in Fig. 1, StyleTailor receives a user-provided reference image and dressing style preference description, and outputs virtual try-on visualization, associated garment visuals, and direct purchase links, providing an end-to-end solution for personalized fashion experiences.

The *Designer* employs a cascade of sequential expert agents, each interpreting the inputs by a VLM to generate a standardized set of fine-grained garment specifications across clothing components. Leveraging these attributes, a search engine (*e.g.*, Google Custom Search API) retrieves curated garment images and associated product links. To enforce text-outfit consistency and enable iterative refinement, we incorporate a two-level negative feedback mechanism: (i) at the item level, during the search phase, a VLM analyzes discrepancies between unsatisfactory results and the original prompt, converting them into negative prompts to guide subsequent searches; and (ii) at the outfit level, where the current expert proposes a complete outfit set—if deemed unsatisfactory, the next expert is activated, incorporating prior suboptimal outputs as explicit negative examples, continuing until convergence on a high-quality result.

The *Consultant* utilizes an advanced image-editing model to enable virtual try-on, synthesizing photorealistic images of the user in recommended outfits conditioned on both visual inputs, *i.e.*, user and garment images, and textual prompts. To achieve precise fashion evaluation and alignment with user preferences, we introduce a higher-level negative feedback paradigm that iteratively refines try-on visualizations. The suboptimal results are scrutinized by a VLM to identify discrepancies, which are then converted into negative prompts guiding subsequent generations until optimal consistency and quality are attained.

To comprehensively evaluate our effectiveness, we propose an assessment suite comprising complementary metrics tailored to personalized fashion styling. Style consistency is quantified via VQAScore (Lin et al. 2024), assessing alignment between synthesized images and user preferences. Visual quality is evaluated using IQAScore (Chen and Mo 2022), verifying high-fidelity generative outputs. Face similarity is measured with InsightFace (Ren et al. 2023), ensuring minimal identity distortion. Finally, aesthetic appraisal leverages VLM-based evaluators for a holistic artistic and stylistic critique. This suite establishes a robust benchmark for agent-driven fashion systems, enabling precise validation of refinement mechanisms.

Our main contributions can be summarized as follows.

- We introduce StyleTailor, the first collaborative agent framework that seamlessly integrates personalized fashion design, shopping recommendation, virtual try-on, and systematic evaluation into a unified pipeline, addressing a key gap in multimodal computer vision for user-centric applications.
- We propose a hierarchical negative feedback mechanism embedded within the agent system, spanning three progressive levels: item-specific refinement, outfit-level coordination, and virtual try-on optimization. This iterative approach leverages vision-language models to enhance accuracy, mitigate hallucinations, and enable adaptive visual refinement.
- We present a comprehensive evaluation suite, assessing style consistency, visual quality, facial similarity, and holistic aesthetic appraisal, thereby establishing a robust benchmark for agent-driven fashion systems.

2. Related Works

2.1. Multimodal Agents

The recent surge in intelligent agents has showcased their versatility across software engineering, scientific computing, and visual content creation. Powered largely by LLMs, these agents decompose complex tasks into actionable steps, enabling sequential reasoning, dynamic interaction, and multimodal generation (Yao et al. 2023; Xu et al. 2025a). They have been applied to slide and poster generation (Zheng et al. 2025; Pang et al. 2025), scientific chart creation (Koh et al. 2025), research and software code development (Wang et al. 2025; Seo et al. 2025), and physics or chemistry simulations (Zi-Yi et al. 2023; Yue et al. 2025), demonstrating strong generalization and coordination capabilities. However, agent-based approaches for personalized fashion workflows—spanning coordinated garment design, recommendation, and virtual try-on—remain underexplored. Furthermore, existing agents often lack explicit mechanisms for iterative refinement and user-centered feedback integration, constraining their adaptability in fashion-oriented tasks.

2.2. Fashion Virtual Try-on Models

Virtual try-on (VTON) methods can be broadly divided into image-based and prompt-based paradigms. Image-based models take a user image and a reference garment as input. Early methods such as VITON (Han et al. 2018) and

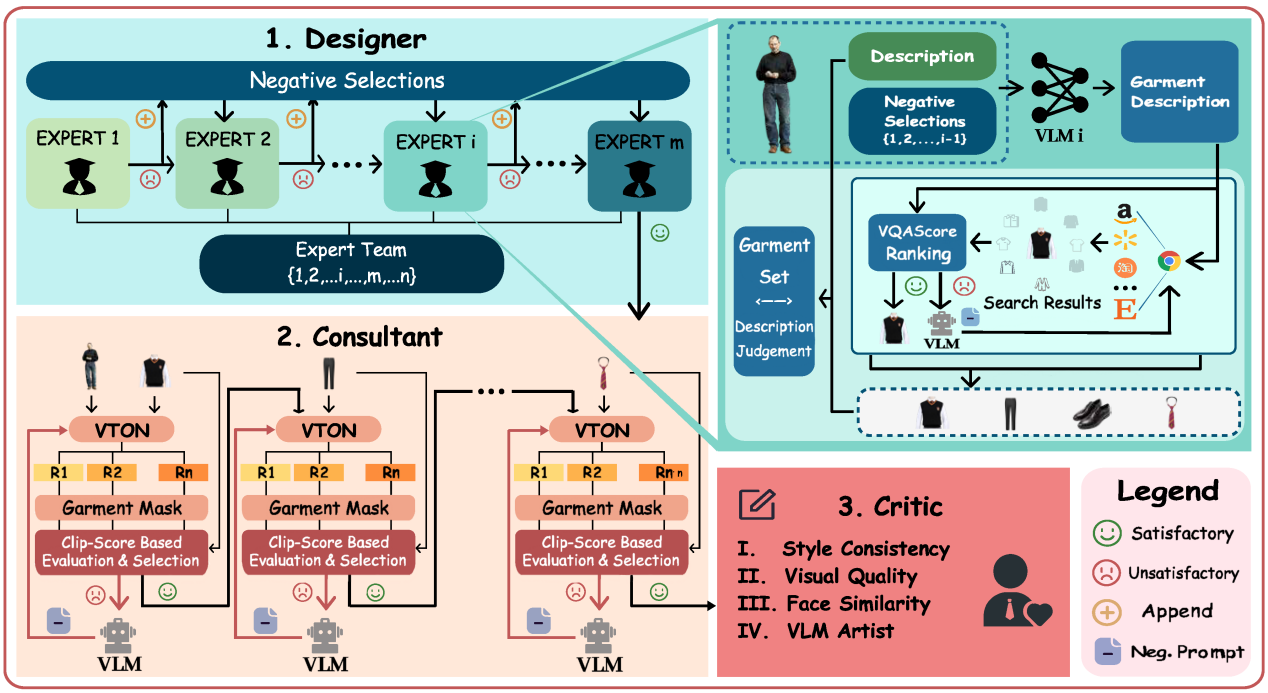

Figure 2: Overview of our agent framework StyleTailor. The *Designer* module analyzes the user-provided image and style preferences, generates garment specifications, and retrieves suitable clothing images. Two hierarchical negative feedback mechanisms within the *Designer* refine retrieval at the item and outfit levels. The *Consultant* module generates virtual try-on results and applies higher-level feedback to further improve alignment with user requirements. The *Critic* quantitatively evaluates the final outputs. This multi-stage feedback ensures the system progressively optimizes recommendations.

CP-VTON introduced warping and composition networks within GAN frameworks, followed by flow-based (Han et al. 2019) and 3D-aware approaches (Kubo et al. 2019) that improved alignment and realism. Recent diffusion-based models (Zhu et al. 2023; Shen et al. 2025a,b) further enhance fidelity via progressive denoising, with additional advances addressing pose variation (Raj et al. 2018) or reducing reliance on paired data (Neuberger et al. 2020). Prompt-based models instead use textual or multimodal cues to guide garment manipulation. Text2Human (Jiang et al. 2022) frames VTON as text-to-image synthesis, whereas multimodal systems (Baldrati et al. 2023) integrate text and reference images for joint control. The rise of diffusion frameworks such as Stable Diffusion (Rombach et al. 2021) and ControlNet (Zhang, Rao, and Agrawala 2023) has further improved precision and scalability in text-guided editing. In this work, we build upon FLUX.1.Kontext (Labs et al. 2025) as our core image-editing model for VTON. Its multimodal conditioning enables fine-grained garment rendering and accessory augmentation through seamless integration of visual and textual guidance.

3. StyleTailor

3.1. Problem Formulation

We define personalized fashion styling as a collaborative generation and selection task. Given a full-body reference image of the user I_0 and a natural language description of their dressing preference P, the objective is to generate a realistic try-on image I_K of the user dressed in a complete outfit, along with the corresponding garment images $\{G_i\}_{i=1}^K$ and their shopping links $\{L_i\}_{i=1}^K$. This task involves multiple challenges, including accurate interpretation of the multimodal input (I_0, P), retrieval of fine-grained garments G_i that align with the described style, composition of coherent outfits across different clothing components, and photorealistic synthesis of the try-on result I_K that preserves user identity. To address these challenges, our framework $\mathcal{T}$ adopts an iterative refinement strategy driven by vision-language feedback, progressively optimizing garment selection and try-on quality to ensure alignment with user intent and enhance the overall personalization experience.

3.2. Framework Overview

As illustrated in Fig. 2, StyleTailor consists of two core agents: *Designer* and *Consultant*, which collaboratively drive the personalized fashion styling workflow. The *Designer* module, denoted as $\mathcal{T}_1$, receives the user's reference image I_0 and their style preference description P. It then retrieves a set of garment images $\{G_i\}_{i=1}^K$ along with corresponding shopping links $\{L_i\}_{i=1}^K$. These garments are selected to align with the user's intended aesthetic and are ready for visualization or purchase. The *Consultant* module, denoted as $\mathcal{T}_2$, takes the initial user image I_0 and the retrieved garments $\{G_i\}_{i=1}^K$ as input and synthesizes a pho-

torealistic try-on image I_K, allowing the user to preview the recommended outfit on themselves. Although the *Designer* and *Consultant* operate independently, their integration forms a unified pipeline where the output of the *Designer* naturally serves as input to the *Consultant*. This modular yet cohesive design enables personalized, end-to-end fashion recommendation and virtual try-on in a structured and adaptive manner.

3.3. Designer

In fashion shopping, a good advisor translates abstract user preferences into specific garment selections. Inspired by this, our *Designer* agent serves as a virtual stylist that interprets user intent and retrieves suitable clothing items from real-world sources. To enhance accuracy and flexibility, we construct a pool of Design Experts, each comprising a *Style Interpreter* and a *Shopping Advisor*, coordinated by a sequential mechanism.

Style Interpreter. Users often express dressing intentions in vague or imaginative terms, rather than specific clothing attributes. The Style Interpreter leverages a vision-language model (VLM) to transform the user's input into a structured representation of garment components. Let I_0 denote the user's full-body reference image and P the textual description of their desired style. The VLM used by the i-th expert is denoted as V_i, and two prompt templates t_1 and t_2 are designed for first-time generation and feedback-refined iterations, respectively. For the first expert ($i = 1$), only the raw input (I_0, P) is provided, while for subsequent experts ($i > 1$), we also include a negative selection set $\{s_j\}_{j=1}^{i-1}$ representing previously rejected results. Each expert generates a set of category-description pairs $\{c_i, d_i\}_{i=1}^{K}$, where c_i refers to the name of a garment component (e.g., "dresses") and d_i its professional-level description:

$$\{c_i, d_i\}_{i=1}^{K} = \begin{cases} V_1(I_0, P, t_1) & \text{if } i = 1 \\ V_i(I_0, P, t_2, \{s_j\}_{j=1}^{i-1}) & \text{otherwise.} \end{cases} \quad (1)$$

The interpreter also produces concise summaries for each component, used by the downstream virtual try-on module.

Shopping Advisor. Given the structured garment descriptions, the Shopping Advisor performs web-based retrieval to simulate an online shopping process. Let d denote the textual description of a garment component. A custom search engine $\mathcal{G}$, built on Google Custom Search API, returns a set of candidate garment images $\{g_i\}_{i=1}^{M}$ and corresponding shopping links $\{l_i\}_{i=1}^{M}$:

$$\{g_i, l_i\}_{i=1}^{M} = \mathcal{G}(d), \quad (2)$$

Each candidate is evaluated using VQAScore, which measures the visual-semantic consistency between the retrieved image g_i and the intended description d. The best-matching image is selected as:

$$g_0 = \arg \max_{1 \le i \le M} \text{VQAScore}(g_i, d). \quad (3)$$

We introduce an *item-level* feedback to iteratively refine the searching results. If the highest score exceeds a predefined threshold τ, the result is accepted. Otherwise, we use a VLM to analyze mismatches between g_0 and d, extract undesirable features as negative cues, and rerun the search with negation terms. This process is repeated until a satisfactory result is found or a preset iteration limit is reached.

While the Shopping Advisor ensures alignment for individual items, it cannot guarantee global coherence across the outfit. We address this by implementing an *outfit-level* feedback to assess the quality of each full outfit using VQAScore. For each generated garment G_i and its corresponding component description d_i, we compute a raw alignment score $s_i = \text{VQAScore}(G_i, P)$, then normalize it by the garment-specific threshold τ_i to obtain a bounded score $s_i' = \min(s_i/\tau_i, 1)$. Assuming independence across components, we aggregate scores using the geometric average:

$$s_0 = \left(\prod_{i=1}^{K} s_i' \right)^{\frac{1}{K}}. \quad (4)$$

If the final score s_0 exceeds an acceptance threshold ω, the result is accepted. Otherwise, the next expert in the ranked pool is invoked, using the failed attempts as additional negative context. Experts are ordered by their VLM capability (e.g., leaderboard performance), enabling a greedy strategy that balances accuracy and computational efficiency. This multi-level feedback ensures that the generated garments are not only locally relevant but also globally consistent with the user's style intent.

3.4. Consultant

After shopping, trying on clothes is a crucial step in determining whether the selected garments truly align with the user's needs and expectations. The *Consultant* module is designed to simulate this virtual try-on process, enabling systematic visual evaluation of the outfit generated by the *Designer*.

Progressive Try-on Generation. Let I_0 denote the original user image and $\{G_i\}_{i=1}^{K}$ the set of selected garment images. The *Consultant* module $\mathcal{F}$ generates the final try-on result I_K by sequentially replacing each garment through K independent sub-processes $\{\mathcal{F}_i\}_{i=1}^{K}$, each responsible for updating one clothing component:

$$\mathcal{F} = \mathcal{F}_1 \circ \mathcal{F}_2 \circ \cdots \circ \mathcal{F}_K. \quad (5)$$

Each sub-process $\mathcal{F}_i$ receives the image I_{i-1} from the previous step and produces the updated output I_i:

$$I_i = \mathcal{F}_i(I_{i-1}) \quad (1 \le i \le K). \quad (6)$$

To reduce visual interference during editing, garments are sorted in descending order of region size—larger components (e.g., outerwear) are replaced first, followed by smaller ones (e.g., accessories).

VLM-Guided Visual Refinement. Each sub-process $\mathcal{F}_i$ is implemented using the image-editing model FLUX.1.Kontext(Labs et al. 2025), denoted $\mathcal{K}$. The model takes as input the current user image I_{i-1}, the corresponding garment image G_i, and a textual prompt z_i summarizing the desired appearance, which is produced by the Style Interpreter of the *Designer*. Let $\mathcal{P}(z_i)$ represent a prompt formatting function. We concatenate I_{i-1} and G_i, and generate l try-on candidates:

$$\{O_i\}_{i=1}^{l} = \mathcal{K}(\text{concat}(I_{i-1}, G_i), \mathcal{P}(z_i)). \quad (7)$$

To select the most accurate result, we apply OpenPose (Cao et al. 2021, 2017; Simon et al. 2017; Wei et al.

Figure 3: Qualitative comparison between our method and the baseline under two conditions: (1) The same user image with different descriptions; (2) The same description with different user images. The visualization results demonstrate both the personalized design capacity of our approach and its superior performance in comparison with the baseline method. The red text indicates the apparently inappropriate garment retrieval by the baseline.

2016) and HumanParsing (Li et al. 2019) models to identify the region corresponding to the target category c_i, and extract that region from each generated image. We then compute the CLIPScore (Hessel et al. 2022) between the masked try-on image and the original garment, and choose the best-matching candidate:

$$O_0 = \arg\max_{1 \le i \le l} \text{CLIPScore}(\text{mask}_{c_i}(O_i), G_i). \quad (8)$$

If the score of O_0 exceeds a predefined threshold σ, the result is accepted. Otherwise, the current candidate and garment are passed into a VLM to diagnose visual inconsistencies. These differences are converted into a negative prompt and injected into $\mathcal{K}$ for regeneration. This *try-on-level* feedback process iterates until the output meets the quality threshold or the maximum number of attempts is reached. The final image I_K is obtained by applying this process to all garments as defined in Eq. (5).

4. Experiments

4.1. Settings

Dataset. To construct the evaluation dataset, we require both input images and text prompts. To balance inclusiveness and testing efficiency, we set the dataset size to 128. The input images are sourced from the LookBook dataset (Yoo et al. 2016), which offers diverse, high-quality photos of individuals wearing various garments. The text prompts are generated using an LLM with a few-shot setup (Brown et al. 2020), encompassing both specific and abstract requests involving suits, dresses, shoes, and accessories. Among the 128 samples, the first 64 are collected under clean backgrounds, while the remaining 64 feature diverse poses and outdoor environments to improve e-commerce applicability. Each 64-sample subset includes 32 male and 32 female models and can be further partitioned by gender based on the criteria described below to enhance diversity.

- **Face Status**: classified as either visible or hidden, with a balanced 1:1 ratio for each gender.
- **Body Status**: indicates whether the full body is shown. Full-body and half-body images follow a 3:1 ratio, consistent across face-status categories.

Baseline. We use a version of the workflow without any negative feedback mechanisms as our baseline. This setup preserves the full functionality of the agent while isolating the effect of the proposed feedback strategies.

Figure 4: Visualizations of diverse user images and style descriptions, along with the corresponding outputs produced by our StyleTailor. These examples demonstrate StyleTailor's ability to effectively handle various user appearances and styling preferences, highlighting its robustness and adaptability to complex, real-world input scenarios.

4.2. Evaluation Metrics

As shown in the *Critic* module of Fig. 2, we employ four metrics to comprehensively evaluate the generated image I_K.

Style Consistency. To assess the alignment between the generated image I_K and user preferences P, we first feed the user's original image I_0 into the VLM V to extract human-related attributes excluding garments, then convert them into textual descriptions. These descriptions are concatenated with the initial preferences and evaluated against the generated image through $VQAScore(I_K, (V(I_0)+P))$.

Visual Quality. We apply IQAScore (Chen and Mo 2022; Chen et al. 2024; Wu et al. 2024) to assess the visual fidelity of the generated images.

Face Similarity. Since facial appearance is essential for preserving personal identity, we use a pre-trained InsightFace model (Ren et al. 2023) to extract facial features from I_0 and I_K, and compute their cosine similarity as the face similarity metric.

VLM Artist. Beyond measuring preference alignment and generative quality, we introduce a VLM-based evaluation agent—referred to as the *VLM Artist*—to provide a comprehensive aesthetic assessment of the final garment synthesis results. The VLM Artist evaluates four aspects of each image (described below), producing both a concise explanation and an integer score from 1 to 10.

- **Design Score**: evaluates the aesthetics of individual garment pieces, including *cut* (silhouette, tailoring) and *elements* (decorative details).
- **Fitness Score**: measures how well the garments fit the wearer's physical attributes (e.g., body shape, proportions).
- **Coherence Score**: assesses the stylistic compatibility and harmony among garments in the ensemble.
- **Mood Score**: evaluates the overall mood, stylistic identity, and visual impact of the final look.

Scores are assigned based on the criteria detailed in the Appendix, and their mean forms the final VLM Artist evaluation. As shown in Tab. 1, our VLM Artist exhibits strong alignment with human preferences.

This multi-dimensional evaluation suite offers a holistic assessment of both the quality of the generated outputs and their consistency with user intent.

Models	Style Consistency	Visual Quality	Face Similarity	VLM Artist	VLM Artist-Human Alignment
Baseline (Clean Background)	0.650	0.699	0.362	7.35	7.09
w/o *Item-level Negative Feedback* (Clean Background)	0.697	**0.767**	0.484	8.41	8.56
w/o *Outfit-level Negative Feedback* (Clean Background)	0.688	0.758	<u>0.532</u>	8.16	8.18
w/o *Try-on-level Negative Feedback* (Clean Background)	0.781	0.708	0.351	8.22	8.06
Ours (Clean Background)	**0.906**	<u>0.764</u>	**0.544**	**8.60**	**8.83**
Ours (Diverse Background)	<u>0.852</u>	0.726	0.483	<u>8.53</u>	<u>8.67</u>

Table 1: Quantitative comparison. The VLM Artist-Human Alignment presents the results after we trained human experts to evaluate the outputs following the VLM artist criteria. **Bold** indicates the **best performance**, <u>underline</u> indicates the <u>second-best performance</u>.

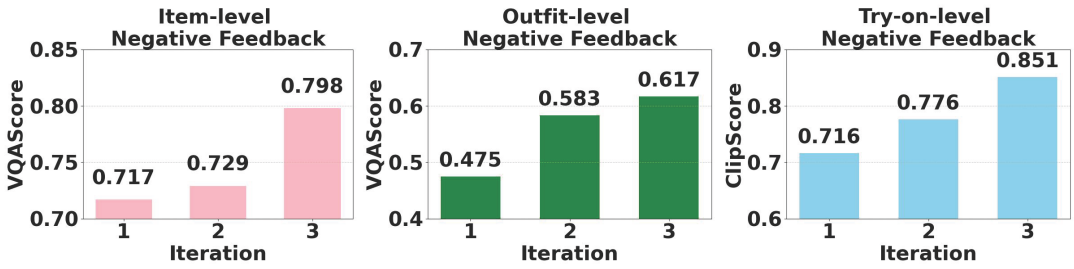

Figure 5: The score changes *w.r.t.* iterations activated by corresponding negative feedback. The consistent improvements demonstrate the effectiveness of our hierarchical negative feedback mechanism.

4.3. Comparison with State-of-the-art Methods

Qualitative Comparison. As shown in Fig. 3, StyleTailor accurately aligns with varied textual specifications for the same individual while maintaining personalization across different users. Comparative analysis reveals that StyleTailor outperforms the baseline, which struggles with specific garment alignment (*e.g.*, shoes and pants). These results demonstrate that our item- and outfit-level negative feedback mechanisms are vital for precise retrieval. Furthermore, unlike the baseline—which produces inconsistent colors and artifacts in facial features or posture—our approach ensures visual and anatomical integrity.

Quantitative Comparison. As shown in Tab. 1, our StyleTailor consistently outperforms the baseline across all metrics. Improved *Style Consistency* confirms that the Designer module's multi-level negative feedback effectively strengthens text-to-image alignment. Simultaneously, gains in *Visual Quality* and *Face Similarity* demonstrate that the *Consultant* module's try-on feedback successfully filters implausible results. Evaluations from the *VLM Artist* further indicate superior alignment with general aesthetic preferences. Finally, the upward score trajectories in Fig. 5 provide procedural evidence of the negative feedback mechanism's iterative effectiveness.

4.4. Diversity

This section illustrates the adaptability of our method across diverse scenarios. We tested our approach using individuals of varying genders, body types, and celebrity resemblances, paired with stylistic themes such as cyberpunk, pirate, vacation, and outdoor military. As shown in Fig. 4, StyleTailor delivers consistently high-quality, adaptive results across all subject and style variations.

4.5. Ablation Study

We perform ablation experiments by removing each type of negative feedback from the full configuration (Tab. 1) and evaluate their individual contributions using the proposed metrics.

Effect of Item-level Negative Feedback. Removing item-level feedback causes substantial drops in *Style Consistency* (0.906 $\rightarrow$ 0.697) and *VLM Artist* score (8.60 $\rightarrow$ 8.41), indicating its key role in aligning garment retrieval with user style preferences.

Effect of Outfit-level Negative Feedback. Eliminating outfit-level feedback similarly reduces *Style Consistency* (0.906 $\rightarrow$ 0.688) and *VLM Artist* score (8.60 $\rightarrow$ 8.16), showing its importance for maintaining global outfit coherence beyond item-level refinement.

Effect of Try-on-level Negative Feedback. Without try-on-level feedback, *Visual Quality* (0.764 $\rightarrow$ 0.708) and *Face Similarity* (0.544 $\rightarrow$ 0.351) drop significantly, confirming that high-level supervision is vital for realism and identity preservation. While item- and outfit-level mechanisms focus on retrieval and style, this feedback directly guides the generative process to ensure faithful, stable results.

5. Conclusion

This paper presents **StyleTailor**, the first agentic framework unifying fashion design, shopping recommendation, and virtual try-on. We introduce a multi-level negative feedback mechanism that strengthens agent reasoning and enables iterative refinement. To evaluate performance, we propose metrics tailored to personalized fashion styling. Extensive experiments show StyleTailor holds strong potential for real-world applications, promising to inspire future advancements in intelligent, user-centric fashion systems.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62502317).

References

Baldrati, A.; Morelli, D.; Cartella, G.; Cornia, M.; Bertini, M.; and Cucchiara, R. 2023. Multimodal Garment Designer: Human-Centric Latent Diffusion Models for Fashion Image Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2021. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1): 172–186.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. arXiv:1611.08050.

Chen, C.; and Mo, J. 2022. IQA-PyTorch: PyTorch Toolbox for Image Quality Assessment. [Online]. Available: https://github.com/chaofengc/IQA-PyTorch.

Chen, C.; Mo, J.; Hou, J.; Wu, H.; Liao, L.; Sun, W.; Yan, Q.; and Lin, W. 2024. TOPIQ: A Top-Down Approach From Semantics to Distortions for Image Quality Assessment. *IEEE Transactions on Image Processing*, 33: 2404–2418.

Durante, Z.; Huang, Q.; Wake, N.; Gong, R.; Park, J. S.; Sarkar, B.; Taori, R.; Noda, Y.; Terzopoulos, D.; Choi, Y.; Ikeuchi, K.; Vo, H.; Fei-Fei, L.; and Gao, J. 2024. Agent AI: Surveying the Horizons of Multimodal Interaction. arXiv:2401.03568.

Han, X.; Hu, X.; Huang, W.; and Scott, M. R. 2019. ClothFlow: A Flow-Based Model for Clothed Person Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Han, X.; Wu, Z.; Wu, Z.; Yu, R.; and Davis, L. S. 2018. VITON: An Image-based Virtual Try-on Network. arXiv:1711.08447.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2022. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. arXiv:2104.08718.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2): 1–55.

Ji, Z.; Yu, T.; Xu, Y.; Lee, N.; Ishii, E.; and Fung, P. 2023. Towards Mitigating Hallucination in Large Language Models via Self-Reflection. arXiv:2310.06271.

Jiang, Y.; Yang, S.; Qiu, H.; Wu, W.; Loy, C. C.; and Liu, Z. 2022. Text2Human: Text-Driven Controllable Human Image Generation. *ACM Transactions on Graphics (TOG)*, 41(4): 1–11.

Koh, W.; Yoon, J.; Lee, M.; Song, Y.; Cho, J.; Kang, J.; Kim, T.; Yun, S.-Y.; Yu, Y.; and Lee, B. 2025. C^2: Scalable Auto-Feedback for LLM-based Chart Generation. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 4525–4566. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.

Kubo, S.; Iwasawa, Y.; Suzuki, M.; and Matsuo, Y. 2019. UVTON: UV Mapping to Consider the 3D Structure of a Human in Image-Based Virtual Try-On Network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

Labs, B. F.; Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; Kulal, S.; Lacey, K.; Levi, Y.; Li, C.; Lorenz, D.; Müller, J.; Podell, D.; Rombach, R.; Saini, H.; Sauer, A.; and Smith, L. 2025. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. arXiv:2506.15742.

Li, J.; Qiu, X.; Xu, L.; Guo, L.; Qu, D.; Long, T.; Fan, C.; and Li, M. 2025. UniF2ace: Fine-grained Face Understanding and Generation with Unified Multimodal Models. *arXiv preprint arXiv:2503.08120*.

Li, M.; Zhou, P.; Liu, J.-W.; Keppo, J.; Lin, M.; Yan, S.; and Xu, X. 2024. Instant3d: instant text-to-3d generation. *IJCV*.

Li, P.; Xu, Y.; Wei, Y.; and Yang, Y. 2019. Self-Correction for Human Parsing. arXiv:1910.09777.

Lin, Z.; Pathak, D.; Li, B.; Li, J.; Xia, X.; Neubig, G.; Zhang, P.; and Ramanan, D. 2024. Evaluating Text-to-Visual Generation with Image-to-Text Generation. *arXiv preprint arXiv:2404.01291*.

Ling, C.; Zhao, X.; Lu, J.; Deng, C.; Zheng, C.; Wang, J.; Chowdhury, T.; Li, Y.; Cui, H.; Zhang, X.; Zhao, T.; Panalkar, A.; Mehta, D.; Pasquali, S.; Cheng, W.; Wang, H.; Liu, Y.; Chen, Z.; Chen, H.; White, C.; Gu, Q.; Pei, J.; Yang, C.; and Zhao, L. 2024. Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey. arXiv:2305.18703.

Liu, S.; Li, J.; Zhao, G.; Zhang, Y.; Meng, X.; Yu, F. R.; Ji, X.; and Li, M. 2025. EventGPT: Event Stream Understanding with Multimodal Large Language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 29139–29149.

Liu, Y.; An, J.; Zhang, W.; Li, M.; Wu, D.; Gu, J.; Lin, Z.; and Wang, W. 2024. RealEra: Semantic-level Concept Erasure via Neighbor-Concept Mining. *arXiv preprint arXiv:2410.09140*.

Miao, C.; Chang, T.; Wu, M.; Xu, H.; Li, C.; Li, M.; and Wang, X. 2025. FedVLA: Federated Vision-Language-Action Learning with Dual Gating Mixture-of-Experts for Robotic Manipulation. *arXiv preprint arXiv:2508.02190*.

Neuberger, A.; Borenstein, E.; Hilleli, B.; Oks, E.; and Alpert, S. 2020. Image Based Virtual Try-On Network From Unpaired Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pang, W.; Lin, K. Q.; Jian, X.; He, X.; and Torr, P. 2025. Paper2Poster: Towards Multimodal Poster Automation from Scientific Papers. arXiv:2505.21497.

Qiao, S.; Zhang, N.; Fang, R.; Luo, Y.; Zhou, W.; Jiang, Y. E.; Lv, C.; and Chen, H. 2024. AutoAct: Automatic Agent Learning from Scratch for QA via Self-Planning. arXiv:2401.05268.

Qwen; :; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Raj, A.; Sangkloy, P.; Chang, H.; Lu, J.; Ceylan, D.; and Hays, J. 2018. SwapNet: Garment Transfer in Single View Images. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Ren, X.; Lattas, A.; Gecer, B.; Deng, J.; Ma, C.; and Yang, X. 2023. Facial Geometric Detail Recovery via Implicit Representation. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Seo, M.; Baek, J.; Lee, S.; and Hwang, S. J. 2025. Paper2Code: Automating Code Generation from Scientific Papers in Machine Learning. arXiv:2504.17192.

Shen, F.; Jiang, X.; He, X.; Ye, H.; Wang, C.; Du, X.; Li, Z.; and Tang, J. 2025a. Imagdressing-v1: Customizable virtual dressing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6795–6804.

Shen, F.; Yu, J.; Wang, C.; Jiang, X.; Du, X.; and Tang, J. 2025b. IMAGGarment-1: Fine-Grained Garment Generation for Controllable Fashion Design. *arXiv preprint arXiv:2504.13176*.

Shi, Y.; Yan, W.; Xu, G.; Li, Y.; Chen, Y.; Li, Z.; Yu, F. R.; Li, M.; and Yeo, S. Y. 2025. Pvchat: Personalized video chat with one-shot learning. *arXiv preprint arXiv:2503.17069*.

Simon, T.; Joo, H.; Matthews, I.; and Sheikh, Y. 2017. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. arXiv:1704.07809.

Su, Z.; Qiu, X.; Xu, H.; Jiang, T.; Zhuang, J.; Yuan, C.; Li, M.; He, S.; and Yu, F. R. 2025. Safe-Sora: Safe Text-to-Video Generation via Graphical Watermarking. *arXiv preprint arXiv:2505.12667*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Wang, X.; Li, B.; Song, Y.; Xu, F. F.; Tang, X.; Zhuge, M.; Pan, J.; Song, Y.; Li, B.; Singh, J.; Tran, H. H.; Li, F.; Ma, R.; Zheng, M.; Qian, B.; Shao, Y.; Muennighoff, N.; Zhang, Y.; Hui, B.; Lin, J.; Brennan, R.; Peng, H.; Ji, H.; and Neubig, G. 2025. OpenHands: An Open Platform for AI Software Developers as Generalist Agents. arXiv:2407.16741.

Wei, S.-E.; Ramakrishna, V.; Kanade, T.; and Sheikh, Y. 2016. Convolutional Pose Machines. arXiv:1602.00134.

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Li, C.; Liao, L.; Wang, A.; Zhang, E.; Sun, W.; Yan, Q.; Min, X.; Zhai, G.; and Lin, W. 2024. Q-Align: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels. *International Conference on Machine Learning (ICML)*. Equal Contribution by Wu, Haoning and Zhang, Zicheng. Project Lead by Wu, Haoning. Corresponding Authors: Zhai, Guangtai and Lin, Weisi.

Xu, B.; Yang, A.; Lin, J.; Wang, Q.; Zhou, C.; Zhang, Y.; and Mao, Z. 2025a. ExpertPrompting: Instructing Large Language Models to be Distinguished Experts. arXiv:2305.14688.

Xu, H.; Yu, C.; Xiao, F.; Xing, J.; Ci, H.; Chen, W.; Wang, F.; and Li, M. 2025b. Cyc3D: Fine-grained Controllable 3D Generation via Cycle Consistency Regularization. *arXiv preprint arXiv:2504.14975*.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*.

Yoo, D.; Kim, N.; Park, S.; Paek, A. S.; and Kweon, I. S. 2016. Pixel-level domain transfer. In *European conference on computer vision*, 517–532. Springer.

Yue, L.; Somasekharan, N.; Cao, Y.; and Pan, S. 2025. Foam-Agent: Towards Automated Intelligent CFD Workflows. arXiv:2505.04997.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models.

Zhao, F.; Li, M.; Xu, L.; Jiang, W.; Gao, J.; and Yan, D. 2025. FaVChat: Unlocking Fine-Grained Facial Video Understanding with Multimodal Large Language Models. *arXiv preprint arXiv:2503.09158*.

Zheng, H.; Guan, X.; Kong, H.; Zheng, J.; Zhou, W.; Lin, H.; Lu, Y.; He, B.; Han, X.; and Sun, L. 2025. PPTAgent: Generating and Evaluating Presentations Beyond Text-to-Slides. arXiv:2501.03936.

Zhu, L.; Yang, D.; Zhu, T.; Reda, F.; Chan, W.; Saharia, C.; Norouzi, M.; and Kemelmacher-Shlizerman, I. 2023. TryOnDiffusion: A Tale of Two UNets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4606–4615.

Zi-Yi, C.; Fan-Kai, X.; Meng, W.; Yang, Y.; Miao, L.; Zong-Guo, W.; Sheng, M.; and Yan-Gang, W. 2023. MatChat: A large language model and application service platform for materials science. *Chinese Physics B*, 32(11): 118104.