Stylo - Semantic Transformation of Your Looks and Outfits using Vision Foundation Models for Virtual Try-On

Felix Förster Andreas Weber Lars Schneider

Abstract

Virtual Try-On (VITON) has the potential to revolutionize the retail and fashion industries by providing consumers with immersive, personalized experiences that bridge the gap between physical and online shopping. With this work, we aim to demonstrate the capabilities of publicly available state-of-the-art vision foundation models and how they can be effectively integrated into a pipeline for semantic image editing. Specifically, our pipeline is capable to generate custom garments and background images which can used to modify a given subject image. For garment fitting, we utilize StableVITON [7] from the KAIST Research Group. The whole pipeline is provided as a user-friendly web-application that is intuitive to use.

1. Introduction

Given an image of a garment and a subject, Virtual Try-On (VITON) aims to generate a realistic, personalized visualization of the garment on the subject, preserving the garment's distinctive features and ensuring a natural, realistic appearance. Inspired by this idea, we developed a comprehensive semantic image editing pipeline that not only enables VITON but also allows users to apply versatile modifications to original images, such as background replacement, lighting adaptation, and garment generation, ultimately ensuring a seamless and immersive user experience.

In this project, we utilize StableVITON [7], which is a follow-up research of the models proposed in VITON-HD [1] and HR-VITON [10]. In addition, it is publicly available, including a codebase, unlike the newest research from the same research group, PromptDresser [8]. There are many other VITON models, but they are either closed source or outdated [12][20][21] [4] [18].

Project Goals: We defined the following goals in our initial presentation:

1. Integration of specific foundation models for semantic image editing into a pipeline, which enables a user to modify an image of a subject. This includes but is not limited to background replacement, light adaptation, and garment modification.
2. Development of a user-friendly user interface to control the image editing pipeline
3. Qualitative analysis of pipeline results by performing a user study.

2. Foundation Models

We leverage several foundation models to build our pipeline for semantic image editing. These are briefly described in the following.

2.1. Contrastive Language-Image Pretraining

Contrastive Language-Image Pretraining (CLIP) is a model developed by OpenAI [13] that jointly learns visual and textual representations through contrastive learning on a large dataset of image–text pairs. Instead of generating images or captions directly, CLIP maps both images and text into a shared embedding space, enabling it to measure how well an image matches a given text description. This makes it highly effective for tasks like zero-shot image classification, prompt-guided image generation, and evaluating text-to-image models.

2.2. Stable Diffusion 3

Stable Diffusion 3 (SD3) is an advanced text-to-image generative model developed by Stability.AI [3], designed to create high-quality, photorealistic images from natural language prompts. Compared to its predecessors, SD1 [16] and SD2 [17], SD3 offers improved image fidelity, better understanding of complex prompts, and enhanced control over composition and style. Its Multimodal Diffusion Transformer consists of two billion parameters that iteratively refine the latent representation of the generated image based on a given prompt. The architecture of SD3 also makes use of different variants of CLIP, i.e., ViT-G/14 and ViT-L/14 [2], to guide and evaluate how well the generated image matches the text prompt by providing a shared semantic embedding space. Additionally, Google's text-to-text transfer transformer T5-XXL [14] is integrated to refine or expand

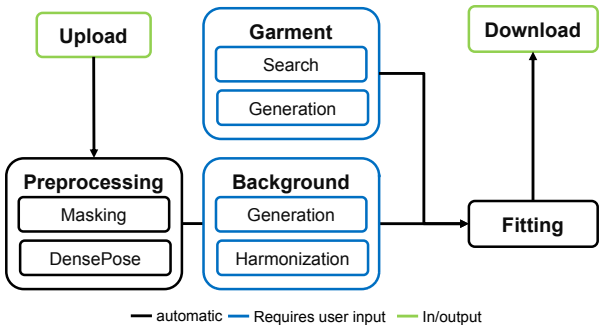


Figure 1. Building block architecture. For intermediate example results, refer to figure 2.

text prompts, improving prompt understanding and expressiveness before image generation. Overall, the SD3 architecture requires around 24 GB of VRAM for inference.

2.3. Segmentation Model

Segment Anything Model 2 (SAM2) [15] is a foundation model for promptable visual segmentation in images and videos. It is the successor to the original SAM model [9] and is one of the most powerful foundation models regarding interactive segmentation tasks with high inference speed and state-of-the-art zero-shot performance. For better promptability, we use a modification called LangSam [11] which allows us to use text prompts for better customization.

2.4. StableVITON

StableVITON is a latent-diffusion model especially developed for Virtual Try-On. It is able to combine clothing parts and a person's image while maintaining realism and characteristics. This is done by learning the correspondences between garments and the human body shape.

To be able to achieve an accurate fitting, StableVITON requires several inputs, such as a garment image, a person image, an agnostic image, an agnostic mask, and the person's DensePose.

3. Method

This section describes the overall pipeline architecture as well as its individual components.

3.1. Overall Pipeline Architecture

Backend: The architecture of the pipeline is based on a Python WebSocket that manages the single components and states. We designed each block in the pipeline to have the same interface for increased flexibility. This interface includes standardized initialization, loading/unloading, and model calls. The intermediate state is then saved and used when required. Figure 1 shows all the building blocks and their relationships.

Frontend: The Frontend is built using the Vue Framework [19]. It exists to enable an intuitive experience that does not need technical knowledge to be understood. The UI uses a step-by-step approach to navigate through the pipeline, giving the user instructions on what to do. It communicates in real-time via WebSockets to ensure as little latency as possible is added by the frontend.

3.2. Garment and Background Generation

We utilize SD3 to generate novel garment and background images based on user-provided prompts. Although SD3 is not specifically fine-tuned for fashion product generation, we achieved notable results by prompt engineering. For garment generation, we apply the following template:

Prompt Template: Garment Generation

```
<user-defined garment prompt> + "
neatly hung in front of a white wall,
isolated product shot, studio lighting,
realistic texture, garment fully
visible, photo-realistic, entire garment
visible, garment centered, size m"
```

This ensures that the generated images clearly display the garment without folds or distortions, and most often depict the garment hanging on a coat hook. Examples of the generated garments based on given prompts can be found in the supplementary material, Figure 6. Similarly, we manually extended the user-specified prompt to improve the results for the background generation.

The following prompt template ensures the creation of photo-realistic, bright images that match the perspective of the model images:

Prompt Template: Background Generation

```
<user prompt> + " photo-realistic,
realistic lighting, highly detailed
texture, viewed from head height, shot
with a DSLR camera, soft shadows"
```

Examples of the generated background images based on given prompts can be found in the supplementary material, Figure 7.

3.3. Masking

Masks are needed for different steps in our pipeline. A full-person mask is required to separate the original background from the person, and then to replace it afterward. Also, an agnostic mask is needed that isolates all the body parts that StableVITON generates. In our case, this means the upper body without hands, but including arms and neck. To achieve this, we use LangSAM [11] to generate isolated masks for (hands, hair, shoes ...), which then get sub-

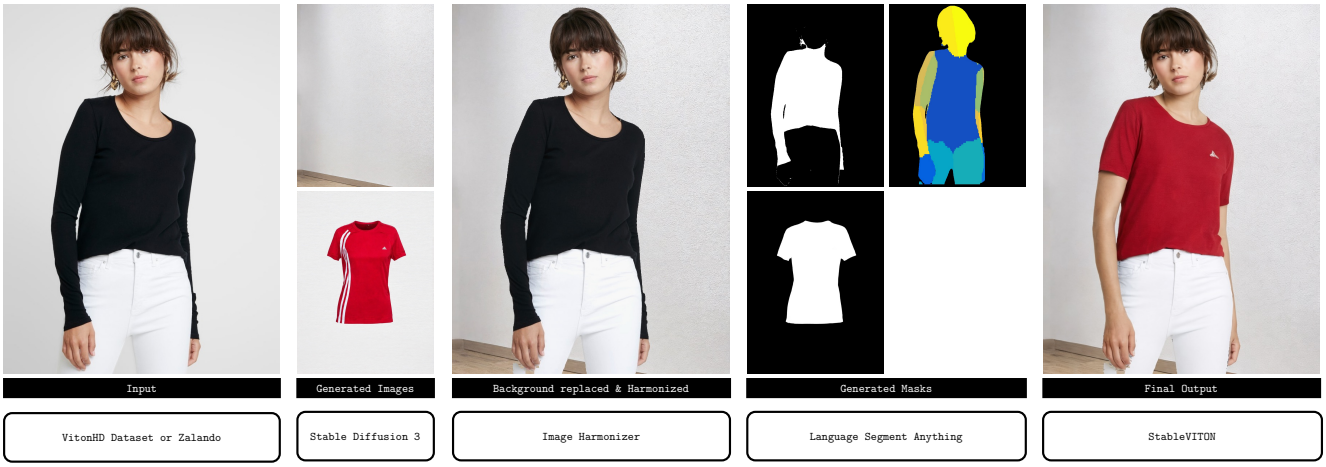


Figure 2. Visualization of pipeline input, intermediate, and final results. The input consists of the subject image, a background, and a garment. Background and garment can be generated by utilizing SD3. The subject image with a replaced background and adapted lighting is an intermediate result. Garment masks (subject and garment image) and dense pose (subject image) serve as additional input for StableVITON to generate the final output.

tracted from the full person mask. Since the resulting mask has grainy artifacts, we use dilation and erosion to get a smoothed-out mask and remove all remaining artifacts. An example of this problem can be found in the supplementary material, Figure 5.

3.4. Garment Fitting

For garment fitting, we take the subject image with its subject mask, agnostic mask, the garment image and garment mask. Once all inputs are prepared, it is straightforward to generate the fitting of the garment with StableVITON. We choose 50 diffusion steps at an image resolution of 1024×768, which offered a good tradeoff between speed and quality.

3.5. Optimization

Image Search: To reduce the total processing time of our pipeline and to make it more user-friendly, we enabled the possibility to search for existing garments that match a given prompt. Therefore, we utilize the CLIP foundation model and pre-computed over 2000 image embeddings from the VITON-HD dataset. When a user provides a garment prompt, we generate a text embedding that is compared with the pre-computed image embeddings. Then the top three matches are returned. The CLIP score is calculated as the cosine similarity between the text embedding vector $\mathbf{T}$ and the image embedding vector $\mathbf{I}$, obtained from CLIP's respective encoders, see Equation 1.

$$\cos(\theta) = \frac{\mathbf{T} \cdot \mathbf{I}}{\|\mathbf{T}\| \cdot \|\mathbf{I}\|} \tag{1}$$

Resource Management and Pre-loading: Limited resources and the size of the model make it difficult to keep loading times low. To guarantee the fastest inference possible, the backend is always moving required models on and off the GPU/CPU. This allows us to run our architecture on workstations with 24GB of VRAM and 32GB of RAM. To achieve this, we make sure that only the SD3 or the StableVITON model is loaded at a time on the GPU. In addition, we keep them loaded in the RAM to avoid enormous loading times from the hard drive.

4. Results

This section presents and qualitatively analyzes selected results from our pipeline. In addition, we conducted a user study to gather human feedback on the perceived quality of the pipeline outputs.

4.1. Qualitative Analysis

For example results please refer to the appendix (section H). We identified the following key strengths and weaknesses.

Strengths:

(+) If LangSam masks are properly generated, StableViton is able to restore covered body parts, e.g., skin and color tone of arms, neck.

(+) The generated garments and background look realistic and align well with the given prompt.

Weaknesses:

(–) LangSam mask generation is highly error-prone. StableViton modifies all parts not masked in the image, leading to especially poor results in the face details of a subject or hands if not properly masked.

(–) Details are not preserved. Logos or special textures on garments are not preserved.

(–) Original colors are not preserved. Garment replacement

	Fidelity	Realism	Aesthetics	Sharpness
Generated Garments	4.27	4.00	4.12	4.00
Fitted Garments	3.76	3.04	3.64	4.19

Table 1. Average user study scores for garment generation and garment fitting, evaluated on realism, fidelity, aesthetics, and sharpness.

from black to white often seems to lead to gray garments in the resulting image.

We did not apply any additional finetuning to the foundation models used in this project, in order to evaluate their out-of-the-box performance. Given that garment generation and fitting are highly specialized tasks, finetuning these models would likely lead to substantial improvements.

4.2. User Study

While automated metrics, such as the Fréchet Inception Distance [6] or CLIP Score [5] provide useful quantitative benchmarks, they fail to capture the nuanced ways in which humans perceive image realism, prompt alignment, and aesthetic quality. For this reason, we decided to evaluate our pipeline results by conducting a conventional user study. Thereby, we directly assess how convincing, engaging, and semantically accurate the generated images appear to human perception, ensuring that the pipeline truly meets user expectations. The following four metrics were chosen to evaluate the images on a scale from one to five. A detailed description of each category can be found in the supplementary material, subsection G.1.

1. **Fidelity** – Does the image match the given input?
2. **Realism** – How realistic does the image appear?
3. **Aesthetics** – How visually pleasing is the image?
4. **Sharpness** – How sharp and detailed is the image?

   Scale – Worst: 1 and Best: 5

Table 1 summarizes the average score for each metric over a total number of 26 user reviews. The detailed statistics can be found in the supplementary material, subsection G.2.

5. Retrospective

Table 2 summarizes the defined target scope at the beginning of this project. We initially set seven different target goals, from which six have been fulfilled. We deprioritized the person selection in images and focused instead on a custom web-based user interface. Additionally, we implemented the possibility to search for existing garments to save some inference time and make the whole process more user-friendly.

6. Conclusion

In this work, we demonstrate how to integrate different publicly available foundation models into a semantic image editing pipeline. While it can be seen as a prototype, the results show that the underlying technologies can be used to build powerful products that possess the ability to impact various businesses, ranging from online shopping to advertising to fashion design.

Component	Target	Achieved
Pipeline Blocks	Background Manipulation	✓
	Fix Lighting	✓ Harmonization
	Modification of Garment	✓ Generation of Custom Garments and Backgrounds
	Person Selection	✗
	Fit Garment	✓
Evaluation	User Study	✓
User Interface	Basic UI (e.g., Gradio)	✓ Custom Web-based UI
Additional Functionalities	-	✓ Performance Optimization, Garment Search

Table 2. Overview of pipeline components, corresponding targets, and achieved status.

References

[1] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. VITON-HD: high-resolution virtual try-on via misalignment-aware normalization. *CoRR*, abs/2103.16874, 2021. 1

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 1

[3] Patrick Esser et al. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis, 2024. 1

[4] Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 7599–7607. ACM, 2023. 1

[5] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning, 2022. 4

[6] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a Better Evaluation Metric for Image Generation, 2024. 4

[7] Jeongho Kim, Gyojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. StableVITON: Learning Semantic Correspondence with Latent Diffusion Model for Virtual Try-On, 2023. 1

[8] Jeongho Kim, Hoiyeong Jin, Sunghyun Park, and Jaegul Choo. Promptdresser: Improving the quality and controllability of virtual try-on via generative textual prompt and prompt-aware mask, 2024. 1

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 2

[10] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions, 2022. 1

[11] Luca Medeiros. https://github.com/luca-medeiros/lang-segment-anything, 2025. 2

[12] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on, 2023. 1

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, 2021. 1

[14] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, 2023. 1

[15] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. 2

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and BjÃ¶rn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[18] Ke Sun, Jian Cao, Qi Wang, Linrui Tian, Xindi Zhang, Lian Zhuo, Bang Zhang, Liefeng Bo, Wenbo Zhou, Weiming Zhang, and Daiheng Gao. Outfitanyone: Ultra-high quality virtual try-on for any clothing and any person, 2024. 1

[19] VueJS. https://github.com/vuejs/, 2025. 2

[20] Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning, 2023. 1

[21] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models, 2022. 1

Stylo - Semantic Transformation of Your Looks and Outfits using Vision Foundation Models for Virtual Try-On

Supplementary Material

G. User Study

To quantify our results, we conducted a user study in which 26 participants rated different metrics of generated garments and the final results of garment fitting.

G.1. Metric Categorization

Garment Generation - Fidelity: Does the generated images match the given prompts?

1. *Very Poor*: The garment does not match the prompt at all and is entirely incorrect.
2. *Poor*: The garment loosely resembles the prompt but lacks most key attributes.
3. *Moderate*: The garment generally matches the prompt, but has notable inaccuracies or missing details.
4. *Good*: The garment matches the prompt well with only minor discrepancies.
5. *Excellent*: The garment precisely and fully matches the prompt in all important aspects.

Garment Generation - Realism: How realistic or natural are the images?

1. *Very Poor*: The garment looks completely artificial, distorted, or obviously fake.
2. *Poor*: The garment is clearly unrealistic, with major artifacts or unnatural textures.
3. *Moderate*: The garment appears somewhat realistic, but still contains noticeable artifacts or unnatural elements.
4. *Good*: The garment looks mostly realistic, with only minor flaws or slight synthetic hints.
5. *Excellent*: The garment is highly realistic and natural, nearly indistinguishable from a real garment.

Garment Generation - Aesthetics: How visually pleasing are the images?

1. *Very Poor*: The garment is visually unappealing and unattractive.
2. *Poor*: The garment has weak aesthetic appeal and looks awkward or poorly designed.
3. *Moderate*: The garment is somewhat visually pleasing but lacks harmony or refinement.
4. *Good*: The garment is visually appealing and well-designed, with minor aesthetic flaws.
5. *Excellent*: The garment is highly attractive, stylish, and beautifully designed.

Garment Generation - Sharpness: How sharp and detailed are the images?

1. *Very Poor*: The garment is extremely blurry, lacking any clear details, and has severe artifacts.
2. *Poor*: The garment is mostly blurry, with very few sharp features and noticeable artifacts.
3. *Moderate*: The garment has some sharp areas, but overall appears soft or partially blurred, with moderate artifacts.
4. *Good*: The garment is generally sharp and detailed, with only minor softness or small artifacts.
5. *Excellent*: The garment is very sharp, highly detailed, and free from noticeable artifacts.

Garment Fitting - Fidelity: Does the fitted garment match the given garment?

1. *Very Poor*: The fitted garment does not match the given garment at all; it is entirely incorrect and unrecognizable.
2. *Poor*: The fitted garment only loosely resembles the given garment; most key features are missing or incorrect.
3. *Fair*: The fitted garment captures the general idea of the given garment, but has several significant inaccuracies or missing details.
4. *Good*: The fitted garment matches the given garment well, with only minor inaccuracies or slight deviations in details.
5. *Excellent*: The fitted garment precisely and fully matches the given garment in all important aspects, with high accuracy and fidelity.

Garment Fitting - Realism: How realistic or natural are the images?

1. *Very Poor*: The garment looks completely unrealistic on the body, with obvious distortions or unnatural fitting.
2. *Poor*: The garment appears clearly fake, with major fitting errors or unnatural body interaction.
3. *Moderate*: The garment looks somewhat realistic but shows noticeable fitting artifacts or awkward draping.
4. *Good*: The garment generally looks realistic on the body, with only minor fitting flaws.
5. *Excellent*: The garment appears highly realistic and naturally integrated with the body, nearly indistinguishable from a real photo.

Garment Fitting — Aesthetics: How visually pleasing are the final images?

1. *Very Poor*: The final images are visually unappealing and unattractive overall.

2. *Poor*: The final images have weak aesthetic appeal; they look awkward, unbalanced, or poorly presented.
3. *Fair*: The final images are somewhat visually pleasing but lack harmony, polish, or refinement.
4. *Good*: The final images are visually appealing and well-composed, with only minor aesthetic imperfections.
5. *Excellent*: The final images are highly attractive, stylish, and beautifully composed, showing strong aesthetic quality.

Garment Fitting - Sharpness: How sharp and detailed are the images?

1. *Very Poor*: The garment is extremely blurry on the body, with no clear details and severe artifacts.
2. *Poor*: The garment is mostly blurry, with major loss of detail and noticeable artifacts.
3. *Moderate*: The garment has some sharp areas, but overall appears soft or partially blurred, with moderate artifacts.
4. *Good*: The garment is generally sharp and detailed on the body, with only minor softness or small artifacts.
5. *Excellent*: The garment is very sharp and crisp, highly detailed, and free from visible artifacts when fitted.

G.2. Detailed Statistics

The detailed evaluation of the user study for garment generation and fitting can be found in Figure 3 and 4, respectively.

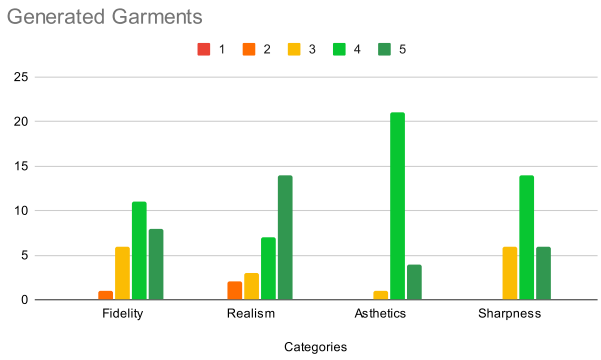


Figure 3. Histogram of user ratings for generated garments, showing distributions across realism, fidelity, aesthetics, and sharpness.

H. Qualitative Comparison

Masking: The quality of masks is an essential part of a realistic output. As described in the main paper 3.3, by using dilation and erosion of masks, we are able to increase the quality significantly. A comparison can be found in figure 5.

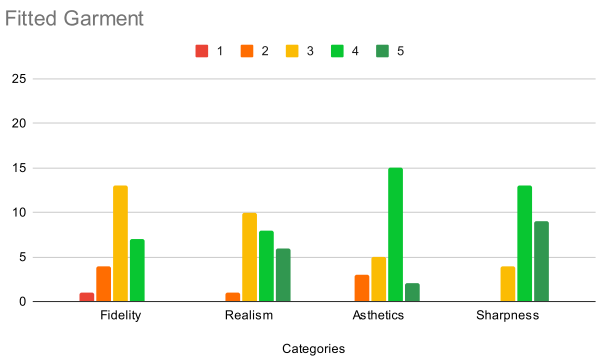


Figure 4. Histogram of user ratings for fitted garments, showing distributions across realism, fidelity, aesthetics, and sharpness.

Figure 5. Comparison of agnostic mask and final image with and without mask dilation. The top image shows results without dilation (severe artifacts on the face), while the bottom image shows improved results with dilation.

Garment and background generation: Figure 6 and 7 show examples for garment and background generation, respectively. There, you can also find the related user prompts used.

Generated Garmens vs VITON-HD: A comparison between pipeline results based on generated garments via Stable Diffusion 3 and garments from the VITON-HD Dataset can be found in figures 9, 8, and 10.

A red, yellow, and pink striped sweater with a round neckline, fitted silhouette, and a zipper on the side.		A yellow flowered shirt with a round neckline, fitted silhouette, and floral pattern	
A black t-shirt with a white heart on the front, made of 100% cotton. The shirt has a regular fit and a classic silhouette.		A white t-shirt with a red rose design on it. The shirt has a fitted silhouette and a round neckline.	
A green sweater with a pink logo on the front, featuring a round neckline and a fitted silhouette		A blue polo shirt with a white collar and a blue and white polo player logo on the chest. The shirt has a fitted silhouette and is ideal for a casual or sporty style.	

Figure 6. Garment generation utilizing SD3 based on user-given prompt. Note that the user-defined prompt is extended by garment template.

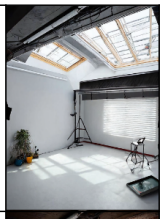
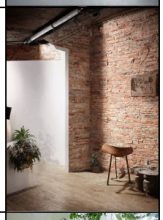
A white elegant wall in a modern house with minimalist décor.		A modern art studio with large skylights.	
A spacious gallery with abstract paintings		A contemporary loft with exposed brick walls.	
A bright living room with large windows.		A sunny conservatory filled with lush plants.	

Figure 7. Background generation utilizing SD3 based on user-given prompt. Note that the user-defined prompt is extended by the background template.

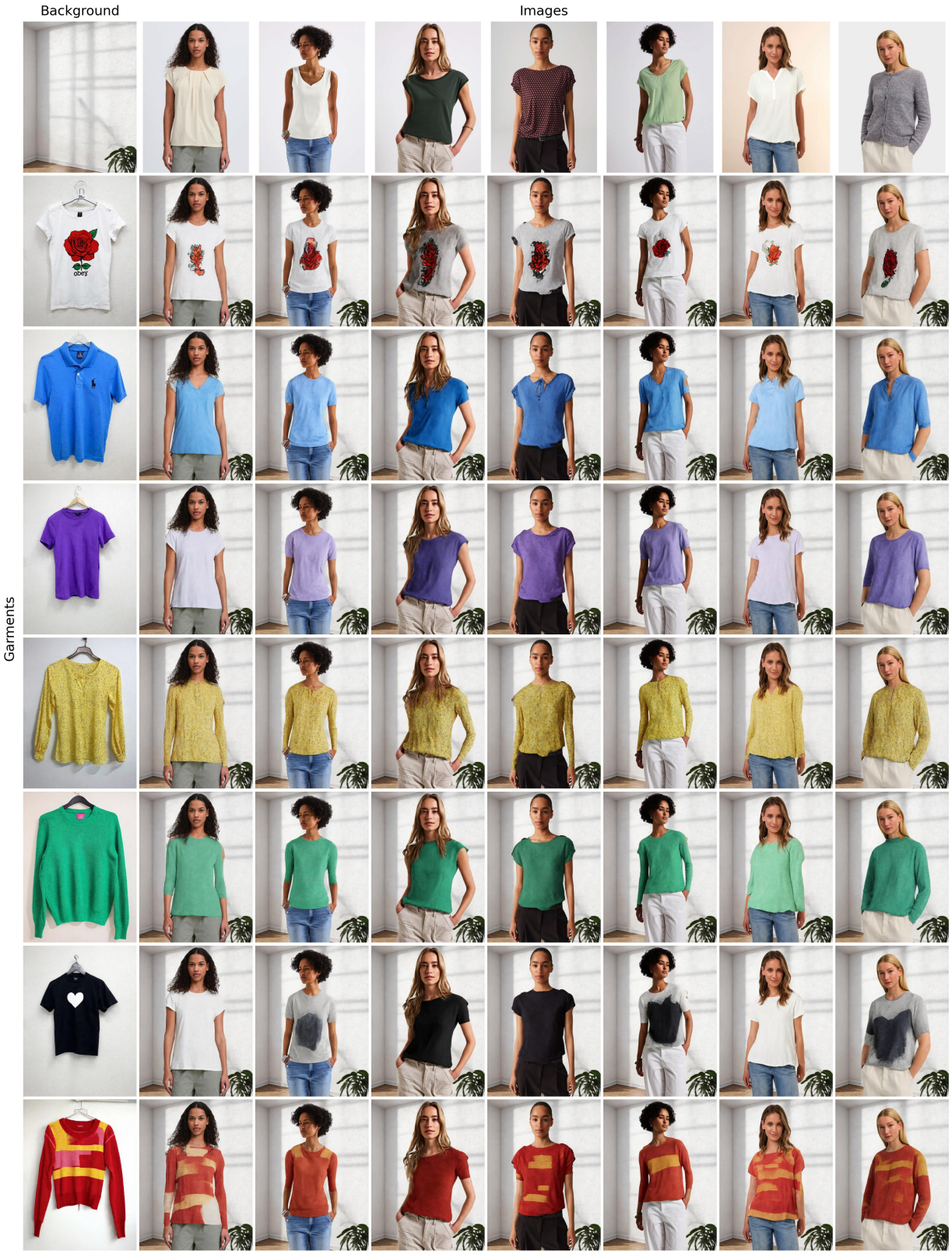


Figure 8. Comparison matrix of pipeline results utilizing SD3-generated garments and subject images from Zalando at a resolution of 1024×768.

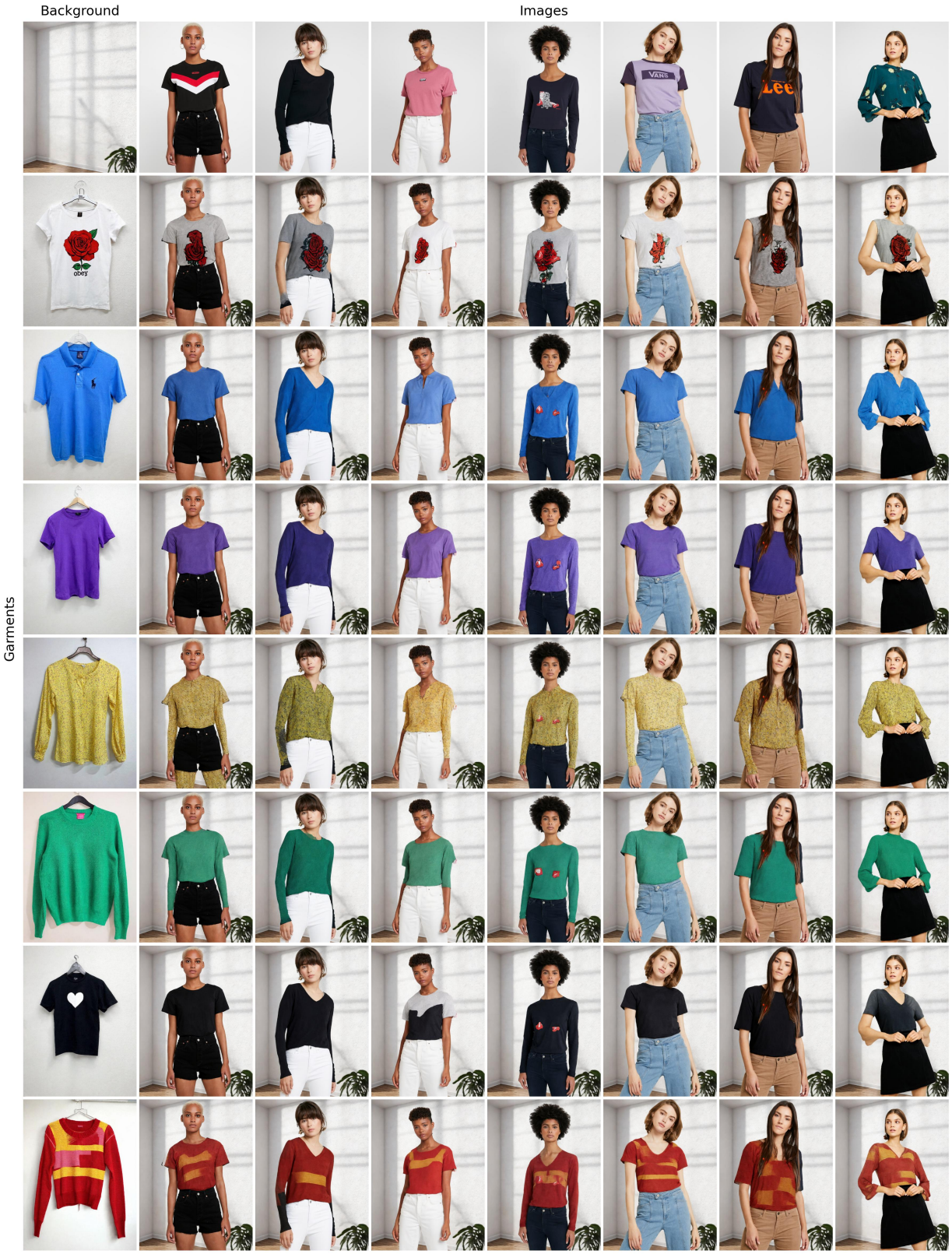


Figure 9. Comparison matrix of pipeline results utilizing SD3-generated garments and subject images from VitonHD at a resolution of 1024×768.

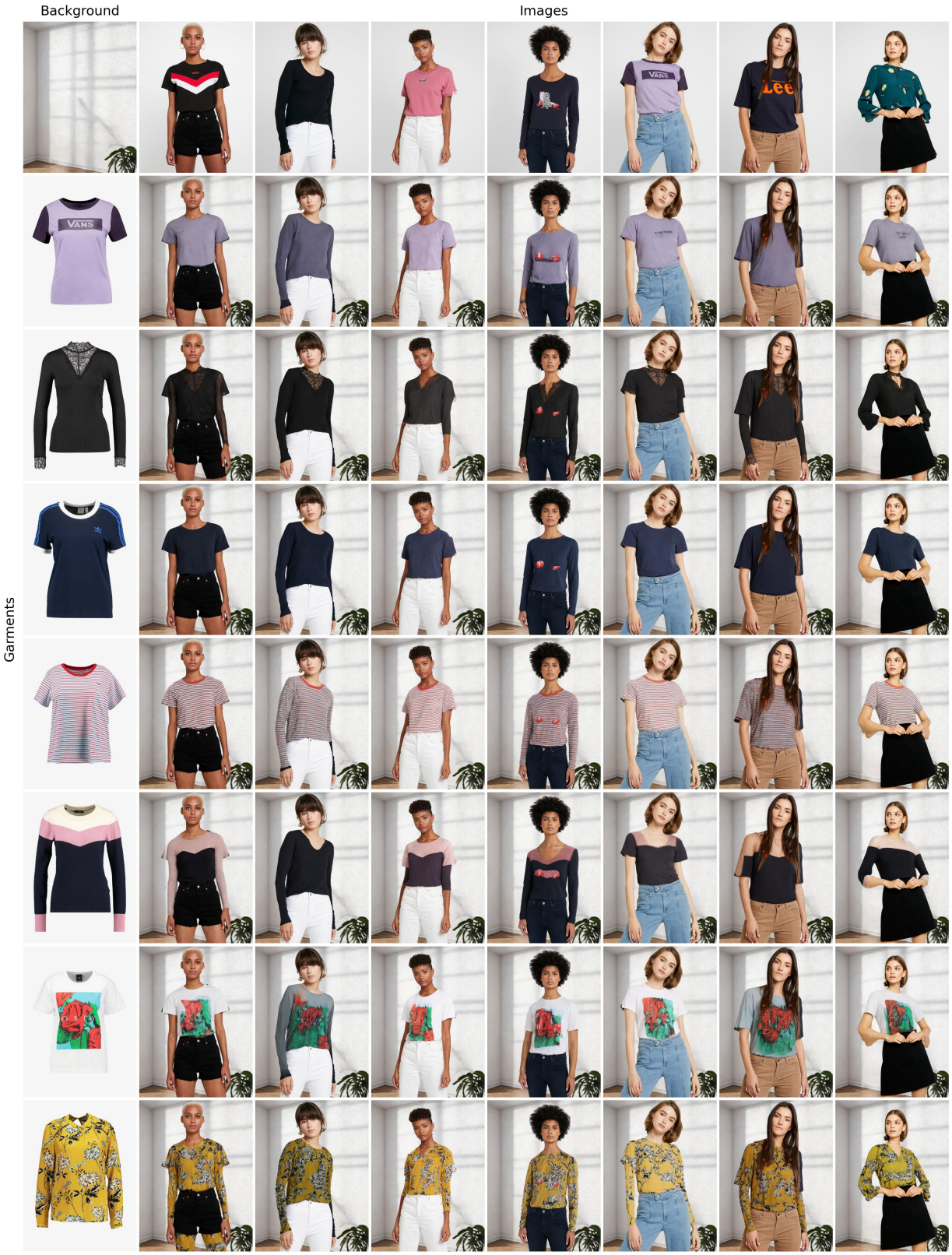


Figure 10. Comparison matrix of pipeline results utilizing garments and subject images from VitonHD at a resolution of 1024×768.