

HiGarment: Cross-modal Harmony Based Diffusion Model for Flat Sketch to Realistic Garment Image

Supplementary Material

7. Dataset details

7.1. Dataset construction

The dataset collection team comprised six members, each assigned to three to five garment brand websites. Images were collected from multiple garment brand websites, with specific requirements to ensure clear display of fabric and color characteristics. Before official collection, all team members were required to carefully study a standardized tutorial and strictly follow the tutorial’s guidelines during the collection process. Images were collected from multiple garment brand websites, with specific requirements to ensure clear display of fabric and color characteristics. For each site, members adhered to predefined standards for image resolution, file format, and garment presentation to ensure consistent image quality across the dataset. Only images with a sufficiently clear display of fabric and color characteristics were selected to enable accurate identification of garment details. The annotation strategy required single-color garments to be labeled with precise color information, whereas garments with complex colors focused on fabric and structural annotations. To guarantee the reliability of the annotations, the team regularly consulted professional garment designers, who provided expert verification of fabric types and other visual features. Fig. 8 and Fig. 9 provide some samples of flat sketches, close-ups, and detailed descriptions from the MMDGarment dataset.

For the fabric vector database, the team also downloaded high-resolution sample images exhibiting distinct fabric characteristics, curated according to fabric types frequently encountered in professional fashion design practice. To ensure practical relevance, the covered fabric types were chosen based on frequency and importance in professional fashion design practice, as identified through consultation with experienced designers and review of contemporary fashion collections. The resulting fabric vector database consists of 150 image-text pairs, covering 11 major fabric categories (e.g., cotton, wool, denim, silk, lace, and synthetic blends), thereby providing a comprehensive and diverse reference for fabric representation tasks.

7.2. Dataset analysis

In addition to the main dataset, the MMDGarment also includes 1,975 close-up images of garment details, such as collars, sleeves, and pockets, with detailed fabric and color annotations essential for training models to capture intricate garment features. The flat sketches are technical

drawings of real garment designs, created by professional designers from our collaborating garment company. Tab. 4 demonstrates a comparison between MMDGarment and other mainstream fashion datasets [4, 21, 34]


Datasets	Public	Close-ups	Fabric	Color	Sketch	# Pairs	Example Image	Example Text
CM-Fashion [34]	✗	✓	✓	✓	✗	500,000		classic collar Navy blue cotton blend button-up trench coat
DressCode [21]	✓	✗	✗	✗	✗	53,795		(Body key points and segmentation masks are provided)
VITON-HD [4]	✓	✗	✓	✓	✗	13,679		Embroidery Lavender lettering prints Cotton Short Sleeve normal-fit Round Neck T-shirts
MMDGarment (Ours)	✓	✓	✓	✓	✓	20,151		White cotton t-shirt with a black round collar, white short sleeves, black sleeve cuff.

Table 4. Comparison of the most widely used datasets for garment synthesis tasks

8. Visualization

We conducted a series of visualization experiments on attention heat maps to demonstrate the effectiveness of our method in accurately generating garment components and attributes. Fig. 10 shows that the generated garment images capture attribute details, such as color and fabric, as specified in the textual input. We also compare the cross-attention visualization results between other methods and HiGarment as shown in Fig. 11. Our method demonstrates significantly more distinct attention to key components such as the pocket and collar. These results validate the critical role of modality harmony in addressing the FS2RG task.

9. MLLM-based evaluation and user study

The MLLM evaluation contains four aspects: structure (35%), color (25%), fabric (25%), and details (15%). Specifically, structure refers to the overall layout and key components of the garment, such as the shape, style, and arrangement of major parts (e.g., collars, sleeves, and hoods), demonstrating the model’s ability to reconstruct the correct garment shape and structural attributes. Color evaluates the accuracy and consistency of the generated garment’s colors with respect to the reference, reflecting the model’s capability in color reproduction and harmony. Fabric assesses whether the generated images correctly represent the type, texture, and appearance of the garment materials, thus veri-



Figure 8. Samples of the flat sketches and corresponding real garment images in the collected MMDGarment dataset.



Figure 9. Samples of the detailed description and close-ups in the collected MMDGarment dataset.

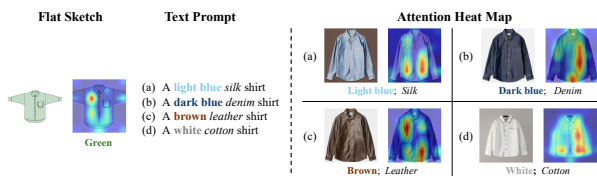


Figure 10. Cross-attention visualization for generated garment images with different color-fabric attribute composition.

ifying the model’s ability to capture subtle material differences. Details focus on fine-grained elements, including stitching, decorative features, and small structural or visual cues, highlighting the model’s proficiency in generating intricate garment details.

Each aspect is scored from 0 to 10 by the MLLM, with

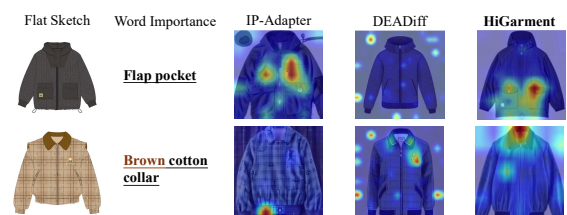


Figure 11. Cross-attention visualization comparison between HiGarment and other methods.

higher scores indicating better fidelity to the ground truth in that dimension. The lowest possible score for a pair is 0 (if all aspects score 0), and the highest possible score is 10 (if all aspects receive the maximum score). These four scores are then combined using a weighted sum to produce

a final score for each pair, which allows for a comprehensive assessment of the generated images across both global and local garment characteristics. For the evaluation process, we first selected 28 pairs of generated images and corresponding product images. For each pair, the associated instruction was provided to the MLLM as a prompt, together with both the generated and real product images. The MLLM was tasked to rate each pair independently on structure, color, fabric, and details according to predefined evaluation criteria. After scoring all four aspects for each pair, the weighted sum was calculated to obtain the pair’s overall score. Finally, the final evaluation result is reported as the average score across all 28 pairs, providing an objective and multi-faceted measure of the model’s performance in garment generation.

For the user study, we invited 2 professional garment designers from the corporate company, 11 fashion design students from NingboTech University, and 20 non-experts majoring in computer science to assess the similarity between generated garment images and real garment photos. Each participant was provided with a questionnaire that mirrored the evaluation criteria used in the MLLM assessment, covering four aspects: structure, color, fabric, and details. For each image pair, the participants were asked to score each aspect on a scale from 0 to 10, following clear evaluation guidelines. This human evaluation not only offers a subjective perspective complementary to the automated MLLM assessment but also enables us to examine the consistency between expert, trained, and layperson judgments. The strong alignment between MLLM scores and human ratings demonstrates the effectiveness of using large models as evaluators in this task. Moreover, the high scores achieved by our method in both objective (MLLM-based) and subjective (human-based) evaluations validate its robustness and reliability from both technical and perceptual standpoints.

fine elements are often lost during the diffusion denoising process, leading to visible discrepancies between the generated images and the ground truth products. To address these issues in the future, we plan to explore the integration of specialized modules or attention mechanisms designed to enhance the preservation of fine-grained patterns in the diffusion process.

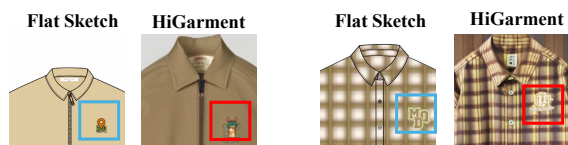


Figure 12. Failure cases for tiny logos.

10. Limitations and failure cases

We analyze the limitations and failure cases in this section. Current evaluation remains limited to design-stage synthesis due to scarce public datasets and garment generation codes. We will expand comparisons when resources [3, 33, 35] permit. Fig. 12 illustrates several representative failure cases observed in HiGarment, particularly in preserving tiny logos and intricate garment details. These