GenAI-Driven Image Generation Pipeline for Sustainable Garment Design and Waste Reduction in Fashion Production

Ilham Ghori [1], **Kayvan Karim** [1], **Dima Alkawadri** [1]

[1] Heriot-Watt University
ilhamghori@hotmail.com, K.Karim@hw.ac.uk, D.Alkawadri@hw.ac.uk

Abstract

The fashion industry's linear production model generates significant pre-consumer textile waste, especially during pattern cutting. In response to the environmental impact of fashion consumption, strategies such as reuse, recycling, and refashioning aim to divert textiles from landfills and promote sustainable practices. However, challenges in the textile sector—such as raw material variability and complex manufacturing—require more targeted solutions. Recent studies have identified Artificial Intelligence (AI) as a promising tool to enhance sustainability, streamline production, and enable personalised design. One such advancement is Generative AI (GenAI), which supports applications like virtual try-ons, fabric-to-garment transformations, and multimodal garment design via tools such as FashionGAN, StyleGAN, and Latent Diffusion Models. Despite these developments, current image generation methods struggle with preserving fabric detail and structural accuracy. This research proposes an image generation pipeline that accurately reflects specific fabric textures and visual attributes, offering designers greater creative control while reducing the need for physical samples—thereby minimising process waste. The system is implemented using ComfyUI and LoRA-enhanced Stable Diffusion 1.5 models to overcome limitations found in existing methods. To evaluate performance, quantitative metrics such as FID, KID, SSIM, LPIPS, and CLIP-S were used to assess visual quality, structural similarity, and semantic alignment. A qualitative comparison was also conducted to evaluate fabric texture preservation and prompt consistency across models. Among the tested models, Realistic Vision v5.1 delivered the best results across most metrics and is recommended for photorealistic applications in sustainable fashion. DreamShaper v8 excelled in preserving fabric texture, while MajicMix v5 produced stylised outputs more suitable for conceptual design stages. This study aims to empower fashion designers with a flexible and sustainable design model, to reduce waste, accelerate prototyping, and explore AI-driven innovation in digital fashion.

Introduction

The garment manufacturing industry has significant environmental impacts, including resource depletion, pollution, and waste generation (Ragab et al. 2024). This impact necessitates sustainable practices. Studies showed that adopting green practices in garment manufacturing is vital for environmental sustainability and economic performance. Principles like Lean production, including 5S, Value Stream Mapping, and Single Minute Exchange of Die, have been shown to improve environmental performance by identifying and eliminating waste in production processes (Marudhamuthu and Krishnaswamy 2011). These practices not only enhance productivity and quality but also lead to reduced emissions and pollution prevention (Marudhamuthu and Krishnaswamy 2011). However, Lean manufacturing implementation in the garment industry addresses challenges, such as high labor costs, short delivery times, and frequent style changes (Kumari, Quazi, and Kumar 2015). Additionally, the garment industry faces challenges in adopting these practices due to unpredictable demand fluctuations and complex manufacturing environments (Abd Jelil 2018).

One approach is to use the advancements in artificial intelligence (AI) in fashion production. These advancements have brought transformative changes across fields like art, design, and fashion (Singh and Patras 2024). Within the fashion industry, image generation models have the potential to revolutionize design workflows and streamline prototyping processes, offering new ways to automate and enhance creativity (Wang et al. 2024). Fashion image synthesis, in particular, enables designers to generate realistic garment images from textual descriptions or visual inputs, allowing them to explore design concepts without needing physical samples (Goodfellow et al. 2020; Ho, Jain, and Abbeel 2020).

Despite these capabilities, existing models face notable limitations. Many struggle to maintain complex patterns, such as stripes and text, and fall short in capturing nuanced fabric textures and material qualities—essential aspects for realistic fashion applications (Wang et al. 2024; Sun et al. 2023). Current methods offer control over basic elements like color but lack the ability to handle detailed textures and materials, which are challenging to accurately describe and model (Zhang, Zhang, and Xie 2024).

Background

The field of image generation has evolved through a variety of model architectures, each contributing uniquely to visual synthesis. Early models like Variational Autoencoders (VAEs) introduced probabilistic latent spaces to gen-

erate diverse, though often blurry, outputs (Kingma 2013). Generative Adversarial Networks (GANs) gained popularity for producing sharper, realistic images through adversarial training between generator and discriminator networks, but they suffer from training instability and mode collapse (Goodfellow et al. 2020). Convolutional Neural Networks (CNNs), while not generative by design, have been foundational in extracting visual features for downstream generative tasks, aiding in structure, texture, and content preservation (Rawat and Wang 2017). These models laid the groundwork for more recent approaches in high-fidelity and controlled image synthesis.

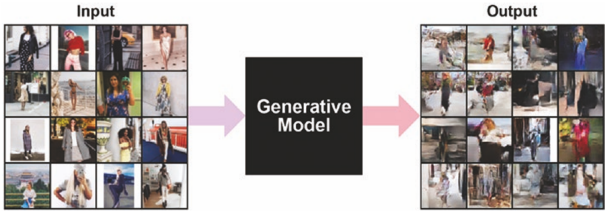


Figure 1: An overview of the generative model process, demonstrating how it takes an input dataset and creates "fake" images that closely mimic the original data (Luce 2018).

Diffusion Models Diffusion models, introduced by (Sohl-Dickstein et al. 2015), are a type of probabilistic generative model that uses a parameterized Markov chain to create samples resembling the original data. It learns to reverse a noise-adding diffusion process, where noise is gradually introduced to the data in small increments until the original signal becomes unrecognizable. By setting the transitions in this sampling process to conditional Gaussian distributions, the model enables a straightforward neural network configuration for effective data generation (Ho, Jain, and Abbeel 2020). This approach has been shown to generate high-quality samples and has applications in various fields, including image synthesis (Singh and Patras 2024).

One of the main advantages of diffusion models is their ability to produce high-quality images with broad distribution coverage, which reduces the risk of mode collapse—a common problem in other generative models like GANs. Diffusion models also benefit from a stationary training objective, making them more stable and easier to scale compared to GANs (Dhariwal and Nichol 2021). However, a notable disadvantage is their slower sampling time, as they require multiple denoising steps, each involving forward passes through the network. This can result in diffusion models being 5-20 times slower than GANs in generating images, especially as diffusion models tend to have larger architectures (Dhariwal and Nichol 2021).

Applications: Diffusion models have been used in high-quality image generation, video synthesis, and generating medical images, where precise detail is critical.

Fashion-Specific Image Generation Models

In fashion image synthesis, generative models like FashionGAN and StyleGAN have been instrumental in tasks such as fabric-to-garment transformation, virtual try-on, and texture-aware garment generation. These models demonstrate how visual and conditional inputs can guide structured image generation and inform the design of pipelines in this research.

FashionGAN FashionGAN was introduced to simplify garment visualization by generating images that combine fabric textures with fashion sketches (Cui et al. 2018). Built on a conditional GAN framework, the model encodes fabric textures into latent representations, which are then combined with sketches to synthesize complete garment visuals. A novel loss function enhances regular pattern generation, making the model useful for structured design outputs.

However, FashionGAN faces challenges in preserving irregular or complex patterns, and its reliance on sketches limits its flexibility when working with real fabric inputs. Despite these constraints, the model remains relevant for exploring texture control and guided garment synthesis.

StyleGAN StyleGAN is a high-resolution image synthesis model capable of fine-grained control over garment features like texture, pose, and style through latent space manipulation (Karras et al. 2020). It introduces a mapping network to transform input noise vectors into intermediate latent codes, allowing control at different synthesis stages using adaptive instance normalization (AdaIN).

This capability enables consistent style and texture manipulation, which is particularly useful in fashion applications. While StyleGAN can produce highly realistic results, training and customization remain computationally intensive, and visual artifacts such as distortions may affect output fidelity (Choi, Park, and Park 2022). Nonetheless, its structured latent manipulation makes it valuable in building guided generation pipelines.

Multimodal Approaches in Image Synthesis

The integration of multimodal inputs, such as text and images, has become a key advancement in generative modeling, especially in creating detailed and customized outputs. The following approaches allow models to better align generated images with specific user intentions, such as generating garments from both descriptive and visual cues, enhancing creative control and application versatility:

CLIP (Contrastive Language–Image Pre-training): Developed by OpenAI, CLIP learns visual concepts from natural language descriptions by pretraining on diverse image-text pairs (Radford et al. 2021). It encodes images and texts into a shared latent space, enabling alignment between visual content and language (Radford et al. 2021). In fashion image generation, CLIP can be paired with generative models to control image outputs based on text prompts. For example, in fashion-specific applications, CLIP can help ensure that generated garment features match specified attributes like color, style, or pattern (Baldrati et al. 2023; Singh and Patras 2024).

ControlNet: ControlNet is an extension of diffusion models that enhances control over generated images by integrating conditional inputs, such as sketches, poses, or segmentation maps, into the image synthesis process (Zhang, Rao, and Agrawala 2023). In fashion image generation, it allows models to fine-tune specific visual elements while maintaining flexibility in the overall composition. This approach is used in models like FashionSD-X, where ControlNet's conditional layers enable designers to visualize garment designs based on both text descriptions and garment outlines, allowing for precise control over garment structure and style.

Related Work

Research in fabric-to-garment generation has produced several notable models that leverage both deep learning and generative techniques to aid the fashion industry. These studies underscore the advancements made in generating high-quality garment images while highlighting ongoing challenges in maintaining fabric realism, structural integrity, and user control over design outputs.

SGDiff One significant advancement in the field is SGDiff, a style-guided diffusion model that incorporates both textual and visual inputs for garment generation. By allowing text-based descriptions of style and fabric images to guide the output, SGDiff provides designers with enhanced creative control (Sun et al. 2023). A notable achievement of SGDiff is its dual-modality input capability, which allows for a richer depiction of aesthetic and material properties in garment images. However, the model struggles to represent intricate fabric textures, especially with complex materials. This limitation arises due to SGDiff's focus on style guidance rather than material-specific features, which often results in less realistic textures in complex fabrics (Sun et al. 2023).

DiffFashion The DiffFashion model addresses structural preservation in garment generation. Unlike SGDiff, which emphasizes style control, DiffFashion combines a garment's source image with a reference texture or pattern to produce new designs that maintain the original garment structure (Cao et al. 2023). This model's strength lies in its ability to retain the garment's shape, making it suitable for designs where structure is crucial. However, DiffFashion's limitation lies in its visual coherence; when reference textures differ significantly from the original material, the final output may appear unnatural, thus limiting its effectiveness in realistic fabric-to-garment synthesis (Cao et al. 2023).

FashionSD-X FashionSD-X expands on multimodal synthesis by incorporating sketches, text prompts, and textures to condition garment generation in a structured way. Through a pipeline combining Low-Rank Adaptation (LoRA) and ControlNet, it synthesizes garments that reflect both visual structure from sketches and thematic cues from text inputs, supporting creative control for fashion designers (Singh and Patras 2024). Although promising, FashionSD-X's reliance on multimodal inputs increases the complexity of training and demands careful calibration to balance each modality's influence on the final output.

StableGarment StableGarment applies a garment-centric approach using stable diffusion with a dedicated garment encoder and ControlNet to maintain intricate textures while performing try-on tasks. This framework allows flexible generation of fashion images by maintaining garment textures even under varied poses or styles, which improves image quality and utility for applications like virtual try-on (Wang et al. 2024). Nonetheless, StableGarment's focus on garment-centric generation leaves room for improvement in more creative, text-driven garment design where designers may seek dynamic visualizations of concepts (Wang et al. 2024).

Purpose	Packages/Tools
Image Generation Pipeline	ComfyUI, Diffusers, ComfyUI-Manager, LoRA
Model Weights	Stable Diffusion v1.5, DreamShaper v8, RealisticVision v5.1, MajicMix v5
Data Handling	os, glob, shutil, Google Drive API
Image Processing	PIL, cv2, torchvision.transforms
Evaluation Metrics	torchmetrics, lpips, skimage, clip, numpy
Environment	Google Colab, NVIDIA A100 GPU

Table 1: Key Libraries and Tools Used

While these models have advanced fabric-to-garment synthesis in various ways, limitations remain in each. Models like SGDiff and DiffFashion struggle with accurately preserving fabric textures and structural realism, especially with complex patterns and materials (Sun et al. 2023; Cao et al. 2023). Similarly, multimodal models like FashionSD-X highlight the need for enhanced control and interpretability in garment generation processes, especially when working with diverse stylistic inputs (Zhang, Zhang, and Xie 2024; Singh and Patras 2024). Addressing these gaps, the proposed approach focuses on integrating textual control, structure-aware synthesis, and texture fidelity to better meet the practical needs of designers in the fashion industry.

Methodology

Development Tools and Environment

The development and experimentation process was conducted using Google Colab Pro+, which provided access to an NVIDIA A100 GPU. This was crucial for efficiently running computationally intensive workflows in ComfyUI, performing evaluations, and preprocessing large datasets. Colab was selected for its powerful GPUs, seamless integration with Google Drive, and support for Python-based workflows.

The entire implementation workflow was structured and executed through modular Google Colab notebooks. Using the A100 GPU environment, key stages of the implementation were separated for clarity, scalability, and ease of reproducibility.

ComfyUI Execution Notebook: Responsible for loading and launching the ComfyUI interface. This notebook handled the setup of the environment, installation of depen-

dencies, model downloads, and initiation of the fabric-to-garment generation workflows.
Evaluation Metrics Notebook: Conducted quantitative evaluations of the generated outputs using metrics such as FID, KID, SSIM, LPIPS, and CLIP-S. The results were recorded for each of the three tested models: DreamShaper v8, Realistic Vision v5.1, and MajicMix v5.
Preprocessing Notebook: Preprocessed all fabric screenshots into model-compatible images.

Dataset

Source and Collection The fabric images used for this research were derived from the Fashion Product Images Dataset available on Kaggle (Aggarwal 2017). A custom dataset was manually created by taking high-quality screenshots of fabric regions from product thumbnails that clearly displayed texture and pattern. These fabric screenshots served as the primary conditioning input for the garment generation pipeline.

Preprocessing To ensure compatibility with model requirements, the following preprocessing steps were applied:
Resizing: All fabric images were resized to a resolution of 512×512 pixels using LANCZOS interpolation, which helps preserve texture fidelity.
Normalization: Pixel values were scaled to the range [0, 1] and cast to uint8 format before being used in ComfyUI.
Filename Consistency: Images were renamed using a fixed naming convention (e.g., fabric_01.png to fabric_10.png) to ensure easy tracking and evaluation alignment.

Prompt Design Each fabric image was paired with a descriptive prompt to guide the model in generating a garment that reflects both the structure and style of the intended output. The prompt template used was:

"A <garment type> made from this fabric."

This design allowed the model to infer both the semantic garment category (e.g., dress, shirt, jacket) and the material context from the accompanying fabric image. By using a fixed, simple sentence structure across all samples, we ensured prompt consistency, which is a best practice when working with diffusion models, especially those using CLIP-based guidance (Developers 2023).

Furthermore, the garment types were manually selected to reflect a variety of fashion categories, ensuring that the generation pipeline could handle different levels of garment complexity (e.g., casual shirts vs. layered jackets) and formality. The simplicity of the prompts was intentional, allowing the image input (i.e., the fabric texture) to remain the primary visual driver, while the text helped anchor the type of output garment desired.

Model Overview

The architecture and workflow for generating garment images from fabric inputs were designed using ComfyUI, a node-based interface for Stable Diffusion that enables a visually interpretable pipeline. To compare performance and identify the most suitable configuration, three Stable Diffusion 1.5-compatible pre-trained models were used in the experimentation phase.

ComfyUI Architecture and Workflow ComfyUI is a modular, graph-based user interface for working with Stable Diffusion models (Contributors 2023). It enables precise control over every step in the diffusion process and supports custom workflows by connecting individual components such as model checkpoints, samplers, VAE decoders, and CLIP text/image encoders.

- **Load Checkpoint Node:** Loaded the pre-trained SD 1.5-compatible model checkpoint.
- **CLIP Text Encoder Node:** Encoded the user-defined prompt (e.g., "A dress made from this fabric").
- **LoRA Node:** Optional node for style guidance using a lightweight fine-tuned LoRA model.
- **Load Image Node:** Provided the preprocessed 512×512 fabric image input.
- **VAE Encode + Decode Nodes:** Encoded and decoded latent representations of the image.
- **KSampler Node:** Controlled generation with parameters like steps, sampler type, CFG scale, denoising, and seed.
- **Save Image Node:** Saved the output garment image.

Model Configurations and Parameters To determine the optimal generation settings for high-quality, fabric-aware garment synthesis, six unique configurations were selected for experimentation. These were not chosen at random but designed to explore variations across the key parameters that influence image generation quality in Stable Diffusion pipelines.

The Classifier-Free Guidance Scale (CFG) controls how strongly the generation adheres to the input prompt. A low value may produce outputs loosely guided by the prompt, while too high a value can lead to over-saturation or loss of realism. We tested a range from 7 to 10 to evaluate this tradeoff.

The Steps parameter defines how many denoising iterations the model performs. More steps typically allow better detail recovery but come at the cost of generation time and potential overfitting. We tested values between 40 and 60.

The Denoise Strength affects how much of the original latent image is preserved during generation. A lower value retains more of the original fabric structure, while a higher value encourages creative deviation. Our tests spanned values from 0.85 to 1.0.

The Sampler Type controls the denoising algorithm. We selected several samplers: Euler (known for smooth and balanced results), DPM++ 2M and DPM++ 2M SDE (advanced samplers offering more control but sometimes producing sharper or chaotic details), and UniPC (a modern sampler that balances noise and quality but may distort finer textures).

The parameter tuning was performed using the DreamShaper v8 model on the input fabric: fabric_7.jpg.

Prompt Used: *"A floral dress made from this fabric."*

Model Used: DreamShaper v8 (Stable Diffusion 1.5)

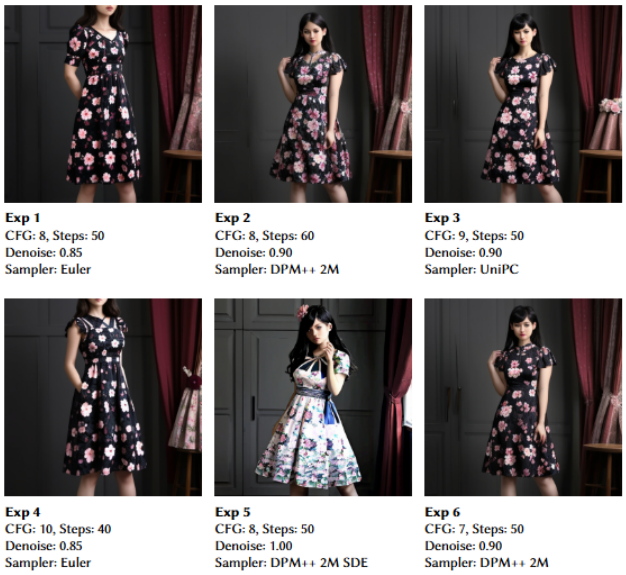


Figure 2: Parameter Experimentation Results using DreamShaper v8 on fabric_7.jpg for the prompt: *"A floral dress made from this fabric."*

Parameter Experimentation Results These six configurations (Exp 1–6) provided us with a comprehensive grid of different stylistic outputs, helping to empirically identify the best tradeoff between prompt fidelity, fabric realism, and garment structure. This rational selection enabled us to fine-tune the generation pipeline with a strong foundation for consistent evaluation across all models.

Among the tested configurations, Experiment 1 produced the most visually accurate and semantically aligned output. The resulting garment most effectively reflected the floral pattern of the input fabric and remained faithful to the prompt: *"A floral dress made from this fabric"*. It offered a good balance of prompt adherence and texture preservation, without the over-sharpness or distortion seen in other configurations. In comparison, Experiment 2 (DPM++ 2M) showed improved clarity but introduced structural inconsistencies in the dress, while Experiment 3 (UniPC) resulted in an overly saturated and slightly distorted output. Experiment 4 used fewer steps, leading to less detailed recovery, and Experiment 5 with Denoise = 1.0 lost too much of the original fabric identity. Finally, Experiment 6 (lower CFG) yielded outputs that lacked strong prompt adherence.

Thus, Experiment 1's configuration was chosen as the fixed setup for evaluating all three models to ensure a fair and consistent comparison.

Parameter	Value
CFG Scale	8
Steps	50
Denoise Strength	0.85
Sampler	Euler
Seed	2024

Table 2: Final fixed parameters selected for model comparison

Selected Pretrained Models from Civitai Three high-quality pre-trained models were selected from Civitai, each built upon the Stable Diffusion 1.5 architecture and known for their strong capabilities in realistic image generation. The models were chosen based on community reputation, performance benchmarks, and their relevance to fashion-focused visual synthesis (Civitai 2023).

All models were downloaded in the .safetensors format to ensure compatibility and secure weight handling in ComfyUI. The selection criteria centered around models' ability to capture fine fabric textures, adhere to text prompts, and maintain garment structure.

- DreamShaper v8
- Realistic Vision v5.1
- MajicMix Realistic v5

Model Name	Version	Description
DreamShaper	v8	Realistic outputs with artistic flair and fabric texture accuracy. Used as the base model.
Realistic Vision	v5.1	Photorealistic model with strong structure and sharp texture. Ideal for detailed synthesis.
MajicMix Realistic	v5	Balanced realism and soft style. Good for visually appealing and consistent garments.

Table 3: Pretrained models selected from Civitai for garment generation

Each model was tested using the exact same set of preprocessed fabric inputs and corresponding descriptive prompts. To ensure consistency and fairness across evaluations, the same fixed configuration—determined during the parameter experimentation phase—was applied to all generations.

Evaluation and Results

The evaluation strategy and results are presented through both quantitative and qualitative analyses to assess how effectively each model generates realistic garments that preserve fabric texture, follow the input prompt, and maintain garment structure. By applying established evaluation metrics from related works, we ensured objective assessment and fair benchmarking across models.

Quantitative Evaluation
The quantitative evaluation included metrics that measured the visual fidelity, structural similarity, and semantic alignment of generated garments with input fabrics and textual descriptions.

Fréchet Inception Distance (FID): FID, introduced by (Heusel et al. 2017), measures the similarity between two distributions of features from real and generated images using the Fréchet distance:

$$\text{FID} = ||\mu_r - \mu_g||^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2 \cdot (\Sigma_r \Sigma_g)^{\frac{1}{2}}\right) \quad (1)$$

- μ_r, μ_g: Mean vectors of features from real (r) and generated (g) images.
- Σ_r, Σ_g: Covariance matrices of features from real and generated images.

FID was used to evaluate the realism of the generated garments, offering insight into the model's ability to synthesize textures and structural garment details.

Kernel Inception Distance (KID): KID, introduced by (Bińkowski et al. 2018), calculates the distance between two feature distributions without assuming Gaussianity, based on the squared Maximum Mean Discrepancy (MMD):

$$\begin{aligned}\text{KID}(X, Y) =& \frac{1}{m(m-1)} \sum_{i \neq j} k(x_i, x_j) \\ &+ \frac{1}{n(n-1)} \sum_{i \neq j} k(y_i, y_j) \\ &- \frac{2}{mn} \sum_{i,j} k(x_i, y_j)\end{aligned} \quad (2)$$

- X, Y: Feature representations of real and generated images.
- k: Polynomial kernel function.
- m, n: Number of samples in each distribution.

KID allowed evaluation of distribution similarity between real and generated garments, complementing the FID evaluation.

SSIM and LPIPS: SSIM (Structural Similarity Index Measure) (Wang et al. 2004) measures local similarities in pixel intensity, capturing luminance, contrast, and structural fidelity:

$$\text{SSIM}(x, y) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (3)$$

LPIPS (Learned Perceptual Image Patch Similarity) (Zhang et al. 2018) computes perceptual similarity by comparing deep network features:

$$\text{LPIPS}(x, y) = \sum_l w_l \, ||\phi_l(x) - \phi_l(y)||_2^2 \quad (4)$$

- x, y: Real and generated image patches.
- ϕ_l: Activation from layer l in a deep neural network.
- w_l: Learned weights for layer l.

SSIM and LPIPS were used to assess how well structural consistency and perceptual quality were preserved in generated garments.

CLIP Score (CLIP-S): (Radford et al. 2021) introduced CLIP Score to evaluate semantic similarity between text and image embeddings using cosine similarity:

$$\text{CLIP-S}(t, i) = \frac{\phi_t \cdot \phi_i}{||\phi_t|| \, ||\phi_i||} \quad (5)$$

- t: Text description embedding.
- i: Generated image embedding.
- ϕ_t, ϕ_i: CLIP embeddings of the text and image.

CLIP-S was used to evaluate alignment between generated outputs and the descriptive garment prompts, ensuring the models generated garments consistent with user intent.

Qualitative Evaluation
The qualitative evaluation focused on visual comparison of generated garments across the three Stable Diffusion 1.5 models. This helped assess texture adherence, garment shape, and overall fidelity with the conditioning fabric and prompt.

Unlike traditional comparisons against state-of-the-art (SOTA) models like SGDiff or StableGarment, the evaluation focused on internal performance across the three models: DreamShaper v8, Realistic Vision v5.1, and MajicMix Realistic v5. These models were tested on identical fabric inputs and prompts using a fixed generation configuration to ensure fair comparison.

Visual Comparisons: Garment outputs were visually inspected to assess fabric structure preservation, prompt relevance, and garment realism. Specific attention was paid to fine details such as texture flow, design consistency, and fabric-structure alignment.

Quantitative Results
The quantitative evaluation compared the three selected models (introduced in the Model Overview section) using the five key metrics implemented in Python and PyTorch-based libraries. Table 4 summarizes the models' performance across these metrics.

To enhance understanding of the performance differences, the results were also visualized using graphical methods. Figure 3 displays the FID scores separately due to their larger numerical scale, clearly showing Realistic Vision v5.1 achieving the best realism. Meanwhile, Figure 4 presents a radar chart comparing the normalized KID, SSIM, LPIPS, and CLIP-S scores, offering a concise multivariate view of each model's performance profile. This visualization reveals that while DreamShaper v8 slightly outperforms in perceptual texture fidelity (LPIPS), Realistic Vision v5.1 consistently dominates across all other metrics, including structural alignment (SSIM), semantic accuracy (CLIP-S), and overall distribution similarity (KID). Together, these charts

Model	FID ↓	KID ↓	SSIM ↑	LPIPS ↓	CLIP-S ↑
DreamShaper v8	411.89	0.1816	0.3933	**0.6076**	0.2688
Realistic Vision v5.1	**362.12**	**0.0894**	**0.4460**	0.6628	**0.2829**
MajicMix Realistic v5	387.61	0.1733	0.3365	0.6432	0.2662

Table 4: Quantitative Evaluation Metrics for Each Model (lower FID, KID, LPIPS = better; higher SSIM, CLIP-S = better)

provide both metric-specific insights and an integrated performance overview, supporting more informed model comparisons.

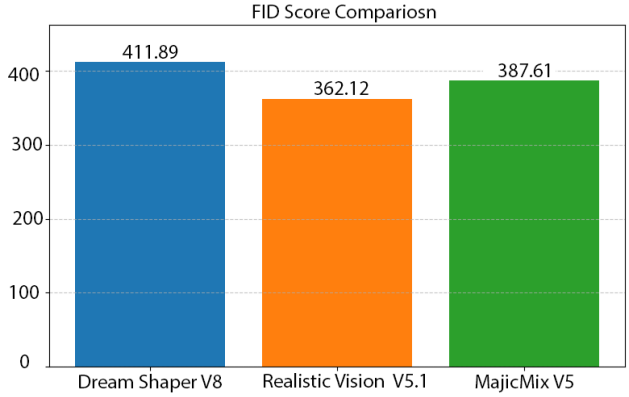


Figure 3: FID Score Comparison Across Models (Lower is Better)

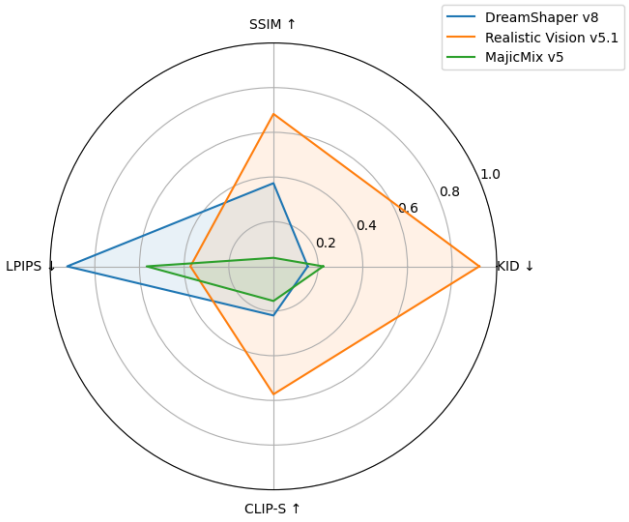


Figure 4: Radar Chart Comparison of Normalized Metrics: KID, SSIM, LPIPS, CLIP-S

Qualitative Results

A side-by-side visual comparison was conducted for each model using the same fabric input and prompt. This allowed a closer assessment of visual quality, texture retention, garment shape, and adherence to prompt semantics.

DreamShaper v8 produced consistent textures and vibrant color mapping across most outputs, making it suitable for maintaining fabric fidelity.

Realistic Vision v5.1 outperformed in structure retention and photorealism, accurately reflecting the intended garment type and silhouette.

MajicMix v5 delivered creative and stylish results, but sometimes exaggerated patterns or deviated slightly from the prompt semantics.

For example, in Row 7 (Floral Dress) of Table 5, both DreamShaper and Realistic Vision preserved the floral motif, but DreamShaper produced a more elegant and wearable garment. In contrast, MajicMix introduced stylized lighting and fashion flair, which, while visually appealing, led to more artistic deviation.

Conclusion

This research successfully explored the application of diffusion-based image generation pipelines for generating garments from fabric images and descriptive prompts, aiming to reduce reliance on physical prototyping and support sustainable fashion design. By building and evaluating a Stable Diffusion 1.5 pipeline in ComfyUI, three pretrained models were compared through quantitative and qualitative assessments. Realistic Vision v5.1 consistently achieved the best performance across realism, structural alignment, and semantic accuracy, making it ideal for photorealistic garment visualization workflows that minimize textile waste. DreamShaper v8, while weaker on realism metrics, excelled in perceptual texture fidelity, making it valuable for early-stage textile design exploration where preserving fabric detail is key. MajicMix v5 delivered balanced, stylized outputs but lagged behind in structural and semantic precision, making it more suitable for conceptual or editorial use. Together, these findings highlight the practical potential of diffusion-based models to transform the fashion design process by accelerating prototyping, enhancing creative flexibility, and supporting more sustainable industry practices.

Limitations

Despite achieving the core objectives, several limitations were encountered. Hardware constraints prevented direct model training or fine-tuning due to insufficient GPU memory, restricting the research to inference-only workflows in ComfyUI. The limited dataset scope, using only 10 fabric-garment pairs, provided a proof-of-concept but reduced exposure to diverse patterns, styles, and materials. Additionally, restricted model customization arose because, while ComfyUI allowed rapid pipeline construction, it lacked support for integrating custom model layers or loss functions without deeper source-level modification. Finally, reproducibility challenges were noted, as although ComfyUI pre-

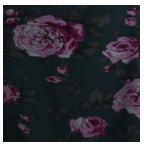
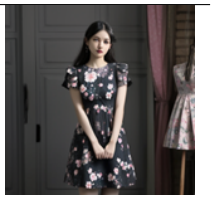
Row	Fabric	Prompt	DreamShaper v8	Realistic Vision v5.1	MajicMix v5
7		*"A floral dress made from this fabric"*			
8		*"A puffer jacket made from this fabric"*			

Table 5: Visual comparison of garments generated by DreamShaper v8, Realistic Vision v5.1, and MajicMix v5 on two selected fabric inputs (Rows 7 and 8).

Model	Strengths	Weaknesses	Best Use Case
Realistic Vision v5.1	Highest realism, structure and prompt alignment (best FID, KID, SSIM, CLIP-S)	Slightly lower texture preservation (LPIPS)	Sustainable design pipelines focused on photorealistic garment visualization
DreamShaper v8	Strong texture fidelity (best LPIPS), vibrant fabric rendering	Lower realism (worst FID), lower prompt alignment (CLIP-S)	Early-stage textile design exploration where fabric details are prioritized
MajicMix v5	Balanced and creative stylization	Lower structural and semantic accuracy across metrics	Editorial or conceptual fashion design where style is prioritized over precision

Table 6: Summary of Model Comparison: Strengths, Limitations, and Application Contexts

serves visual node graphs, detailed parameter logging and prompt-image matching documentation were not systematically maintained.

Future Work

Several opportunities remain to extend and improve this work. Custom model training using higher-end GPUs could enable fine-tuning or training diffusion models for enhanced texture transfer and garment fidelity. Stronger conditioning techniques, such as contrastive learning or multi-stage pipelines, may improve the model's ability to preserve intricate fabric patterns and structure. Expanded evaluation involving larger fashion industry datasets and direct feedback from designers would support more practical, user-centered assessments. Additionally, developing an interactive interface, such as a web-based frontend for real-time fabric uploads, prompt entry, and output generation, could make the system more accessible to design professionals. Finally, further exploration of LoRA tuning may enhance style transfer consistency across diverse garment types and prompts.

The dataset and project materials are available at: https://github.com/Ilham7x/fabric2garment-generation

References

Abd Jelil, R. 2018. Review of artificial intelligence applications in garment manufacturing. *Artificial Intelligence for fashion industry in the big data era*, 97–123.

Aggarwal, P. 2017. Fashion Product Images Dataset. https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-dataset. Accessed: 2025-03-27.

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*, 23393–23402.

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Cao, S.; Chai, W.; Hao, S.; Zhang, Y.; Chen, H.; and Wang, G. 2023. Difffashion: Reference-based fashion design with structure-aware transfer by diffusion models. *IEEE Transactions on Multimedia*.

Choi, I.; Park, S.; and Park, J. 2022. Generating and modifying high resolution fashion model image using StyleGAN. In *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, 1536–1538. IEEE.

Civitai. 2023. Civitai: AI Model Sharing Platform. https://civitai.com/.

Contributors, C. 2023. ComfyUI - A powerful and modular stable diffusion GUI. https://github.com/comfyanonymous/ComfyUI. Accessed: 2025-03-27.

Cui, Y. R.; Liu, Q.; Gao, C. Y.; and Su, Z. 2018. FashionGAN: Display your fashion design using conditional generative adversarial nets. In *Computer Graphics Forum*, volume 37, 109–119. Wiley Online Library.

Developers, C. 2023. ComfyUI Documentation – Prompting Best Practices. https://docs.comfy.org/get_started/introduction. Accessed: 2025-03-27.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.

Kingma, D. P. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kumari, R.; Quazi, T.; and Kumar, R. 2015. Application of lean manufacturing tools in garment industry. *International Journal Of Mechanical Engineering And Information Technology*, 3(1): 976–982.

Luce, L. 2018. *Artificial intelligence for fashion: How AI is revolutionizing the fashion industry*. Apress.

Marudhamuthu, R.; and Krishnaswamy, M. 2011. The development of green environment through lean implementation in a garment industry. *Journal of Engineering and Applied Sciences*, 6(9): 104–111.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ragab, A. G.; Al-Gizawy, A. S. H.; Al-Minyawi, O. M. A.; Hassabo, A. G.; and Mahmoud, M. N. I. 2024. Environmental impact of clothing manufacturing and the fashion industry. *Journal of Textiles, Coloration and Polymer Science*.

Rawat, W.; and Wang, Z. 2017. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9): 2352–2449.

Singh, A. K.; and Patras, I. 2024. FashionSD-X: Multimodal Fashion Garment Synthesis using Latent Diffusion. *arXiv preprint arXiv:2404.18591*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Sun, Z.; Zhou, Y.; He, H.; and Mok, P. 2023. Sgdiff: A style guided diffusion model for fashion synthesis. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8433–8442.

Wang, R.; Guo, H.; Liu, J.; Li, H.; Zhao, H.; Tang, X.; Hu, Y.; Tang, H.; and Li, P. 2024. StableGarment: Garment-Centric Generation via Stable Diffusion. *arXiv preprint arXiv:2403.10783*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, Y.; Zhang, T.; and Xie, H. 2024. TexControl: Sketch-Based Two-Stage Fashion Image Generation Using Diffusion Model. *arXiv preprint arXiv:2405.04675*.