Semi-supervised Latent Disentangled Diffusion Model for Textile Pattern Generation

Chenggong Hu[1*], Yi Wang[1*], Mengqi Xue[2], Haofei Zhang[1], Jie Song[1], Li Sun[3†]
[1]Zhejiang University
[2]Hangzhou City University
[3]Ningbo Global Innovation Center, Zhejiang University
{huchenggong, y_w, sjie, haofeizhang, lsun}@zju.edu.cn, mqxue@hzcu.edu.cn,

Abstract

Textile pattern generation (TPG) aims to synthesize fine-grained textile pattern images based on given clothing images. Although previous studies have not explicitly investigated TPG, existing image-to-image models appear to be natural candidates for this task. However, when applied directly, these methods often produce unfaithful results, failing to preserve fine-grained details due to feature confusion between complex textile patterns and the inherent non-rigid texture distortions in clothing images. In this paper, we propose a novel method, SLDDM-TPG, for faithful and high-fidelity TPG. Our method consists of two stages: (1) a latent disentangled network (LDN) that resolves feature confusion in clothing representations and constructs a multi-dimensional, independent clothing feature space; and (2) a semi-supervised latent diffusion model (S-LDM), which receives guidance signals from LDN and generates faithful results through semi-supervised diffusion training, combined with our designed fine-grained alignment strategy. Extensive evaluations show that SLDDM-TPG reduces FID by 4.1 and improves SSIM by up to 0.116 on our CTP-HD dataset, and also demonstrate good generalization on the VITON-HD dataset.

Introduction

Contemporary image-to-image diffusion models demonstrate remarkable generative capabilities (Zhang et al. 2024; Li, Li, and Hoi 2024; Wang et al. 2024; Qi et al. 2024; Ju et al. 2024) across many tasks, such as virtual try-on (Choi et al. 2024) and scene generation (Huang et al. 2025). However, in the inherently visual and artistic field of fashion design, the adoption of such powerful generative tools for creating fashion designs remains limited. Progress in this area could substantially lower design costs while preserving creative fidelity. In this work, we refer to this relatively underexplored area as *Textile Pattern Generation* (TPG), a design process that integrates varied shapes, colors, and textures to produce novel and visually appealing patterns. TPG is central to the textile sector, with wide-ranging influence on fashion design, apparel creation, and textile manufacturing.

The core goal of TPG is to extract the complex textile pattern contents and design from a natural clothing image and reconstruct them into a highly realistic textile pattern image (simply referred to as the pattern image), as illustrated in Figure 1. The synthesized pattern image is expected to meet the following desiderata: (1) the generation should faithfully retain clear and consistent visual details from the clothing image, including not only low-level features such as color and texture, but also complex elements like floral motifs, geometric patterns and so on; (2) the results should exhibit a well-structured overall layout, similar to real pattern images; (3) the outputs must be free from any texture defects, such as *deformations*, *blurriness*, and *occlusions*. However, clothing images often suffer from texture defects, which contrast sharply with well-structured textile pattern images. Therefore, it is essential to eliminate these defects and accurately extract the underlying textile pattern content and design embedded in clothing images. This process naturally leads to feature disentanglement, which is key to enabling faithful and fine-grained reference-based textile pattern generation.

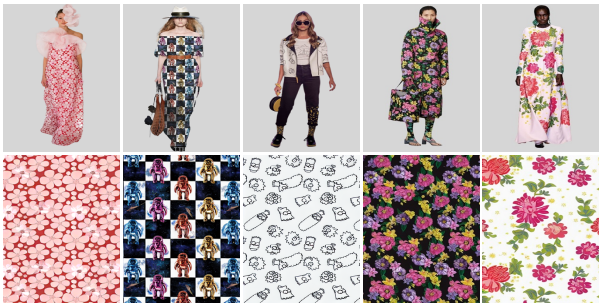


Figure 1: Textile pattern image results generated by our SLDDM-TPG from real worn clothing on our CTP-HD dataset. The first row show the clothing images and the second row show the generated textile pattern images.

Recently, popular image-to-image (I2I) diffusion models (Zhang et al. 2024; Gou et al. 2023; Gao et al. 2024) appear to be a natural choice for TPG. However, they rarely consider feature disentanglement required for processing clothing images, often leading to unfaithful generations. Specifically, I2I diffusion models typically input a whole image as a reference into the feature encoder and then use its output, the entangled features, to guide the generation. In TPG scenarios, clothing images contain not only the intended pattern

*These authors contributed equally.
†Corresponding author.

content but also unwanted texture defects. Conditioning generation directly on entangled features increases task complexity for diffusion models, making it harder to isolate informative signals and thus impeding the generation of high-fidelity pattern images with fine-grained details. Although text-guided approaches (Zhang et al. 2024) have been proposed to localize semantic features for disentanglement, the inherent visual complexity of textile pattern designs makes them difficult to describe precisely using text. Moreover, due to the limited availability of training data specific to TPG, these fully supervised learning methods often suffer from overfitting and exhibit poor generalization performance.

To address these challenges in TPG, we propose a semi-supervised latent disentangled diffusion model (SLDDM-TPG), a two-stage framework for generating faithful pattern images from clothing images. In the first stage, to address the unfaithful generation caused by feature confusion in clothing images, we propose a latent disentangled network (LDN) to disentangle useful and irrelevant information. LDN first extracts the **shared semantic content** between clothing and pattern images using a similarity contrast module (SCM), which serves as generation guidance. Meanwhile, a reverse attention module (RAM) captures **texture defects** specific to clothing images, which are used as negative prompts. However, clothing images often lack the well-structured layout specific to pattern images. We define this **structured information** as *flatness*, *clarity*, and *full visibility*. To recover such features, we introduce structure affine transformations (SATs), which map low-level defect features to structured representations. SATs are jointly trained with RAM to ensure feature stability and prevent drift from the original latent space. In the second generation stage, we propose a semi-supervised latent diffusion model (S-LDM) that strategically leverages labeled and additional unlabeled data to improve generation quality and generalization, guided by LDN features. The framework adopts a cascaded two-stage architecture consisting of a denoising process and an alignment process. The denoising process learns the distribution of pattern images through denoising training. The alignment process treats the model-predicted image as a semi-supervised optimization target, encouraging the output to align not only with the general data distribution but also with the specific content of each clothing image, guided by our proposed stable transformation domain (STD) loss. Recognizing local similarity as an intrinsic property of pattern images, we propose a convolutional local similarity (CLS) module to enforce local consistency via localized feature matching. We also present CTP-HD, a high-resolution dataset of paired clothing and pattern images for training.

In summary, our main contributions are as follows: (1) we propose SLDDM-TPG, the first model to achieve faithful and fine-grained clothing-based pattern generation; (2) we introduce CTP-HD, the first high-resolution paired dataset bridging clothing and pattern images; (3) our proposed LDN addresses the feature confusion in clothing images by disentangling their representations into multi-dimensional feature spaces; (4) we propose S-LDM that enables to utilize unlabeled data and we design an alignment process with a novel STD loss and CLS module to refine generation quality.

Related Work

Textile Pattern Generation. There are existing methods similar to TPG (Wu and Li 2024; Pang et al. 2019), but they all focus on unconditional synthesis without clothing image guidance. As a result, they are inherently incapable of generating garment-specific patterns from real-world clothing images, often producing uncontrollable outputs with low fidelity. Most GAN-based approaches (Fayyaz, Maqbool, and Hanif 2020; Wang and Sun 2021) additionally suffer from mode collapse and limited pattern diversity. Meanwhile, 3D-to-2D methods (Lu, Mok, and Jin 2017; Choi et al. 2007) rely on expensive 3D scanning data, which significantly limits the practical applicability. In contrast, our framework, SLDDM-TPG, enables controllable generation of real-world textile pattern images conditioned on RGB clothing images through a latent diffusion model. It achieve precise content alignment with the reference clothing image while preserving the well-structured overall layout of the real patterns.

Image-to-Image Diffusion Models. Recent advancements in diffusion models have significantly enhanced image-to-image tasks, including image inpainting, reference-based generation, and style transfer. For image inpainting, StrDiffusion (Liu et al. 2024) uses structure-guided denoising to preserve semantic consistency and DCI-VTON (Gou et al. 2023) treats virtual try-on as structure-aware inpainting guided by parsing maps. In reference-based generation, SSR-Encoder (Zhang et al. 2024) encodes image subject across multiple scales, IP-Adapter (Ye et al. 2023) injects image features into diffusion models via cross-attention for subject-driven synthesis, and UniCon (Li et al. 2025) models joint distributions over image pairs for direct reference-guided generation. For style transfer, StyleShot (Gao et al. 2024) introduces a style-aware encoder to extract style embeddings across diverse styles, while OSASIS (Cho et al. 2024) leverages Diff-AE (Preechakul et al. 2022) for one-shot facial style transfer. However, these methods overlook feature confusion in clothing images, resulting in unfaithful outputs. In contrast, our LDN addresses this issue by disentangling garment features for more faithful generation.

Method

Our method targets faithful and fine-grained generation of pattern images from clothing images. We adopt Stable Diffusion V1-5 (Rombach et al. 2022) as the backbone and introduce an adapter (Ye et al. 2023) to incorporate guidance. As shown in Figure 2, our method consists of two stages: (1) a latent disentangled network (LDN), employing three collaborative modules to resolve feature confusion and extract diverse features from clothing images; (2) a semi-supervised latent diffusion model (S-LDM), guided by the extracted features and an alignment process, and trained with unlabeled and limited labeled data to produce faithful results.

Latent Disentangled Network

As shown in Figure 2, our three-module latent disentangled network (LDN) effectively addresses feature confusion in clothing representations. Specifically, given a clothing image C, LDN disentangles the textile pattern content feature

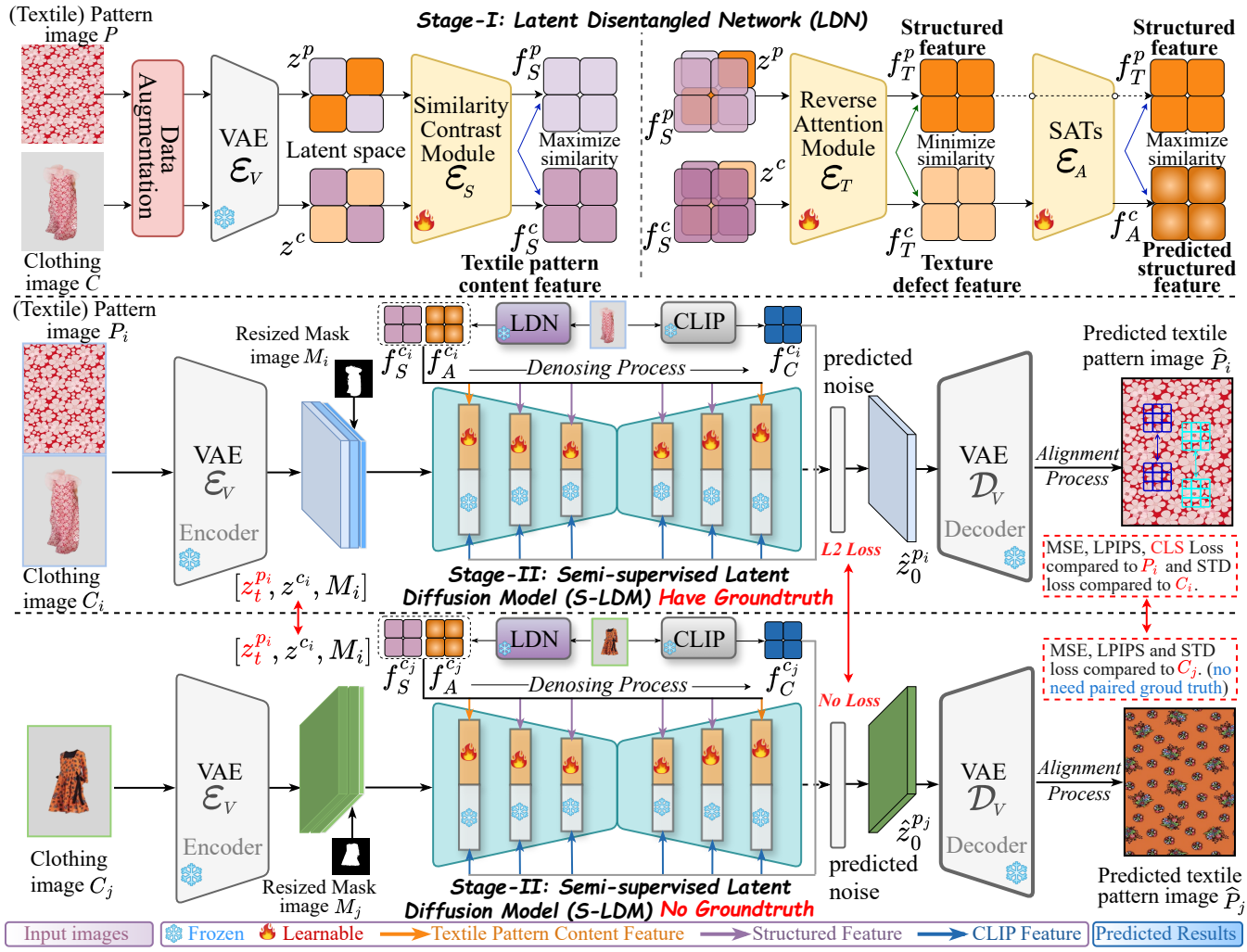


Figure 2: The framework of SLDDM-TPG. During training, LDN is first trained to disentangle features of clothing images C to **textile pattern content feature** f_S^c, predicted **structured feature** f_A^c, and **texture defect feature** f_T^c. Then S-LDM is trained, guided by LDN's features to generate pattern images P aligned with image C. There are data with and without ground truth in a batch. The two S-LDM share the same network and parameters, but their inputs, calculated losses, and comparison objects of losses are different, as can be seen in red parts in *stage-II*. During inference, a clothing image C is fed into LDN, whose output features are passed to S-LDM as conditions, and T-step denoising is then applied to the noise input to generate the final result.

f_S^c shared with the pattern image P, extracts the clothing-specific texture defect feature f_T^c, and predicts the structured feature f_A^c, which is absent in the clothing image C but important for faithful reconstruction.

Similarity Contrast Module. It is evident that clothing images C and pattern images P share semantic content that requires targeted reconstruction. To extract this shared information from image C, we introduce a similarity contrast module (SCM), which adopts SimSiam (Chen and He 2021) to perform contrastive learning between clothing images C and pattern images P for extracting the shared textile pattern content feature f_S^c. SimSiam applies different data augmentations to a sample to produce two views, analogous to a pair of clothing image C and pattern image P, for contrastive learning, as shown in Figure 2. SimSiam supports small-batch training and does not rely on negative samples, achieving strong performance via a stop-gradient operation. In our setting, training is conducted in the latent space, and we modify the original augmentation strategies to better suit our clothing and pattern images, thereby improving generalization performance. The optimization objective is defined as follows:

$$\mathcal{L}_{SCM} = -\frac{1}{2}(\text{sim}(s_{\mathcal{P}}^p, \text{sg}(s_{\mathcal{E}}^c)) + \text{sim}(s_{\mathcal{P}}^c, \text{sg}(s_{\mathcal{E}}^p))), \quad (1)$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and $\text{sg}(\cdot)$ represents the stop-gradient operation. Here, $s_{\mathcal{E}}^c$ and $s_{\mathcal{E}}^p$ denote the encoder outputs of SimSiam for the clothing and pattern images, respectively, while $s_{\mathcal{P}}^c$ and $s_{\mathcal{P}}^p$ denote the corresponding predictor outputs. Like SimSiam, we adopt the encoder output as the textile pattern content feature f_S^c after training.

Reverse Attention Module with SATs. Besides the shared content feature f_S^c, the clothing image C contains texture defects and lacks the structured feature specific to P. To avoid interference caused by these defects, we introduce the reverse attention module (RAM) to extract them as negative prompts. To train RAM, we first freeze the SCM. For C, we use f_S^c as the query (Q), and the latent feature z^c as both the key (K) and value (V). The original N cross-attention layers use Q to query semantically similar features from K, and compute attention weights A_i to aggregate the corresponding values V (Zhang et al. 2024), as defined below:

$$A_i = \text{Softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d}}\right), \quad i \in [1, N]. \tag{2}$$

However, in RAM, we reverse A_i as $A_i^r = \text{Normalize}(1 - A_i)$ in each layer to amplify coarse dissimilarities between Q and K. We then compute the new weighted sum and concatenate it with the latent feature z^c as a residual, forming the initial texture defect feature f_T^c, as follows:

$$\text{Initial: } f_T^c = z^c + \sum A_i^r \cdot V_i, \quad i \in [1, N]. \tag{3}$$

The same operation is applied to P to obtain the initial structured feature f_T^p. We then train the model to push f_T^c and f_T^p apart in the feature space, as they exhibit inherently opposite characteristics. However, the single constraint often leads to feature drift beyond the original latent space, resulting in the loss of basic semantics (details are provided in our ablation study). To ensure stable convergence during RAM training and address the absence of structured features in clothing images, we leverage the observation that both texture defect feature f_T^c and structured feature f_T^p are both low-level representations, but exhibit contrasting properties. This motivates us to apply affine transformations to convert f_T^c into f_T^p. To this end, we propose structured affine transformations (SATs) and jointly train them with RAM. SATs consist of a set of learnable affine transformation networks. For example, we design a convolutional filtering network for sharpness-adjusting transformations (details are provided in the Appendix). Inspired by (Schroff, Kalenichenko, and Philbin 2015), we design a texture triplet loss to enforce separation between the texture defect feature f_T^c and the structured feature f_T^p, while encouraging f_T^c to be transformable toward f_T^p. The loss is defined as:

$$\mathcal{L}_{TRIPLET} = \|f_A^c - f_T^p\|_2^2 - \|f_T^c - f_T^p\|_2^2 + \alpha, \tag{4}$$

where f_A^c is the structured feature predicted by SATs from f_T^c of C, and α is a margin that is enforced between f_T^c and f_T^p pair. In summary, the loss of LDN is as follows:

$$\mathcal{L}_{LDN} = \mathcal{L}_{SCM} + \mathcal{L}_{TRIPLET}. \tag{5}$$

Details of the data augmentations and network structures for the three modules are provided in the Appendix.

Semi-supervised Latent Diffusion Model

Our S-LDM includes denoising distribution learning using labeled data, and an alignment process that enables semi-supervised (Yin et al. 2025; He et al. 2024) training by leveraging unlabeled data through similarity constraints with reference images.

Denoising Distribution Learning Process. As shown in Figure 2, we freeze the backbone and use the features of the frozen LDN to train an added adapter. Inspired by (Voynov et al. 2023), the different layers of cross attention in the denoising UNet (Ronneberger, Fischer, and Brox 2015) are responsible for synthesizing different aspects of content. Thus, we feed the high-resolution coarse layer of the UNet with the low-level structured feature f_A^c and the low-resolution fine layer with the high-level textile pattern content feature f_S^c to generate different parts of the pattern images. We then use the texture defect feature f_T^c as a negative prompt, and perform conditional generation through CFG (Ho and Salimans 2022). During training, S-LDM (denoted as $\boldsymbol{\epsilon}_\theta$) requires the t-step noised latent feature z_t^p of ground truth P. We then concatenate it with the latent feature z^c of clothing image C and its resized mask M to form the input $\boldsymbol{\phi}_t^p = [z_t^p, z^c, M]$ as shown in Figure 2, where $[\cdot, \cdot, \cdot]$ denotes channel-wise concatenation. The denoising process is trained to learn the distribution of the general pattern images P as follows:

$$\mathcal{L}_{DP} = \|\boldsymbol{\epsilon}_\theta(\boldsymbol{\phi}_t^p, f_S^c, f_A^c, f_T^c, f_C^c, t) - \boldsymbol{\epsilon}\|_2^2, \tag{6}$$

where f_C^c is the feature of the frozen CLIP (Radford et al. 2021) image encoder, and $\boldsymbol{\epsilon}$ is the actual t-step noise added to the latent feature z^p of the ground truth image P.

Alignment Process. This process enables generation alignment and supports semi-supervised learning. Specifically, by our designed module and loss functions that constrain the predicted output $\hat{P}$ to be similar to either the ground truth P (if available) or the input clothing image C, we achieve supervision even without annotations. In this way, enforcing alignment with C allows the model to learn from unlabeled data, as illustrated by the *red arrows and dashed boxes* in Figure 2. The predicted $\hat{P}$ is computed as follows:

$$\hat{P} = \mathcal{D}_V\left(\frac{z_t^c - \sqrt{1-\bar{\alpha}_t} \cdot \boldsymbol{\epsilon}_\theta(\boldsymbol{\phi}_t^*, f_S^c, f_A^c, f_T^c, f_C^c, t)}{\sqrt{\bar{\alpha}_t}}\right), \tag{7}$$

where $\boldsymbol{\phi}_t^*$ can be either $\boldsymbol{\phi}_t^p$ or $\boldsymbol{\phi}_t^c = [z_t^c, z^c, M]$ (no ground truth). z_t^c is the t-step noised latent feature of C. The details of our designing module and loss functions are as follows:

(1) Stable Transformation Domain (STD) Loss. To ensure stable transformation from a clothing image C to the generated pattern $\hat{P}$ during semi-supervised training, we propose a STD loss that jointly aligns cross-sample transformation behavior and constrains intra-domain consistency. The first part of the loss compares the distributional transformation vector ($\mathbf{v}_{pred} = f_S^{\hat{p}} - f_S^c$) from C to $\hat{P}$ with a reference transformation ($\mathbf{v}_{real} = f_S^{p_r} - f_S^{c_r}$) obtained from **randomly sampled** real paired data C_r and P_r in each mini-batch, encouraging the model to mimic realistic shifts in content space even in the absence of ground truth supervision. The second part further promotes structural stability by enforcing similarity between the variations within each domain—comparing C and C_r ($\mathbf{d}_{ref} = f_S^c - f_S^{c_r}$) on the one hand, and $\hat{P}$ and P_r ($\mathbf{d}_{gen} = f_S^{\hat{p}} - f_S^{p_r}$) on the other. This dual objective prevents overfitting to synthetic transformations and encourages reliable structure retention. Here, f_S^* denotes the textile pattern content feature. The full transformation loss is:

$$\mathcal{L}_{STD} = \|\mathbf{v}_{real} - \mathbf{v}_{pred}\|_2^2 + \|\mathbf{d}_{gen} - \mathbf{d}_{ref}\|_2^2 \tag{8}$$

Together, these constraints stabilize the semi-supervised learning process and enable consistent textile pattern generation, even in the absence of paired data supervision.

(2) Convolutional Local Similarity (CLS) Module. We observe that pattern images exhibit content periodicity (local similarity), characterized by repeating contents, textures and structures. Based on this, we propose a CLS module through local region similarity maps alignment to enhance the self-similarity of the generation. Specifically, we randomly select N region matrices of size 64×64 from a real pattern image P as kernels $\mathcal{K}_i$, where $i \in [1, N]$, and compute the cosine similarity with other same size regions of itself to produce N local similarity maps, as shown in Figure 2. Then, entries in the N maps that exceed the threshold $\mathcal{T}$ are set to 1, indicating regional similarity, others are set to 0, indicating dissimilarity, yielding N binary local similarity maps W_i^p as supervision labels. Local similarity maps $W_i^{\hat{p}}$ of the predicted $\hat{P}$ are obtained in the same way. To evaluate whether P and $\hat{P}$ share similar local structures, we compare them using the Dice coefficient (Milletari, Navab, and Ahmadi 2016):

$$\mathcal{L}_{CLS} = -\sum_{i=1}^{N} \frac{2\sum(W_i^{\hat{p}} W_i^p)}{\sum W_i^{\hat{p}} + \sum W_i^p}, \tag{9}$$

where $\sum$ in the fraction denotes the element-wise sum over each map. We set $N = 4$ and $\mathcal{T} = 0.7$.

(3) Perceptual and Pixel-Level Consistency. We include LPIPS loss (Dosovitskiy and Brox 2016) and MSE loss to ensure the perceptual and pixel-level similarity between $\hat{P}$ and the ground truth P (if available) or the clothing image C. The loss function of the alignment process is defined as:

$$\mathcal{L}_{AP} = \lambda_1 \mathcal{L}_{STD} + \lambda_2 \mathcal{L}_{CLS} + \lambda_3 \mathcal{L}_{LPIPS} + \lambda_4 \mathcal{L}_{MSE}. \tag{10}$$

Here, we set $\lambda_1 = \lambda_2 = 1e^{-4}, \lambda_3 = 1e^{-2}, \lambda_4 = 1e^{-1}$. The total S-LDM training loss is defined as follows:

$$\mathcal{L}_{S-LDM} = \mathcal{L}_{DP} + \mathcal{L}_{AP}. \tag{11}$$

The S-LDM preserves the distributional stability of the pattern images and ensures content alignment with the clothing images. The details of MSE and LPIPS losses are provided in the Appendix.

Experiments

Experimental Settings

Datasets. We introduce a novel high-resolution paired dataset of clothing and textile pattern images for TPG, named the clothing textile pattern dataset (CTP-HD). It contains $9,804$ annotated pairs (clothing images with corresponding pattern ground truths) and over $10,000$ unannotated clothing images. All images are of uniform resolution 501×821 pixels, making it the first large-scale dataset for TPG. We also use the virtual try-on dataset, VITON-HD (Choi et al. 2021), for generalization testing only.

Evaluation Metrics. We use Learned Perceptual Image Patch Similarity (LPIPS) (Dosovitskiy and Brox 2016), Structural Similarity Index (SSIM) (Wang et al. 2004), and Fréchet Inception Distance (FID) (Heusel et al. 2017) to evaluate visual quality of generated images. In addition, we introduce a new metric, Fourier Periodic Similarity (FPS), to measure content periodicity based on local similarity, as detailed in the Appendix. For generalization testing without ground truth, we use LPIPS and MSE to evaluate visual similarity (VLS), and compute content similarity (CTS) using the textile pattern content feature f_S extracted from SCM.

Baseline Methods. We compare our model with representative image-to-image generation methods relevant to TPG, including image inpainting (DCI-VTON (Gou et al. 2023), Paint-by-Example (Yang et al. 2023), StrDiffusion (Liu et al. 2024)), reference-based generation (SSR-Encoder (Zhang et al. 2024), IP-Adapter (Ye et al. 2023)), UniCon (Li et al. 2025) and style transfer (StyleShot (Gao et al. 2024), OSASIS (Cho et al. 2024)). Details on the CTP-HD dataset, implementation, and baseline methods can respectively be found in the Appendix.

Generation on CTP-HD Dataset

Qualitative Comparisons. Figure 3 shows the qualitative results of different methods. Image inpainting methods often result in color distortion and misaligned pattern contours. Reference-based generation methods yield coarse, structurally inconsistent patterns due to inaccurate semantic extraction. Style transfer methods often introduces style-color artifacts (e.g. foreground–background hue swapping) and suffers from pattern collapses. By contrast, our method restores fine-grained details and faithfully reconstructs clothing appearance and well-structured overall layout. Besides, we provide more visualization results in the Appendix.

Quantitative Comparisons. Table 1 presents the quantitative results of different methods. Style transfer methods perform poorly as they fail to capture complex texture content. In addition, the significant domain gap between pattern and natural images makes modeling especially difficult under limited labeled data. Image inpainting methods enable content reconstruction but still fail to deliver satisfactory results due to feature confusion of the clothing images. Reference-based generation methods also struggle to capture complex clothing content when relying on text-driven subject extraction, image encoding, or joint distribution modeling, often resulting in inaccurate features and coarse outputs. In contrast, our method disentangles clothing features into more independent and accurate feature spaces, and improves generation quality through the alignment process, consistently outperforming baselines across all metrics.

Generalization Testing on VITON-HD Dataset

For generalization testing, we randomly sampled 1,000 images from VITON-HD and chose a representative strong baseline from each of three distinct domains for comparisons. As shown in Table 2, our method achieved the best performance in both visual similarity (VLS) and content similarity (CTS). Figure 4 presents qualitative comparisons, showing that our model effectively captures diverse clothing content and structured pattern information. The fine-grained details in the generated pattern images align closely with the input clothing images, unlike baseline methods that overlook feature confusion, resulting in visual artifacts.

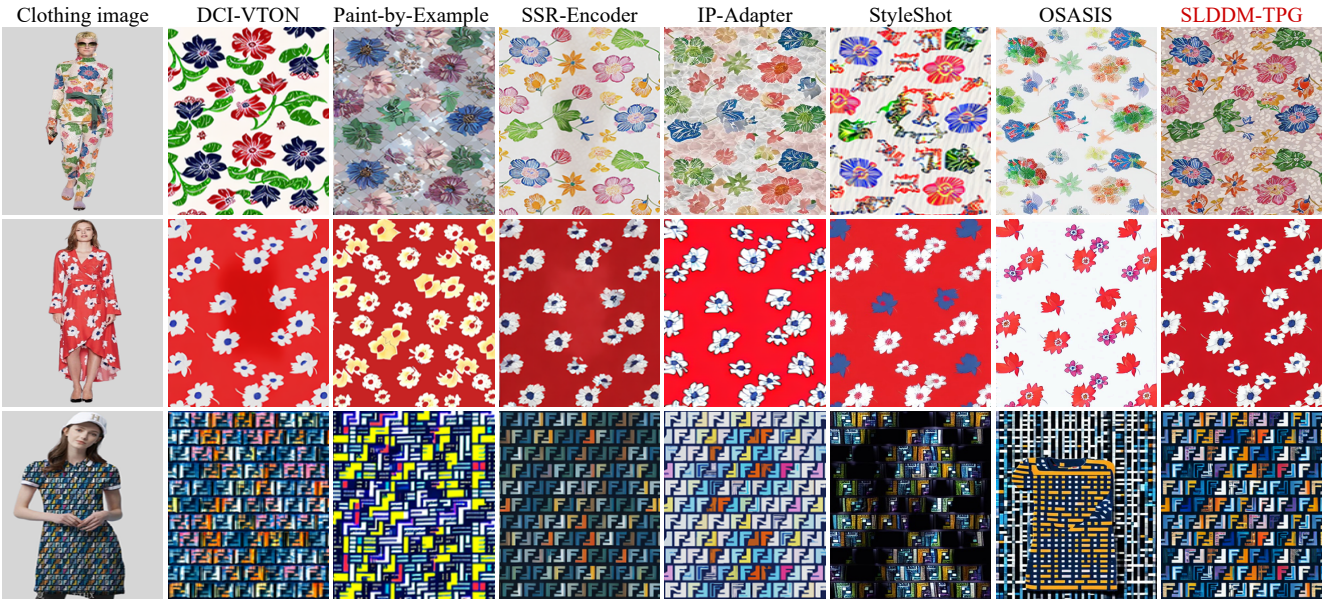


Figure 3: Qualitative comparisons of different methods on TPG on our *CTP-HD* dataset *with and without GT*.

Method	FID↓	SSIM↑	LPIPS↓	FPS↑
DCI-VTON	36.39	0.175	0.521	0.758
Paint-by-Example	37.52	0.167	0.536	0.746
StrDiffusion	39.31	0.164	0.529	0.733
SSR-Encoder	19.69	0.278	0.425	0.770
IP-Adapter	19.98	0.272	0.431	0.773
UniCon	21.98	0.270	0.431	0.768
StyleShot	39.76	0.149	0.597	0.681
OSASIS	42.36	0.140	0.619	0.651
SLDDM-TPG (Ours)	**15.59**	**0.394**	**0.280**	**0.875**
improvement	*4.10*	*0.116*	*0.145*	*0.102*

Table 1: Performance comparison for textile pattern generaion on our *CTP-HD* dataset *with Ground Truth (GT)*.

Method	LPIPS (VLS)↓	MSE (VLS)↓	CTS↑
DCI-VTON	0.621	0.139	0.554
SSR-Encoder	0.513	0.115	0.641
StyleShot	0.597	0.158	0.615
SLDDM-TPG (Ours)	**0.406**	**0.080**	**0.736**
improvement	*0.107*	*0.035*	*0.095*

Table 2: Comparisons of model generalization performance on the *VITON-HD* dataset *without GT*.

Ablation Study

Multi-dimensional Features in LDN. We conduct an ablation study to evaluate the effect of each component in our LDN for resolving feature confusion. As shown in Table 3, removing LDN, especially the textile pattern content feature f_S^c from SCM, model performance drops significantly, since this feature captures key semantic information crucial for reconstruction. The texture defect feature f_T^c from RAM and the structured feature f_A^c from SATs further refine the output by helping the model avoid defects and recover a well-structured layout. When all features are used together, the model achieves the best performance. Qualitative results in Figure 5 support this. Without disentanglement, derectly using the entangled features lead to blurry, distorted outputs, while excluding structure or defect features causes deformation of visual elements, such as distorted floral shapes, blurry texture, and color degradation. In contrast, incorporating all the disentangled features from LDN yields pattern images with finer granularity, better structure, and higher fidelity.

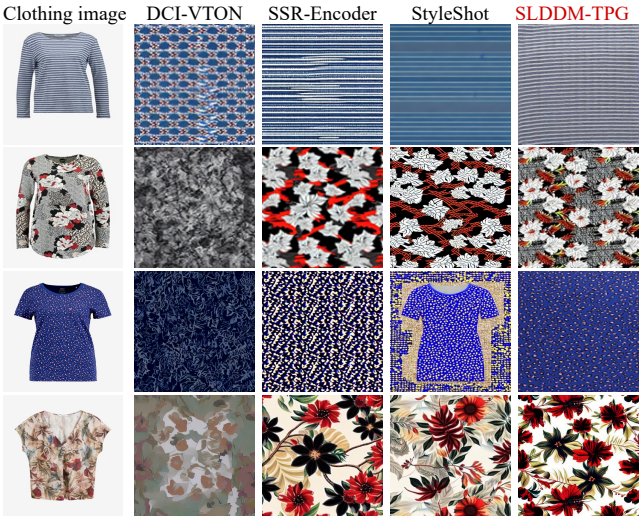


Figure 4: Generalization performance comparisons of different methods on the *VITON-HD* dataset *without GT*.

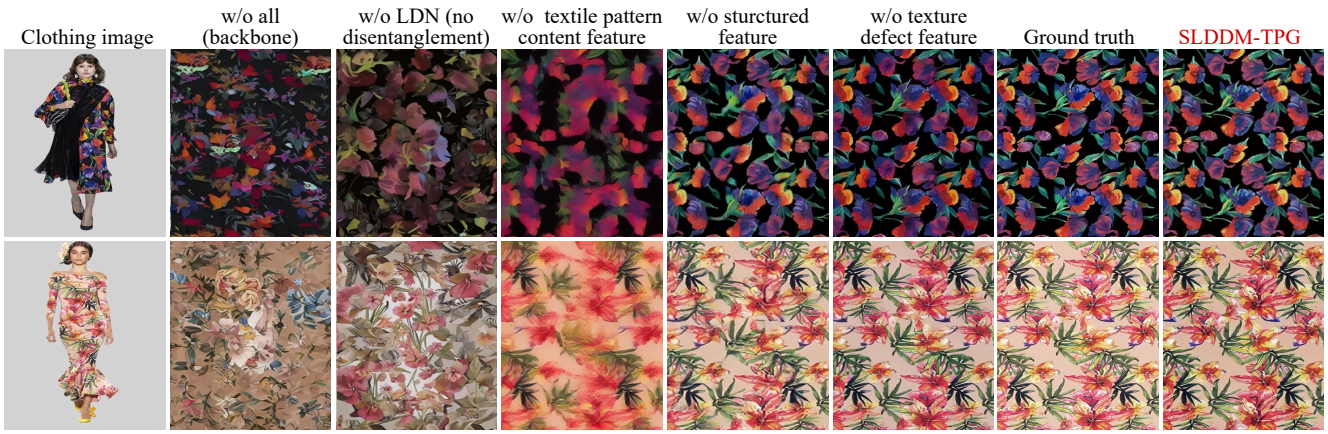


Figure 5: Qualitative ablation study on LDN using each individual feature of LDN and the original undisentangled feature.

Component	Method	FID↓	SSIM↑	LPIPS↓	FPS↑
LDN	w/o all	41.53	0.122	0.623	0.621
	w/o LDN	27.15	0.183	0.479	0.703
	w/o content feature	23.59	0.219	0.468	0.719
	w/o structured feature	16.04	0.376	0.315	0.863
	w/o defect feature	16.89	0.376	0.320	0.857
S-LDM	w/o alignment process	18.61	0.350	0.318	0.852
	w/o CLS module	16.23	0.364	0.299	0.727
	w/o STD loss	16.86	0.359	0.311	0.863
	SLDDM-TPG (Ours)	**15.59**	**0.394**	**0.280**	**0.875**

Table 3: Ablation study on the Latent Disentangled Network (LDN) and the alignment process in S-LDM.

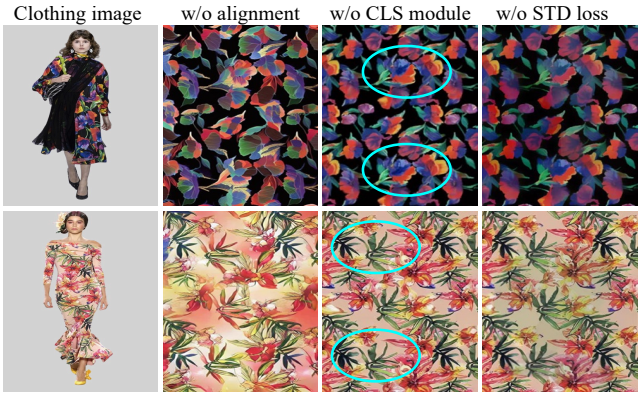
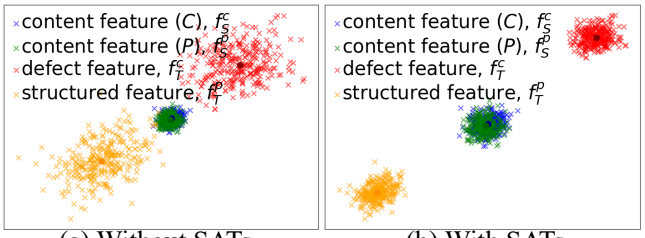


Figure 6: Qualitative ablation study on alignment process.

Feature Distribution Movement Analysis. To evaluate the effect of co-training RAM with SATs, we visualize the LDN's feature distributions with and without SATs. We quantify the distributional difference using the Euclidean distance. As shown in Figure 7, without SATs, the texture defect and structured features are separated but fail to converge, indicating drift from the latent space and loss of semantics. In contrast, with SATs, the affine transformations encourage the two features towards convergence, verifying that SATs co-training ensures proper disentanglement.



(a) Without SATs. (b) With SATs.

Figure 7: The visualization of the distribution movement of LDN's features with and without SATs.

Ablation on Alignment Process. Table 3 shows that the CLS module enhances content periodicity (local similarity) of the generated results, as quantified by FPS. Meanwhile, the two ellipses in Figure 6 highlight shape inconsistencies in regions that were expected to be locally aligned, due to the lack of the CLS module. Removing the STD loss results in fidelity degradation (e.g., color darkening), as the model loses guidance from real paired-data transformations to supervise the mapping from clothing images to generated pattern images. In contrast, incorporating STD loss preserves alignment and consistently improves generation quality. Additional experiments (user study and model efficiency test), extended visualization results, the limitation of our work and future work are provided in the Appendix.

Conclusion

We propose textile pattern generation (TPG), a new task that synthesizes textile pattern images based on clothing images, and introduce SLDDM-TPG as an effective solution. SLDDM-TPG combines a latent disentangled network (LDN) to resovle feature confusion and a semi-supervised latent diffusion model (S-LDM) to leverage both unlabeled and limited labeled data. This framework enables high-fidelity pattern generation while preserving faithful clothing details. Extensive experiments on our CTP-HD dataset and the widely used VITON-HD dataset confirm the superior performance and good generalization of our approach.

Acknowledgments

This work was sponsored by National Natural Science Foundation of China (62576305), Zhejiang Provincial Natural Science Foundation of China (LQ24F020020, LD24F020011), "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2024C01167), and the Fundamental Research Funds for the Central Universities (No. 226-2025-00057).

References

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15750–15758.

Cho, H.; Lee, J.; Chang, S.; and Jeong, Y. 2024. One-Shot Structure-Aware Stylized Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8302–8311.

Choi, S.; Park, S.; Lee, M.; and Choo, J. 2021. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14131–14140.

Choi, Y.; Kwak, S.; Lee, K.; Choi, H.; and Shin, J. 2024. Improving diffusion models for authentic virtual try-on in the wild. In *European Conference on Computer Vision*, 206–235. Springer.

Choi, Y. L.; Nam, Y.; Choi, K. M.; and Cui, M. H. 2007. A method for garment pattern generation by flattening 3D body scan data. In *Digital Human Modeling: First International Conference on Digital Human Modeling, ICDHM 2007, Held as Part of HCI International 2007, Beijing, China, July 22-27, 2007. Proceedings 1*, 803–812. Springer.

Dosovitskiy, A.; and Brox, T. 2016. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29.

Fayyaz, R. A.; Maqbool, M.; and Hanif, M. 2020. Textile design generation using GANs. In *2020 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, 1–5. IEEE.

Gao, J.; Liu, Y.; Sun, Y.; Tang, Y.; Zeng, Y.; Chen, K.; and Zhao, C. 2024. Styleshot: A snapshot on any style. *arXiv preprint arXiv:2407.01414*.

Gou, J.; Sun, S.; Zhang, J.; Si, J.; Qian, C.; and Zhang, L. 2023. 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*, 7599–7607.

He, J.; Cheng, L.; Fang, C.; Feng, Z.; Mu, T.; and Song, M. 2024. Progressive feature self-reinforcement for weakly supervised semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 2085–2093.

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.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Huang, Z.; Guo, Y.-C.; An, X.; Yang, Y.; Li, Y.; Zou, Z.-X.; Liang, D.; Liu, X.; Cao, Y.-P.; and Sheng, L. 2025. Midi: Multi-instance diffusion for single image to 3d scene generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 23646–23657.

Ju, X.; Liu, X.; Wang, X.; Bian, Y.; Shan, Y.; and Xu, Q. 2024. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*, 150–168. Springer.

Li, D.; Li, J.; and Hoi, S. 2024. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36.

Li, X.; Herrmann, C.; Chan, K. C.; Li, Y.; Sun, D.; and Yang, M.-H. 2025. A Simple Approach to Unifying Diffusion-based Conditional Generation. In *The Thirteenth International Conference on Learning Representations*.

Liu, H.; Wang, Y.; Qian, B.; Wang, M.; and Rui, Y. 2024. Structure Matters: Tackling the Semantic Discrepancy in Diffusion Models for Image Inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8038–8047.

Lu, S.; Mok, P. Y.; and Jin, X. 2017. A new design concept: 3D to 2D textile pattern design for garments. *Computer-Aided Design*, 89: 35–49.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571. Ieee.

Pang, Z.; Wu, S.; Zhang, D.; Gao, Y.; and Chen, G. 2019. NAD: Neural network aided design for textile pattern generation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2081–2084.

Preechakul, K.; Chatthee, N.; Wizadwongsa, S.; and Suwajanakorn, S. 2022. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10619–10629.

Qi, T.; Fang, S.; Wu, Y.; Xie, H.; Liu, J.; Chen, L.; He, Q.; and Zhang, Y. 2024. DEADiff: An Efficient Stylization Diffusion Model with Disentangled Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8693–8702.

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.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Voynov, A.; Chu, Q.; Cohen-Or, D.; and Aberman, K. 2023. p+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*.

Wang, S.; and Sun, Z. 2021. Dyeing creation: a textile pattern discovery and fabric image generation method. *Multimedia Tools and Applications*, 80(17): 26511–26530.

Wang, X.; Fu, S.; Huang, Q.; He, W.; and Jiang, H. 2024. MS-Diffusion: Multi-subject Zero-shot Image Personalization with Layout Guidance. *arXiv preprint arXiv:2406.07209*.

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.

Wu, X.; and Li, L. 2024. An application of generative AI for knitted textile design in fashion. *The Design Journal*, 27(2): 270–290.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 2023. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18381–18391.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Yin, Y.; Cheng, L.; Zhou, W.; Deng, J.; Yu, Z.; and Li, H. 2025. Self-Classification Enhancement and Correction for Weakly Supervised Object Detection. *arXiv preprint arXiv:2505.16294*.

Zhang, Y.; Song, Y.; Liu, J.; Wang, R.; Yu, J.; Tang, H.; Li, H.; Tang, X.; Hu, Y.; Pan, H.; et al. 2024. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8069–8078.