Accepted: 03 January 2026
Revised: 02 December 2025
Received: 04 September 2025

RESEARCH-ARTICLE

Multimodal-Conditioned Latent Diffusion Models for Fashion Image Editing

ALBERTO BALDRATI, University of Florence, Florence, FI, Italy

DAVIDE MORELLI, University of Modena and Reggio Emilia, Modena, MO, Italy

MARCELLA CORNIA, University of Modena and Reggio Emilia, Modena, MO, Italy

MARCO BERTINI, University of Florence, Florence, FI, Italy

RITA CUCCHIARA, University of Modena and Reggio Emilia, Modena, MO, Italy

Multimodal-Conditioned Latent Diffusion Models for Fashion Image Editing

ALBERTO BALDRATI*, University of Florence, Italy and University of Pisa, Italy
DAVIDE MORELLI*, University of Modena and Reggio Emilia, Italy and University of Pisa, Italy
MARCELLA CORNIA, University of Modena and Reggio Emilia, Italy
MARCO BERTINI, University of Florence, Italy
RITA CUCCHIARA, University of Modena and Reggio Emilia, Italy and IIT-CNR, Italy

Fashion illustration is a crucial medium for designers to convey their creative vision and transform design concepts into tangible representations that showcase the interplay between clothing and the human body. In the context of fashion design, computer vision techniques have the potential to enhance and streamline the design process. Departing from prior research primarily focused on virtual try-on, this paper tackles the task of multimodal-conditioned fashion image editing. Our approach aims to generate human-centric fashion images guided by multimodal prompts, including text, human body poses, garment sketches, and fabric textures. To address this problem, we propose extending latent diffusion models to incorporate these multiple modalities and modifying the structure of the denoising network, taking multimodal prompts as input. To condition the proposed architecture on fabric textures, we employ textual inversion techniques and let diverse cross-attention layers of the denoising network attend to textual and texture information, thus incorporating different granularity conditioning details. Given the lack of datasets for the task, we extend two existing fashion datasets, Dress Code and VITON-HD, with multimodal annotations. Experimental evaluations demonstrate the effectiveness of our proposed approach in terms of realism and coherence concerning the provided multimodal inputs.

CCS Concepts: • **Computing methodologies** → **Appearance and texture representations**; **Computer vision**; **Computer vision tasks**; • **Applied computing** → *Media arts*.

Additional Key Words and Phrases: Fashion Product Design, Latent Diffusion Models, Textual Inversion, Generative AI, Multimodal Learning.

1 Introduction

In recent years, the intersection of computer vision and fashion has garnered significant attention, with a surge in research mainly dedicated to adapting or re-designing state-of-the-art computer vision models for fashion images. Previous studies have primarily focused on tasks such as clothing item recognition and retrieval [8, 21, 40, 91], garment and outfit recommendation [29, 64, 84], and virtual try-on [12, 24, 48, 77, 87]. While these works have advanced research in the field, limited attention has been paid to text-conditioned fashion image editing, mainly

*Both authors contributed equally to this research.

Authors' Contact Information: Alberto Baldrati, alberto.baldrati@unifi.it, University of Florence, Florence, Italy and University of Pisa, Pisa, Italy; Davide Morelli, davide.morelli@unimore.it, University of Modena and Reggio Emilia, Modena, Italy and University of Pisa, Pisa, Italy; Marcella Cornia, marcella.cornia@unimore.it, University of Modena and Reggio Emilia, Reggio Emilia, Italy; Marco Bertini, marco.bertini@unifi.it, University of Florence, Florence, Italy; Rita Cucchiara, rita.cucchiara@unimore.it, University of Modena and Reggio Emilia, Modena, Italy and IIT-CNR, Pisa, Italy.

ACM 1551-6865/2026/1-ART
https://doi.org/10.1145/3789212

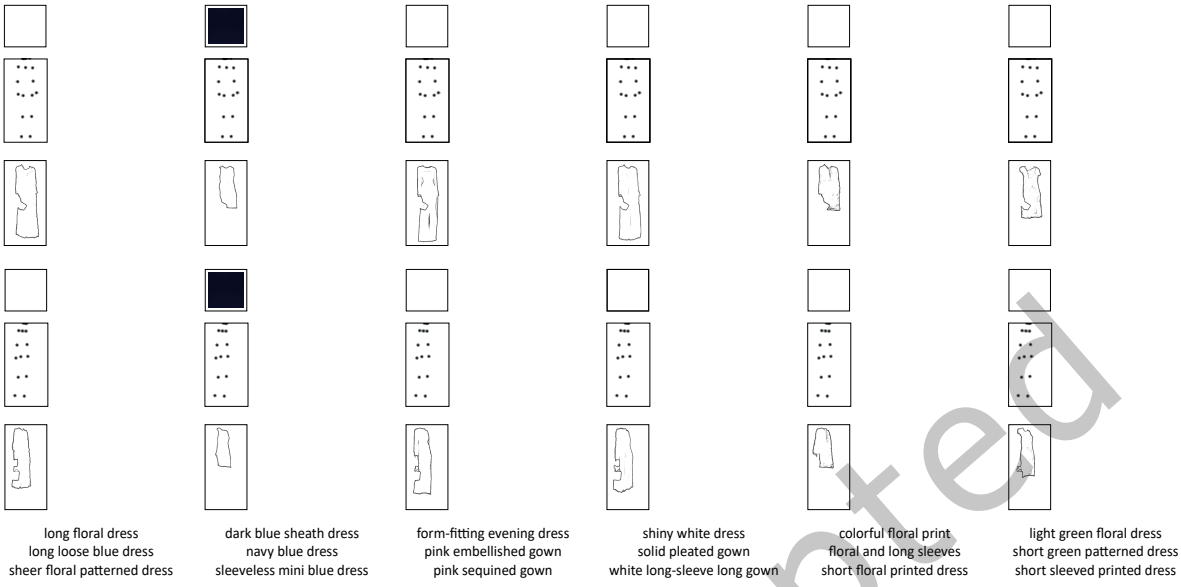

Fig. 1. Example of images generated using the proposed Textual-inverted Multimodal Garment Designer (Ti-MGD) method, with each row featuring the same model edited using different inputs. For each generated image, we show the generation input conditions: texture (top left), keypoints (middle left), sketch (bottom left), and text (bottom of each column).

due to the specificity of the fashion lexicon, the lack of existing datasets, and the complexity of the task itself. Among the few works that have addressed the task, some attempts [32, 53, 93] have been dedicated to the use of GAN-based methods to generate images of models wearing clothing items only exploiting the condition of textual descriptions. Recently, diffusion models [16, 28, 50, 59] have shown exceptional generation capabilities compared to GANs, enabling better control over the synthesized output. However, the applicability of these models to the fashion domain remains largely underexplored.

In this work, we go beyond standard text-conditioned generation and introduce *multimodal-conditioned fashion image editing*, a new challenging task that involves the generation of new garment images worn by a given person, leveraging the conditioning of multiple multimodal constraints including human pose, garment sketches, textual descriptions, and garment fabric textures. Integrating such diverse prompts is particularly complex in the fashion domain, where images vary significantly due to factors like target gender, garment category, and target market dynamics (*i.e.* whether the garment is a luxury or economical item). At the same time, this task can have a significant impact on creative industries, as it can enable fashion designers to empower the design of new fashion items, facilitating the exploration of the interplay between their sketches, the available fabric textures, and diverse human body shapes.

To tackle the newly proposed task, we present a novel approach that enables the generative process to be guided by multimodal prompts (*i.e.* text, human pose, garment sketches, and fabric textures) while preserving the identity and body shape of the subject (Fig. 1). Specifically, we leverage latent diffusion models [59], which define the forward and reverse processes in the latent space of a pre-trained autoencoder instead of the pixel space, and propose a denoising network that can be conditioned by multiple modalities, also incorporating pose consistency between input and generated images. A first attempt at fashion image editing conditioned by multimodal inputs has previously been proposed by us in [4]. Compared to the previous version, we improve the architecture by enabling it to also deal with fabric texture input while retaining the capability to remove

any constraint at inference time. In particular, we design a novel textual inversion-based component that can project texture images to the textual space of the diffusion model. We then let diverse cross-attention layers of the denoising network capture diverse granularity details, enabling simultaneous conditioning of both text and fabric textures through the same layers of the denoising network. We denote this new version as Textual-inverted Multimodal Garment Designer (Ti-MGD).

The task of multimodal-conditioned fashion image editing is new and no datasets are available both for training and testing. To effectively address the task, we also define a semi-automatic framework for extending existing fashion datasets with multimodal data. Specifically, we leverage two well-known virtual try-on datasets, Dress Code [48] and VITON-HD [11], and augment them with textual descriptions, garment sketches, and fabric texture. To evaluate the impact of conditioning signals, we introduce three novel evaluation metrics that measure human pose, sketch, and fabric texture coherence between input and generated images. Through extensive experiments on the proposed multimodal fashion benchmarks, we demonstrate the quantitative and qualitative effectiveness of our proposed approach in generating high-quality images based on multimodal inputs. As quantitative metrics and human evaluations confirm, our method outperforms state-of-the-art competitors and baselines.

In summary, our contributions are as follows:

- We propose the novel task of multimodal-conditioned fashion image editing, which utilizes multimodal prompts to guide the generative process.
- To tackle the task, we design a semi-automatic annotation framework to extend two existing fashion datasets with textual data, garment sketches, and fabric textures.
- We introduce a new human-centric generative architecture based on latent diffusion models capable of incorporating multimodal prompts while preserving the input person's characteristics. Specifically, we let the denoising network take multimodal prompts as input and design a novel textual inversion-based component that effectively integrates fabric texture by projecting texture images into the textual space of the diffusion model.
- To the best of our knowledge, we are the first to use, in a concrete working case, the property that distinct cross-attention layers of the denoising network can capture diverse granularity conditioning details. This method enables concurrent textual and texture generation conditioning by sharing the same layers.
- Extensive experiments demonstrate that our approach outperforms state-of-the-art competitors in terms of realism and input coherence in generating images with multimodal conditioning. Source code and trained models are available at: https://github.com/aimagelab/Ti-MGD.

2 Related Work

Text-Guided Image Generation. Text-to-image synthesis aims to generate images that accurately reflect a given textual prompt. Early methods were based on GANs [72, 82, 88], while recent work has shifted toward diffusion models [16, 51, 54, 57, 59]. Nichol *et al.* [51] introduced a diffusion model with local editing capabilities for handling complex prompts. Ramesh *et al.* [57] proposed a two-stage system with a prior generating CLIP embeddings [55] and a diffusion decoder. Similarly,the approach proposed in [62] leverages the T5 language model [56] followed by a cascade of super-resolution diffusion models to improve the generation process. Rather than operating in pixel space, recent approaches favor latent diffusion [59], where the diffusion process occurs in the latent space of a pre-trained autoencoder, improving both efficiency and image quality.

Only a few attempts of text-to-image synthesis have been conducted for the fashion domain [32, 53, 93]. Notably, Zhu *et al.* [93] introduced a GAN-based solution that generates the final image based on both textual data and semantic layouts. A different approach is the one presented in [53], where a latent code regularization technique is employed to enhance the GAN inversion process. This involves leveraging CLIP textual embeddings [55] to guide the image generation process. Differently, Jiang *et al.* [32] proposed to synthesize full-body images by

mapping textual descriptions of clothing items into one-hot vectors. However, this approach imposes limitations on the expressive capacity of the conditioning signal.

Multimodal Image Generation with Diffusion Models. A correlated set of studies seeks to incorporate various modalities into existing diffusion models, thereby enhancing control over the generation process [9, 10, 33, 41, 44, 49, 79, 89]. In this context, Choi *et al.* [10] proposed refining the generative mechanism of an unconditional denoising diffusion probabilistic model [50] by aligning each latent variable with a given reference image. Conversely, the approach proposed by Mang *et al.* [44] introduces noise to a stroke-based input and applies the reverse stochastic differential equation to generate images, without additional training. Instead, Wang *et al.* [79] suggested learning a deeply semantic latent space and conducting conditional fine-tuning for each downstream task to correlate guidance signals with the pre-trained space. Other recent studies have suggested incorporating sketches as additional conditioning signals, either by concatenating them with the model input [9] or by training an MLP-based edge predictor to map latent features to spatial maps [73].

Among other works proposing conditioning strategies for pre-trained latent diffusion models, Zhang *et al.* introduced ControlNet [89], which extends Stable Diffusion [59] with an additional conditioning input. This involves creating two copies of the model parameters: one fixed (locked) and one trainable. The trainable copy learns the new condition, while the locked copy retains the original model's knowledge. Conversely, the model in [49] uses modality-specific adapters to condition Stable Diffusion on new modalities. Similarly, Ye *et al.* [86] propose a lightweight adapter that enables conditioning on image prompts via the cross-attention layers of the denoising network. In contrast, our focus lies within the fashion domain, where we propose a human-centric architecture based on latent diffusion models, leveraging direct conditioning from textual sentences and other modalities like human body poses, garment sketches, and fabric textures.

Diffusion-based Generative Models for Fashion. While some approaches for in-shop garment generation using diffusion models have been explored [20, 90], most research in the fashion domain has focused on virtual try-on, where the goal is to synthesize a person wearing a given garment. To overcome the limitations of traditional garment warping inherited from GAN-based try-on systems [11, 17, 24, 77], recent works rely on pre-trained diffusion models as generative priors [19, 47, 63, 92]. Several methods adopt ControlNet-style architectures [34, 87], while others employ dual-UNet pipelines [35, 92] or exemplar-based inpainting formulations [19, 83]. StableGarment [78] and IDM-VTON [12] further refine garment alignment and detail preservation via customized attention mechanisms or garment encoders. AnyFit [39] addresses multi-garment harmonization by enabling upper–lower garment combinations.

More recent diffusion-based VTON models further refine controllability and garment preservation by introducing more expressive garment-aware conditioning. For instance, GarDiff [76] incorporates garment-focused adapters and an appearance-oriented objective to better maintain high-frequency details, while CatVTON [13] demonstrates that efficient input concatenation and lightweight adaptation suffice for competitive try-on performance. IMAGDressing-v1 [66] expands controllability by integrating garment and text features through hybrid attention, enabling scene- and prompt-conditioned dressing. Complementary to these approaches, SPM-Diff [75] formulates garment guidance through semantic point correspondences and 3D-aware cues to improve structural alignment, and FitDiT [31] leverages DiT-based models enriched with texture- and frequency-aware modules to strengthen both fidelity and size-aware fitting.

Although these approaches achieve impressive realism, they all require the garment image as input, inherently tying them to virtual try-on scenarios. In contrast, our model generates full-body fashion images directly from textual descriptions, human poses, garment sketches, and fabric textures, enabling creative and design-oriented applications beyond try-on.

Textual Inversion. Textual inversion, as introduced in the recent work by Gal *et al.* [18], is a novel technique aimed at learning pseudo words within the embedding space of a text encoder to represent visual concepts

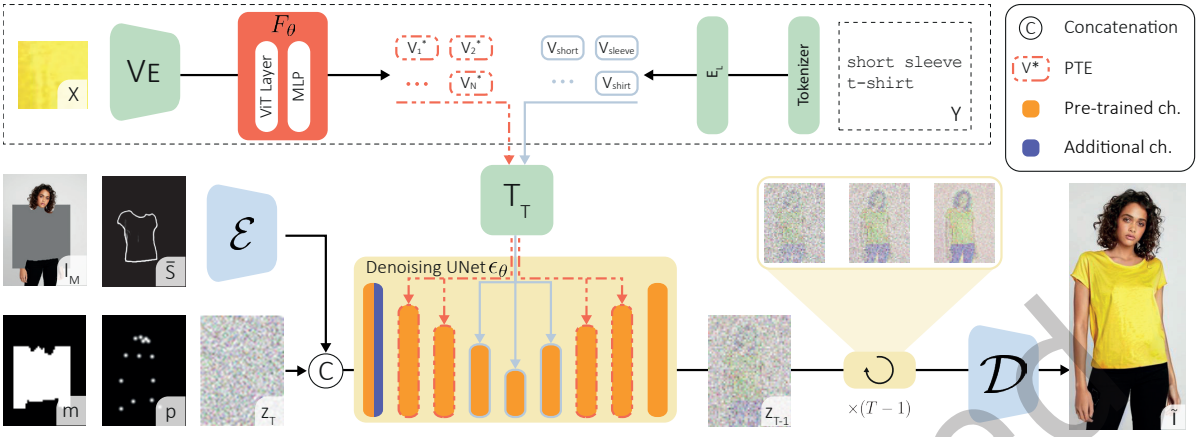

Fig. 2. Overview of the proposed Textual-inverted Multimodal Garment Designer (Ti-MGD) approach, a human-centric latent diffusion model conditioned on multiple modalities, including text, human pose, garment sketch, and fabric texture. The denoising UNet ϵ_θ takes as input the latent variable z_T and the spatial conditioning inputs (*i.e.* encoded masked model $\mathcal{E}(I_M)$, inpainting mask m, body keypoints p, and encoded sketch $\mathcal{E}(\bar{S})$). We incorporate text conditioning Y using Stable Diffusion cross-attention capabilities, extending this mechanism to condition the generated image on the texture image X by projecting it into the CLIP pseudo-word token embedding space. For this purpose, we utilize distinct cross-attention layers dedicated to text and texture conditioning.

effectively. Building on this, several promising methods have been developed for personalized image generation and editing [14, 22, 46, 61]. Among them, Ruiz *et al.* [61] specifically introduced a fine-tuning technique that associates an identifier with a subject represented by a few images, incorporating a class-specific prior preservation loss to address language drift. Similarly, Kumari *et al.* [37] proposed an alternative fine-tuning method for enabling multi-concept composition, demonstrating that updating only a small subset of model weights suffices to integrate new concepts. Instead, Han *et al.* [22] decomposed the CLIP [55] embedding space based on semantics, facilitating image manipulation without the need for further fine-tuning. In this work, we adapt textual inversion techniques to effectively condition latent diffusion models on garment fabric textures.

3 Proposed Method

This section proposes a novel task to automatically edit a human-centric fashion image conditioned on multiple modalities. Specifically, given the model image $I \in \mathbb{R}^{H \times W \times 3}$, its pose map $P \in \mathbb{R}^{H \times W \times 18}$ where each channel represent a human keypoint, a textual description Y of a garment, a sketch of the same $S \in \mathbb{R}^{H \times W \times 1}$, and a sample image of a fabric texture $X \in \mathbb{R}^{H_X \times W_X \times 3}$, we want to generate a new image $\tilde{I} \in \mathbb{R}^{H \times W \times 3}$ that retains the information of the input model while substituting the target garment according to the multimodal inputs. To tackle the task, we propose a novel latent diffusion approach, denoted as Textual-inverted Multimodal Garment Designer (Ti-MGD), that effectively combines multimodal information when generating the new image $\tilde{I}$.

To the best of our knowledge, this is the first proposed approach in literature to constrain fashion image editing on text, pose, sketch, and fabric texture. We strongly believe this task can foster research in the field and enhance the design process of new fashion items with greater customization. An overview of our model is shown in Fig. 2.

3.1 Preliminaries

Stable Diffusion. While diffusion models [67] are latent variable architectures that work in the same dimensionality of the data (*i.e.* in the pixel space), latent diffusion models (LDMs) [59] operate in the latent space of a

pre-trained autoencoder achieving higher computational efficiency while preserving the generation quality. In our work, we leverage the Stable Diffusion model [59], a text-to-image implementation of LDMs, as a starting point to perform multimodal conditioning for human-centric fashion image editing. Stable Diffusion is composed of an autoencoder with an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$, a text-time-conditional UNet denoising model ϵ_θ, and a CLIP-based text encoder T_E taking as input a text Y. The encoder $\mathcal{E}$ compresses an image I into a lower-dimensional latent space defined in $\mathbb{R}^{h\times w\times 4}$, where $h = H/8$ and $w = W/8$. The decoder $\mathcal{D}$ performs the opposite operation, decoding a latent variable into the pixel space. For the sake of clarity, we define the ϵ_θ convolutional input (*i.e.* z_t in this case) as spatial input γ, because of the property of convolutions to preserve the spatial structure, and the attention conditioning input as ψ. The denoising network ϵ_θ is trained according to the following loss:

$$L = \mathbb{E}_{\mathcal{E}(I),Y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(\gamma,\psi)\|_2^2\right], \tag{1}$$

where t is the diffusing time step, $\gamma = z_t$, $\psi = [t; T_E(Y)]$, and $\epsilon \sim \mathcal{N}(0,1)$ is the Gaussian noise added to $\mathcal{E}(I)$.

CLIP. This vision-language model [55] aligns visual and textual inputs in a shared embedding space. In particular, CLIP consists of a visual encoder V_E and a text encoder $T_E = E_L \circ T_T$, where E_L is the embedding lookup layer, which maps each tokenized word of Y to the token embedding space $\mathcal{W}$, and T_T is the CLIP text Transformer that maps the token embedding features to the CLIP shared embedding space. CLIP extracts feature representations $V_E(I) \in \mathbb{R}^d$ and $T_E(Y) \in \mathbb{R}^d$ for an input image I and its corresponding text caption Y, respectively. Here, d is the size of the CLIP shared embedding space.

The proposed approach introduces a novel textual inversion technique to generate a representation of the fabric texture X. We feed this representation to the CLIP text Transformer T_T to condition the diffusion process. It consists in mapping the visual features of X into a set of N new token embeddings $V_n^* \in \mathcal{W}, n = \{1,\dots,N\}$. Following the terminology introduced in [3], we refer to these embeddings as Pseudo-word Tokens Embeddings (PTEs) since they do not correspond to any linguistically meaningful entity but rather are a representation of the fabric texture visual features in the token embedding space $\mathcal{W}$.

3.2 Human-Centric Image Editing

The proposed task aims to generate a new image $\tilde{I}$, by replacing the target garment in the input image I using multimodal inputs while preserving the model's identity and physical characteristics. As a natural consequence, this task can be identified as a particular type of conditional inpainting tailored for human body data. Instead of using a standard text-to-image model, we perform inpainting concatenating along the channel dimension of the denoising network input z_t an encoded masked image $\mathcal{E}(I_M)$ and the relative resized binary inpainting mask $m \in \{0,1\}^{h\times w\times 1}$, which stems from the original inpainting mask $M \in \{0,1\}^{H\times W\times 1}$. Since here, the spatial input of the denoising network is $\gamma = [z_t; m; \mathcal{E}(I_M)], \gamma \in \mathbb{R}^{h\times w\times 9}$.

To give users more precise control over the generation of garments, we propose extending the input capabilities of the denoising UNet by enabling constraints on multiple modalities. Essentially, we exploit spatial information such as pose and sketch to feed into the UNet spatial input γ, while we inject semantic information such as textual descriptions and fabric textures as attention conditioning input ψ. This allows for a more refined and accurate garment generation process.

The fully convolutional nature of the encoder $\mathcal{E}$ and the decoder $\mathcal{D}$ allows LDM-based architectures to preserve the spatial information in the latent space. Our method can thus optionally add conditioning constraints to the generation by exploiting this feature. In particular, we propose to add two spatial generation constraints: the model pose map P to preserve the original human pose of the input model and the garment sketch S to condition the shape of the generated garment. In addition, we leverage the Stable Diffusion textual information conditioning mechanism for two purposes: condition on plain text and condition on fabric texture information. While the

former is intrinsic in the Stable Diffusion model by design, we propose a novel forward-only textual inversion method to tackle the latter without adding additional parameters in the denoising network.

Pose Map Conditioning. In most cases [38, 43, 71], inpainting is performed with the objective of either removing or entirely replacing the content of the masked region. However, in our task, we aim to remove all information regarding the garment worn by the model while preserving the model's body information and identity. Thus, we propose to improve the garment inpainting process by using the bounding box of the segmentation mask along with pose map information representing human body keypoints. This approach enables the preservation of the model's physical characteristics in the masked region while allowing the inpainting of garments with different shapes. Differently from conventional inpainting techniques, we focus on selectively retaining and discarding specific information within the masked region to achieve the desired outcome. To enhance the performance of the denoising network with human body keypoints, we modify the first convolution layer of the network by adding 18 additional channels, one for each keypoint. Adding new inputs usually would require retraining the model from scratch, thus consuming time, data, and resources, especially in the case of data-hungry models like the diffusion ones. Therefore, we propose extending the kernels of the pre-trained input layer of the denoising network and retraining the whole network. This consistently reduces the number of training steps, allowing training with less data. We extend these kernels using zero-initialized weights [89], which allows us to retain the knowledge embedded in the original denoising network while enabling the model to deal with the newly proposed inputs. Our experiments show that such improvement enhances the consistency of the body information between the generated in-painted region and the original image.

Incorporating Sketches. Fully describing a garment using only textual descriptions is a challenging task due to the complexity and ambiguity of natural language. While text can convey specific attributes of a garment, like style and color, it may not provide sufficient information about its spatial characteristics, such as shape and size. This limitation can hinder the customization of the generated clothing item other than the ability to match the user's intended style accurately. Therefore, we propose to leverage garment sketches to enrich the textual input with additional spatial fine-grained details. We achieve this following the same approach described for pose map conditioning. The final spatial input of our denoising network is $\gamma = [z_t; m; \mathcal{E}(I_M); p; s], [p; s] \in \mathbb{R}^{h \times w \times (18+4)}$, where p is obtained by resizing P to match the latent space dimensions, while $s = \mathcal{E}(\bar{S})$ in which $\bar{S}$ is the sketch S repeated along the channel dimension to match the $\mathcal{E}$ input channel shape. In the case of sketches, we only condition the early steps of the denoising process as the final steps have little influence on the shapes [2].

Adding Texture. While text conditioning can provide a high-level constraint over the generated garment style, it still misses the ability to express the high-frequency visual details of the garment fabric. This requirement is fundamental to give the user fine-grained control over the garment generation. We propose to enable the model to generate a garment coherent with a user-given fabric texture sample, denoted as X.

Starting with a given fabric texture sample image X, our objective is to condition the generation of the LDM utilizing the non-constrained receptive field of the attention mechanism. As the texture sample lacks spatial information and is intended to serve as a pattern for the generated garments, we propose using the existing cross-attention layers originally trained for textual conditioning, thus avoiding additional layers in the denoising network. To this aim, starting from a given fabric texture sample image X, we leverage a forward-only textual inversion technique to predict a set of fine-grained Pseudo-word Token Embeddings (PTEs) describing the fabric texture X itself. These PTEs are processed by the CLIP text transformer T_T to generate feature vectors that can condition the diffusion model generation. In particular, we feed a given fabric texture sample image X to the CLIP visual encoder V_E and extract the features of its last hidden layer. We learn to project these features into the CLIP token embedding space $\mathcal{W}$ as a set of PTEs $V^* = \{V_1^*, \dots, V_N^*\}$. This is achieved by training a textual inversion adapter module F_θ. The overall mathematical formulation is as follows:

$$V^* = \{V_1^*, \dots, V_N^*\} = F_\theta(V_E(X)). \tag{2}$$

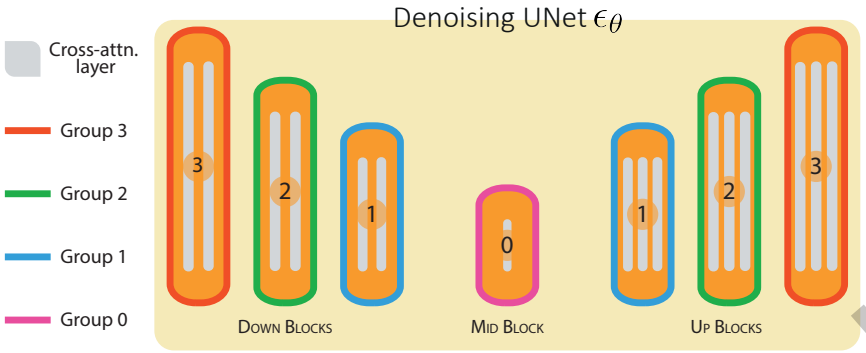

Fig. 3. Detail of cross-attention layers of the denoising network, that are categorized into four groups based on spatial resolution. Group 3 contains the highest-resolution layers, while Group 0 comprises the lowest-resolution ones.

We then use the predicted PTEs V^* to condition the Stable Diffusion denoising network ϵ_θ and obtain the final image $\tilde{I}$ where the model in I is wearing the garment filled with the texture X. For clarity, a set of PTEs represents a fabric texture well if the model conditioned on the predicted pseudo-words can reconstruct the fabric texture of the target image itself.

We leverage the intuition in [74] where distinct Stable Diffusion cross-attention layers capture diverse granularity conditioning details, introducing an innovative approach. Our method enables concurrent textual and texture generation conditioning by leveraging the inherent capabilities of the existing Stable Diffusion layers. Importantly, this strategy avoids the introduction of extra parameters, ensuring a streamlined and efficient process. To the best of our knowledge, this study marks the first instance in which a textual inversion approach is used for texture conditioning in the fashion image generation domain. The proposed approach diverges from conventional textual inversion methods like [18, 37, 61]. Instead of iteratively optimizing pseudo-word token embeddings, our solution trains the adapter F_θ to generate these embeddings in a single forward pass.

3.3 Training and Inference

Following the standard LDM approach, the proposed denoising network predicts the noise added stochastically to the encoded input, $z = \mathcal{E}(I)$. The objective function can be specified as

$$L = \mathbb{E}_{\mathcal{E}(I),Y,\epsilon\sim\mathcal{N}(0,1),t,\mathcal{E}(I_M),m,p,s,V^*}\left[\|\epsilon - \epsilon_\theta(\gamma,\psi)\|_2^2\right], \tag{3}$$

where $\gamma = [z_t; m; \mathcal{E}(I_M); p; s]$ and $\psi = [t; T_E(Y); T_T(V^*)]$.

Classifier-Free Guidance. Classifier-free guidance is a technique in which the denoising network works both conditionally and unconditionally. This procedure adjusts the final predicted noise of the model so that it moves from the predicted unconditional noise toward the direction of the predicted conditioned one. Given the time step t and a generic condition c, the predicted diffusion process follows the below equation:

$$\hat{\epsilon}_\theta(z_t|c) = \epsilon_\theta(z_t|\emptyset) + \alpha \cdot (\epsilon_\theta(z_t|c) - \epsilon_\theta(z_t|\emptyset)), \tag{4}$$

where $\epsilon_\theta(z_t|c)$ is the predicted noise at time t given the condition c and $\epsilon_\theta(z_t|\emptyset)$ is the predicted noise at time t given the null condition. The guidance scale α is a hyperparameter that controls the degree of extrapolation towards the condition.

We use the fast variant multi-condition classifier-free guidance proposed in [1] to speed up the inference time while dealing with multiple input conditions (*i.e.* text, pose map, sketch, fabric texture). Instead of performing the classifier-free guidance according to each condition independently, the fast classifier-free guidance computes the

direction considering all the conditions jointly $\Delta^t_{\text{joint}} = \epsilon_\theta(z_t|\{c_i\}_{i=1}^{i=K}) - \epsilon_\theta(z_t|\emptyset)$:

$$\hat{\epsilon}_\theta(z_t|\{c_i\}_{i=1}^{i=K}) = \epsilon_\theta(z_t|\emptyset) + \alpha \cdot \Delta^t_{\text{joint}}. \tag{5}$$

where K is the number of conditioning prompts. This reduces the number of feed-forward executions from $K+1$ to 2.

Unconditional Training. To enhance the performance of the denoising model with and without specific conditions, we randomly drop them at training time. This method enables the model to adapt to both conditional and unconditional samples, enhancing mode coverage and sample fidelity. Additionally, it allows for the optional use of conditioning signals at inference time. Since our approach considers several conditioning signals, we propose to mask each condition independently. Our experiments demonstrate that adjusting the amount of masked data can significantly improve the output quality.

Inference Modality-Aware Group Conditioning. In our task, we want to condition the generation on multiple prompts. As already stated, we cluster the input prompts in two groups: inputs with spatial information γ, that we feed to the denoising network convolutional input, and attention conditioning input ψ, that contain only semantic information and that leverage the cross-attention conditioning. Since the fabric texture does not contain spatial information, we categorize it as an attention conditioning input ψ. We leverage the existing cross-attention blocks originally trained for textual conditioning to avoid adding additional parameters in the denoising UNet and reduce the computational load (*i.e.* layers inside the denoising network ϵ_θ are executed T times during inference, where T is the number of the denoising steps).

Our idea builds upon the intuition that different cross-attention layers in the Stable Diffusion denoising UNet process the input prompts differently according to the layer resolution [74]. More in detail, higher resolution layers (*i.e.* external layer in the UNet architecture) capture small-level details, while lower-resolution layers (*i.e.* internal layers) capture more coarse information, like shapes. Therefore, we propose to condition the generation using the fabric texture information in the higher-resolution layers and textual information in the lower-resolution ones. This allows the condition of the generation on both textual and texture prompts without losing input information. We experimentally show that leveraging the fabric texture conditioning in each cross-attention layer leads to comparable results to conditioning only external ones. Specifically, given the denoising network ϵ_θ, we categorize its attention layers into four groups, as illustrated in Fig. 3. We name Group 3 the cross-attention layers with the highest resolution (the outermost layers) and sequentially assign lower group numbers down to Group 0 as the resolution decreases. Group 0 comprises the lowest resolution cross-attention layers (the innermost layers).

To maintain the flexibility of independently conditioning the generation on each modality (*e.g.* exclusively on texture or text), we adopt a training strategy involving prompt-exclusive conditioning alternation across samples. In other words, each sample is trained using either exclusive text conditioning or texture conditioning across all cross-attention layers.

4 Collecting Multimodal Fashion Datasets

Current fashion image generation datasets often feature low-resolution images and lack the necessary multimodal information for the task we want to address. Therefore, creating new multimodal datasets is essential for advancing research in the fashion domain. To this aim, we start from two recent high-resolution fashion datasets, Dress Code [48] and VITON-HD [11], used for virtual try-on, and extend them by adding textual descriptions, garment sketches, and fabric textures. Both datasets contain image pairs with a resolution of 1024×768, each composed of a garment image, a corresponding model image wearing it, and 18 person keypoints extracted with OpenPose [7]. In this section, we present a framework for semi-automatically adding multimodal information to fashion images, detailing how we extend the Dress Code and VITON-HD datasets with multimodal annotations. The extended

Fig. 4. Sample images and multimodal data from our newly collected Dress Code Multimodal and VITON-HD Multimodal datasets.

versions of these datasets are named Dress Code Multimodal and VITON-HD Multimodal, respectively. Examples of images and multimodal data from these datasets are shown in Fig. 4.

4.1 Dataset Collection and Annotation

Data Preparation. We start to construct our dataset by annotating over 53k model-garment pairs from Dress Code [48]. Inspired by fashion-specific linguistic structures [5], we annotate each garment using short, informative noun chunks (*e.g.* "striped midi skirt", "denim jacket") that concisely capture garment characteristics. To facilitate scalable annotation, we adopt a semi-automatic pipeline. First, we extract noun chunks from lemmatized captions collected from FashionIQ [81] and Fashion200k [23], yielding 60k+ unique phrases across three garment categories. We then use multiple vision-language models (CLIP [55], OpenCLIP [80]) to match images with relevant noun chunks via cosine similarity and prompt ensembling.

Data Annotation. To ensure quality and diversity, we manually verify and refine automatic annotations. For 26,400 garments (8,800 per category), annotators select the top-3 most accurate noun chunks from a set of 25 candidates or insert custom phrases using a dedicated user interface. To annotate the rest of the dataset, We fine-tune OpenCLIP ViT-B/32 on our validated image-text pairs to improve generalization. We then apply the fine-tuned model to label the remaining Dress Code items and upper-body garments in VITON-HD.

Sketch Extraction. To supplement text with visual detail, we extract garment sketches using PiDiNet [70]. For unpaired virtual try-on settings, we warp the in-shop garment to the target pose using a thin-plate spline transformation [58] refined by a UNet [60], ensuring sketch alignment (see Appendix A for technical details).

Fabric Texture Extraction. To enable users precise control over garment generation, including the fabric texture samples in the dataset is crucial. Given an in-shop garment C and its garment mask M_C, we extract fabric textures leveraging a sliding window mechanism. For each in-shop garment C and its corresponding mask M_C, we extract fabric textures using a sliding window of 128×128 pixels, selecting only patches X fully within the garment mask M_C. To prevent patch redundancy, we employ a stride of $\frac{128}{2} = 64$ pixel horizontally and vertically. We use high-resolution dataset images (*i.e.* 1024×768 pixel) for this process. When the algorithm cannot find a suitable texture patch (*e.g.* mostly in short pants), we reduce the window size to 64×64 pixels to guarantee at least one patch X for each garment C.

Table 1. Comparison of Dress Code Multimodal and VITON-HD Multimodal with other fashion datasets featuring multi-modal annotations. Here T stands for Text, P for Pose, S for Sketch, and F for Fabric texture.

Dataset	T	P	S	F	# Images	# Products	# Unique Texts	# Unique Words
VITON-HD [11]	✗	✓	✗	✗	27,358	13,679	-	-
Dress Code [48]	✗	✓	✗	✗	107,584	53,792	-	-
Be Your Own Prada [93]	✓	✓	✗	✗	78,979	N/A	3,972	445
DF-Multimodal [32]	✓	✓	✗	✗	44,096	N/A	10,253	77
VITON-HD Multimodal	✓	✓	✓	✓	27,358	13,679	5,143	1,613
Dress Code Multimodal	✓	✓	✓	✓	107,584	53,792	25,596	2,995

4.2 Comparison with Other Datasets

The only two text-to-image generation datasets in the fashion domain, referenced in [93] and [32], both utilize images from the DeepFashion dataset [40]. The dataset from [93] includes brief textual descriptions, while DeepFashion-Multimodal [32] features attributes (*e.g.* category, color, fabric) for crafting longer captions. In Table 1, we compare the textual annotation statistics of these publicly available datasets with our newly extended datasets. Along with the number of images and fashion products, we report the number of unique textual items, either noun chunks or textual sentences and the number of unique words excluding stop words and punctuation. Notably, our datasets exhibit a greater diversity in textual items and words, validating the effectiveness of our annotation approach and facilitating more customized control over the generation process. It is also important to note that the other datasets lack in-shop garment images, which limits their utility in our setting making it impossible to extract garment sketches for an unpaired and more realistic setting.

5 Experimental Evaluation

5.1 Implementation Details and Competitors

Training and Inference. All models are trained on the original splits of the Dress Code [48] and VITON-HD [11] datasets using a single NVIDIA A100 GPU. Specifically, Dress Code contains around 48k training items and 5,400 test ones, instead VITON-HD is divided into 11,647 and 2,032 products respectively belonging to the training and test set.

In all experiments, we use an image resolution of 512×384. When trained on Dress Code Multimodal, models undergo 200k training steps, while for VITON-HD Multimodal, they are trained for 75k steps. As diffusion model, we use Stable Diffusion inpainting v2[1]. To ensure a fair comparison with other models, we also develop a version of Ti-MGD based on Stable Diffusion inpainting v1[2]. During training, we use a batch size of 16 and a learning rate of 10^{-5}, with a linear warm-up in the first 500 iterations. AdamW [42] is employed as optimizer, with a weight decay of 10^{-2}. To speed up training and save memory, we use mixed precision [45]. We set the unconditional portion of data during training and the sketch conditioning rate during inference to 0.2 each.

Training involves using textual conditions half of the time and texture conditioning for the remaining half, allowing the network to adapt to both independently. At inference time, when we leverage both texture and textual conditions, we use textual features to condition Groups 0 and 1, while Groups 2 and 3 are conditioned with texture features, following the notation of Fig. 3. The textual inversion network F_θ comprises a single ViT layer and an MLP projection. The MLP includes three fully connected layers, each separated by GELU non-linearity [25] and a dropout layer [69]. The network outputs $N = 16$ PTEs. As the visual encoder V_E, we use OpenCLIP

[1]https://huggingface.co/stabilityai/stable-diffusion-2-inpainting
[2]https://huggingface.co/runwayml/stable-diffusion-inpainting

ViT-H/14 [80], pre-trained on the English portion of the LAION-5B dataset [65]. We use the DDIM [68] with 50 steps as our noise scheduler during inference, setting the classifier-free guidance parameter α to 7.5. To improve the high-frequency details in the region outside the inpainting area, we leverage the EMASC module defined in [47].

Baselines and Competitors. To ensure fair comparisons between our model and competitors, we train a version of our model using the same backbone of the competing approaches and compare results against approaches specialized on different subsets of modalities. For text-only inputs, we compare Ti-MGD with the Stable Diffusion inpainting pipeline available on Huggingface[2]. In scenarios involving text and pose inputs, Ti-MGD is compared with Stable Diffusion v1.5 integrated with ControlNet [89] for pose conditioning[3]. For inputs of text, pose, and sketch, we set Ti-MGD against an adapted version of SDEdit [44] and Stable Diffusion v1.5 integrated with ControlNet with pose and sketch adapters[4]. Specifically for SDEdit, we follow the approach in [44], using our model trained with only text and human poses and guiding the shape with a noise-added sketch image as the starting latent variable, setting the strength parameter to 0.9. For completeness, we also include the results of the previous version of our model, *i.e.* MGD [4]. For the full input set modalities (*i.e.* text, pose, sketch, and texture), we employ ControlNet for text, pose, and sketch, and the IP-Adapter [86][5] for texture, as ControlNet handles only inputs with spatial information. We set the conditioning scale for all ControlNet networks at 0.5 and condition on sketches for only the first 0.2 fraction of denoising steps. The IP-Adapter scale is set to 0.8. Note that our proposed methods and IP-Adapter both leverage OpenCLIP ViT-H/14 as the visual encoder.

Table 2. Quantitative results of Ti-MGD against competitors on the Dress Code Multimodal and VITON-HD Multimodal datasets for both paired and unpaired settings. When analyzing the modalities, T stands for Text, P for Pose, S for Sketch, and F for Fabric texture. Best results are in bold, second best are underlined.

	Modalities				Dress Code Multimodal						VITON-HD Multimodal					
Model	**T**	**P**	**S**	**F**	**FID ↓**	**KID ↓**	**CLIP-S ↑**	**PD ↓**	**SD ↓**	**TS ↑**	**FID ↓**	**KID ↓**	**CLIP-S ↑**	**PD ↓**	**SD ↓**	**TS ↑**
Paired setting																
SD v1.5 [59]	✓				19.14	11.0	28.55	9.20	0.372	0.490	16.14	5.38	29.58	11.04	0.405	0.507
ControlNet [89]	✓	✓			17.71	9.87	28.88	7.68	0.353	0.491	16.65	5.73	29.46	8.10	0.387	0.509
SDEdit [44]	✓	✓	✓		7.17	2.97	30.19	5.47	0.263	0.545	12.22	4.09	28.63	6.51	0.307	0.570
ControlNet [89]	✓	✓	✓		24.40	15.39	27.88	7.85	0.354	0.476	22.06	10.02	28.71	8.21	0.377	0.484
ControlNet+IP [86, 89]	✓	✓	✓	✓	16.02	7.79	28.22	8.12	0.363	0.545	14.98	5.52	28.61	8.12	0.354	0.574
MGD [4]	✓	✓	✓		5.74	2.11	**31.68**	4.72	0.188	0.571	10.60	3.26	**32.39**	5.94	0.243	0.587
Ti-MGD (SDv1)	✓	✓	✓	✓	**3.46**	**0.66**	31.26	**4.34**	0.176	0.578	6.14	0.78	31.16	4.72	**0.179**	0.616
Ti-MGD (SDv2)	✓	✓	✓	✓	3.79	0.99	31.01	**4.34**	**0.172**	**0.595**	**6.04**	**0.63**	31.30	**4.67**	0.187	**0.624**
Unpaired setting																
SD v1.5 [59]	✓				21.77	12.9	27.15	10.00	0.492	0.476	17.87	6.37	27.73	11.81	0.588	0.498
ControlNet [89]	✓	✓			20.16	11.62	27.60	8.39	0.469	0.481	19.17	7.47	27.60	8.83	0.557	0.495
SDEdit [44]	✓	✓	✓		8.79	3.67	27.65	**6.13**	0.354	0.532	15.14	5.99	24.95	7.10	0.446	0.559
ControlNet [89]	✓	✓	✓		26.66	17.33	26.65	8.51	0.462	0.469	23.84	11.92	26.93	8.88	0.547	0.480
ControlNet+IP [86, 89]	✓	✓	✓	✓	17.79	8.89	27.04	8.84	0.475	0.534	17.73	7.25	26.64	8.84	0.507	0.561
MGD [4]	✓	✓	✓		7.73	2.82	**30.04**	6.79	0.342	0.554	12.81	3.86	**30.75**	7.22	0.331	0.578
Ti-MGD (SDv1)	✓	✓	✓	✓	5.69	1.33	29.44	6.19	0.222	0.577	10.18	1.96	28.56	6.59	**0.239**	0.608
Ti-MGD (SDv2)	✓	✓	✓	✓	**5.68**	**1.32**	29.78	6.26	**0.218**	**0.597**	**9.30**	**1.26**	29.43	**6.56**	0.247	**0.630**

[3]https://huggingface.co/lllyasviel/control_v11p_sd15_openpose
[4]https://huggingface.co/lllyasviel/control_v11p_sd15_softedge
[5]https://huggingface.co/h94/IP-Adapter

5.2 Evaluation Metrics

To evaluate the realism of generated images, we use the Fréchet Inception Distance (FID) [27] and the Kernel Inception Distance (KID) [6], following the implementation proposed in [52]. For assessing how well the images adhere to textual conditioning, we apply the CLIP Score (CLIP-S) [26] from the TorchMetrics library [15], using the OpenCLIP ViT-H/14 model as cross-modal architecture. We compute the score on the inpainted region of the generated output pasted on a 224×224 white background. Additionally, we employ three evaluation metrics to assess the adherence of the generated image with respect to pose, sketch, and texture modalities.

Pose Distance (PD). We introduce a novel pose distance metric to assess the consistency of human body poses between original and generated images by measuring the distance between the keypoints. Specifically, we employ OpenPifPaf [36] and compute the ℓ_2 distance between each pair of real-generated corresponding estimated keypoints. This metric focuses only on the keypoints within the generation mask M and adjusts each keypoint distance based on the confidence scores from the detector to account for possible estimation inaccuracies.

Sketch Distance (SD). To quantify the adherence to the sketch constraint, we propose a novel sketch distance metric. We first segment the generated garments using an off-the-shelf clothing segmentation network[6]. Then, we paste the segmented garment area onto a white background (512×384) and use the PIDInet [70] edge detector to extract sketches. The final score is the mean squared error between the generated and input sketch S, weighting each result by the inverse frequency of the activated pixels in S to ensure a fair comparison. We avoid sketch thresholding to ensure a more effective comparison with hand-drawn grayscale sketches, enhancing the evaluation of sketch-guided image generation methods.

Texture Score (TS). We introduce a new metric to assess how well the generated garment matches the input fabric texture. This similarity is determined by extracting and comparing visual features from the input patch and the generated garment texture. We use the same segmentation network as in the sketch distance calculation to isolate the garment information. Then, we crop a 64×64 portion of the image to represent the texture of the generated garment. The adherence of the generated image texture to the input is evaluated using the CLIP cosine similarity with the OpenCLIP ViT-H/14 model, the same model used for computing the CLIP score.

5.3 Experimental Results

In the following, we present our main qualitative and quantitative experimental results. For additional results, please refer to Appendix B.

Comparison with Other Methods. We test our proposed method for each dataset under paired and unpaired settings. In the paired setting, the input conditions (text, sketch, fabric texture) correspond to the garment worn by the model. In the unpaired setting, they describe a different garment. Table 2 presents quantitative results of our models benchmarked against the aforementioned competitors on Dress Code Multimodal and VITON-HD Multimodal datasets. As it can be seen, the proposed Ti-MGD model consistently outperforms competitors in terms of realism (*i.e.* FID and KID) and coherency with input modalities (*i.e.* CLIP-S, PD, SD, and TS).

When considering text-only conditioned methods, we notice that Stable Diffusion [59] can produce images fairly consistent with text conditioning, as underlined by the CLIP-S, while struggling to maintain the original model pose. Constraining the generation on pose using ControlNet [89] helps alleviate this issue, resulting in a lower pose distance while also boosting sketch distance and realism performances. We argue that the improvement related to SD depends on the correlation between the pose and the garment sketch, while the boost in realism stems from the additional details provided by the input. Incorporating sketch constraints shows mixed results when considering ControlNet [89] and SDEdit [44]. The former slightly improves sketch coherence at the expense of realism, while SDEdit enhances both input coherence and realism. Note that we use our text-pose

[6]https://github.com/levindabhi/cloth-segmentation

Table 3. Category-wise quantitative results of Ti-MGD on the Dress Code Multimodal dataset for both paired and unpaired settings. When analyzing the modalities, T stands for Text, P for Pose, S for Sketch, and F for Fabric texture. Best results are in bold, second best are underlined.

	Modalities				Upper-body						Lower-body						Dresses					
Model	**T**	**P**	**S**	**F**	**FID ↓**	**KID ↓**	**CLIP-S ↑**	**PD ↓**	**SD ↓**	**TS ↑**	**FID ↓**	**KID ↓**	**CLIP-S ↑**	**PD ↓**	**SD ↓**	**TS ↑**	**FID ↓**	**KID ↓**	**CLIP-S ↑**	**PD ↓**	**SD ↓**	**TS ↑**
Paired setting																						
SD v1.5 [59]	✓				21.61	8.95	29.37	8.03	0.309	0.480	29.37	15.29	27.67	9.64	0.358	0.496	37.83	22.51	28.59	9.88	0.449	0.493
ControlNet [89]	✓	✓			21.51	8.97	29.36	6.32	0.287	0.479	25.35	11.25	28.38	8.54	0.334	0.499	36.34	20.64	28.91	8.25	0.439	0.495
SDEdit [44]	✓	✓	✓		12.20	2.40	30.28	4.40	0.232	0.527	12.59	2.72	29.48	6.60	0.266	0.542	16.48	6.09	30.81	5.63	0.291	0.565
ControlNet [89]	✓	✓	✓		27.04	13.79	28.51	6.25	0.279	0.466	34.46	19.71	27.22	8.88	0.326	0.482	45.70	27.96	27.92	8.52	0.457	0.480
ControlNet+IP [86, 89]	✓	✓	✓	✓	18.10	6.25	28.71	6.42	0.291	0.544	21.95	8.72	27.61	8.90	0.337	0.543	38.98	20.72	28.34	9.07	0.462	0.548
MGD [4]	✓	✓	✓		12.42	3.71	**31.90**	3.72	0.180	0.547	10.70	2.01	**31.10**	5.70	0.200	0.567	11.38	1.89	32.02	4.93	0.182	0.592
Ti-MGD (SDv1)	✓	✓	✓	✓	**7.92**	**0.76**	31.07	3.38	0.161	0.559	**7.22**	**0.59**	30.60	5.28	0.192	0.572	**9.14**	**0.92**	**32.10**	**4.52**	0.166	0.620
Ti-MGD (SDv2)	✓	✓	✓	✓	8.01	0.97	30.78	**3.33**	**0.160**	**0.568**	7.56	0.95	30.66	**5.25**	**0.191**	**0.582**	9.69	1.63	31.60	4.57	**0.164**	**0.637**
Unpaired setting																						
SD v1.5 [59]	✓				25.08	11.01	27.76	8.60	0.427	0.470	33.16	17.87	26.22	10.86	0.463	0.478	40.85	24.67	27.48	10.62	0.587	0.480
ControlNet [89]	✓	✓			24.59	10.80	27.72	6.72	0.397	0.469	29.53	13.74	27.17	9.50	0.440	0.485	38.83	22.54	27.90	9.06	0.570	0.487
SDEdit [44]	✓	✓	✓		14.52	3.18	27.40	4.76	0.325	0.518	15.73	3.58	27.36	7.51	0.347	0.526	18.99	8.02	28.20	6.36	0.391	0.553
ControlNet [89]	✓	✓	✓		30.58	16.35	26.79	6.94	0.388	0.459	37.98	22.35	26.12	9.50	0.420	0.473	47.32	29.30	27.05	9.18	0.578	0.474
ControlNet [89]+IP [86]	✓	✓	✓	✓	21.32	7.96	26.98	7.03	0.396	0.528	25.20	9.64	26.77	9.69	0.430	0.533	40.33	22.34	27.37	9.84	0.541	0.599
MGD [4]	✓	✓	✓		15.99	4.50	**29.76**	5.41	0.327	0.532	14.82	2.81	**29.96**	7.96	0.352	0.561	14.71	3.63	30.41	7.15	0.348	0.568
Ti-MGD (SDv1)	✓	✓	✓	✓	12.33	1.71	28.50	**4.59**	0.223	0.555	**12.93**	**1.51**	29.34	**7.52**	0.236	0.566	12.65	**1.96**	30.49	**6.62**	0.208	0.609
Ti-MGD (SDv2)	✓	✓	✓	✓	**12.01**	**1.32**	29.08	4.63	**0.220**	**0.573**	13.31	1.90	29.57	7.55	**0.231**	**0.582**	**12.56**	2.02	**30.69**	6.72	**0.203**	**0.635**

conditioned denoising network as the SDEdit backbone. When texture conditioning is added, we compare our Ti-MGD method against ControlNet combined with the IP-Adapter [86]. Notably, ControlNet+IP-Adapter not only boosts texture coherence metrics but also realism. Nevertheless, Ti-MGD surpasses this combination in both realism and adherence to input conditions in both paired and unpaired settings. When comparing the proposed Ti-MGD approach using SDv1 and SDv2 as backbone, we find comparable performance across all metrics, except in texture adherence, where the SDv2-based model shows a more significant improvement. Regarding instead the comparison with the previous version of our model, it is worth noting that adding texture conditioning leads to improved results across all metrics, except CLIP-S in which the previous version of our model achieves slightly improved performance.

Table 3 extends the previous analysis providing a detailed category-wise evaluation on the Dress Code Multimodal dataset. Due to the limited size of the test split for each category, containing only 1,800 images, FID exhibits variance in its results [6]. In contrast, KID delivers more reliable results. Despite this, our method consistently surpasses all competitors across most metrics. The only exception is in the pose metrics under unpaired settings, which can be attributed to the challenges in aligning the predicted warped unpaired sketch with the model's body shape and pose. Also in this case, the previous version of our model achieves better results only in terms of CLIP-S.

To qualitatively validate our results, Fig. 5 compares Ti-MGD with ControlNet+IP-Adapter using SDv1 for fairness. Row1-col1 and row2-col2 highlight improved adherence of Ti-MGD to texture and sketch inputs. Row3-col1 and row4-col1 show how Ti-MGD better integrates text and texture, yielding higher input coherence. Finally, row1-col2, row2-col1, row3-col2, and row4-col2 illustrate the ability of our approach to blend sketch and texture inputs into realistic garments while respecting multimodal guidance.

Varying Input Modalities. In Table 4, we analyze the behavior of our SDv2-based model when conditioned with various combinations of input modalities, either by masking inputs (*i.e.* using a zero tensor for pose and sketch) or by omitting them entirely (*i.e.* replacing texture cross-attention conditioning with text). Since text provides global semantic guidance, it anchors the CLIP-S metric across most configurations. When texture is

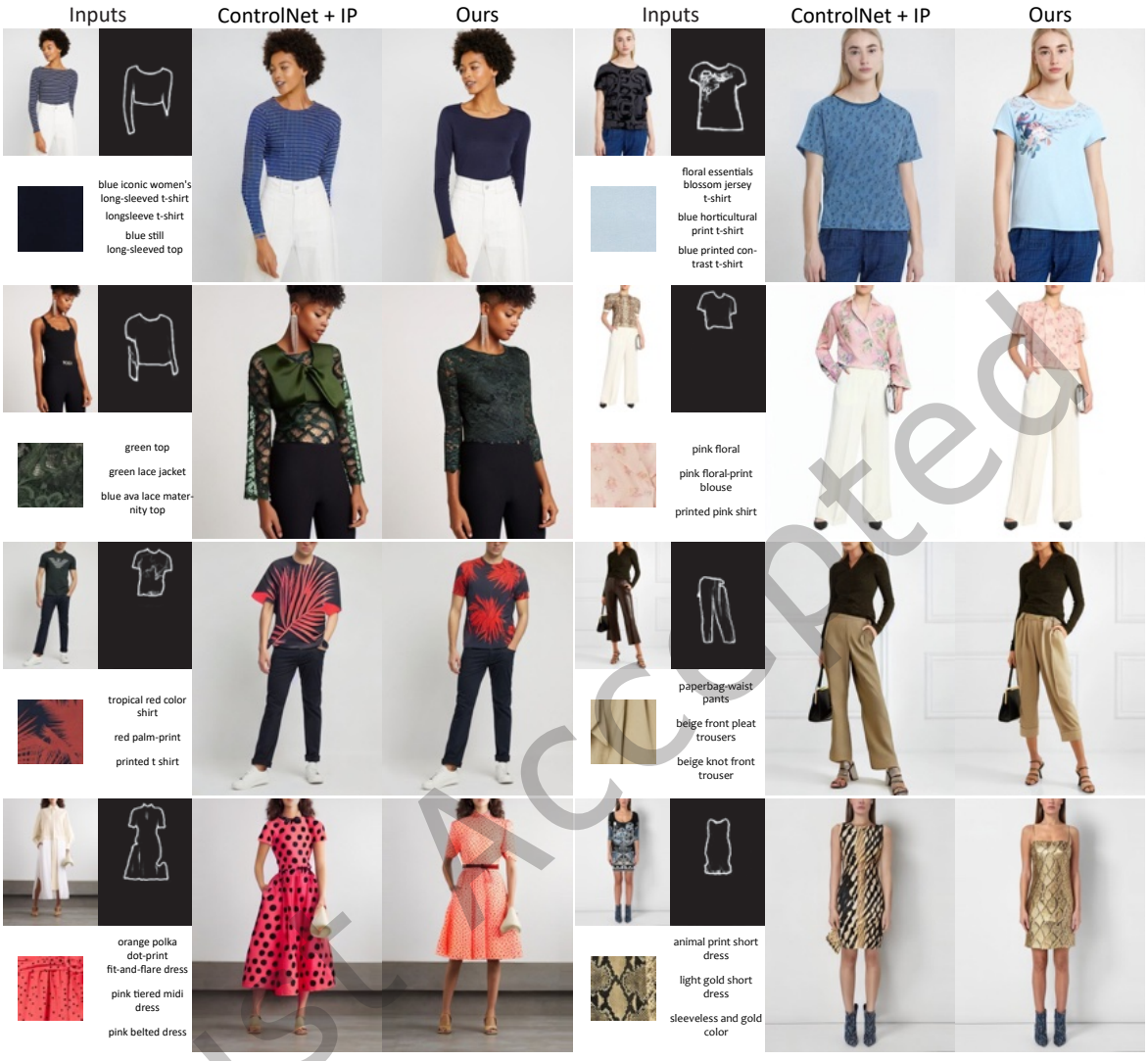

Fig. 5. Qualitative comparison of images generated using our approach versus ControlNet with IP-Adapter. The smaller images represent the model inputs, while the bigger images depict the generated outputs.

used, we observe a slight drop in CLIP-S due to the intrinsic mismatch between the texture appearance cues and the textual description.

Starting from the fully conditioned model (text, pose, sketch, texture), replacing the texture input with text reduces texture similarity and slightly affects realism, highlighting the importance of appearance-level conditioning. Masking the sketch input increases sketch distance and also slightly elevates pose distance due to the structural information implicitly conveyed by sketches. Further masking the pose map decreases pose adherence while keeping realism metrics largely unchanged. In contrast, configurations relying solely on pose, sketch, and texture, without text, exhibit a noticeable decrease in CLIP-S for both Dress Code Multimodal and VITON-HD Multimodal, reflecting the absence of semantic guidance. Despite this drop in semantic alignment,

Table 4. Performance analysis of our proposed model (Ti-MGD) on the unpaired setting of both Dress Code Multimodal and VITON-HD Multimodal datasets as input modalities vary.

Modalities				Dress Code Multimodal						VITON-HD Multimodal					
Text	Pose	Sketch	Texture	FID ↓	KID ↓	CLIP-S ↑	PD ↓	SD ↓	TS ↑	FID ↓	KID ↓	CLIP-S ↑	PD ↓	SD ↓	TS ↑
✓				6.43	1.35	30.41	7.24	0.404	0.539	10.37	1.54	29.37	8.18	0.493	0.559
✓	✓			6.45	1.50	30.18	6.42	0.374	0.539	10.53	1.71	29.31	7.26	0.472	0.560
✓	✓	✓		6.53	1.87	30.44	6.28	0.225	0.553	10.22	1.86	29.62	6.56	0.249	0.581
	✓	✓	✓	5.63	1.36	27.13	6.23	0.215	0.606	9.19	1.11	26.93	6.49	0.250	0.640
✓	✓	✓	✓	5.68	1.32	29.78	6.26	0.218	0.597	9.30	1.26	29.43	6.55	0.247	0.630

Table 5. Ablation study of our complete model varying the unconditional portion during training and the sketch conditioning rate at inference time. Results refer to the unpaired setting.

		Dress Code Multimodal					
Uncond. Portion	Sketch Cond.	FID ↓	KID ↓	CLIP-S ↑	PD ↓	SD ↓	TS ↑
0.1	1.0	9.91	4.94	27.15	6.42	0.148	0.576
0.2	1.0	9.64	4.55	27.04	6.48	0.155	0.576
0.3	1.0	9.95	4.77	27.25	6.49	0.164	0.573
0.2	0.8	8.75	3.81	27.61	6.48	0.166	0.580
0.2	0.6	7.91	3.03	28.10	6.44	0.175	0.586
0.2	0.4	6.56	1.87	28.76	6.35	0.182	0.593
0.2	0.2	5.68	1.32	29.78	6.26	0.218	0.597
0.2	0.0	5.64	1.19	29.30	6.40	0.370	0.584

these settings achieve competitive realism scores (FID/KID) and high texture similarity, indicating that structural and appearance cues alone can drive coherent and visually convincing generation.

Overall, these results collectively confirm that each modality contributes complementary information: text governs high-level semantics, pose and sketch provide structural cues, and texture refines appearance. Combining multiple modalities yields the best trade-off across all metrics, demonstrating the effectiveness of our model in handling heterogeneous conditioning signals. Fig. 6 shows how masking specific modalities influences the generated outputs.

Unconditional Training and Sketch Conditioning. Table 5 explores the performance of our fully conditioned network by varying the amount of unconditional training and the fraction of steps used to sketch conditioning. Specifically, we train three different models for unconditional training with fractions equal to 0.1, 0.2, and 0.3. To evaluate the sketch conditioning rate, we test our model over a range from 0 to 1 with a stride of 0.2. We find that optimal results are obtained when both parameters are set to 0.2, providing an ideal balance between neglecting the sketch (at lower rates) and compromising realism (at higher rates). This is also confirmed from a qualitative point of view, as shown in Fig. 7.

Inference Modality-Aware Group Conditioning. Before analyzing the results, it is important to note that a textual description of a given texture image is a high-level representation of it. Hence, the same textual information can refer to hypothetically infinite texture images other than the given one, *e.g.* replicating a specific texture based on text alone can be challenging. We can clearly see the effect of this asymmetry when analyzing images obtained by conditioning only on text or texture. For example, when creating images solely based on text input, the texture of the produced garment may not precisely replicate the specified texture image (Fig. 6), resulting

Table 6. Quantitative results of our proposed approach on Dress Code Multimodal dataset when using different cross-attention groups for conditioning on texture and text input. For a visual representation of the cross-attention groups, see Fig. 3. Here, ■ stands for the cross-attention groups conditioned on the texture, while ■ are cross-attention groups conditioned on the text.

Cross-Attention Groups	Dress Code Multimodal					
	FID ↓	KID ↓	CLIP-S ↑	PD ↓	SD ↓	TS ↑
	5.63	1.36	27.13	6.23	0.215	0.606
	5.63	1.36	27.13	6.25	0.215	0.605
	5.68	1.32	29.78	6.26	0.218	0.597
	6.31	1.77	30.45	6.25	0.223	0.560
	6.53	1.87	30.44	6.28	0.225	0.553
	6.53	1.87	30.44	6.27	0.225	0.553
	6.22	1.77	28.15	6.21	0.217	0.572
	5.62	1.38	27.22	6.22	0.215	0.604

in a sub-optimal texture similarity score. However, this approach tends to maximize the CLIP-S. On the other hand, if the generation process focuses exclusively on the texture, the resulting images might closely resemble the intended texture, achieving a high texture similarity. Yet, this method might lead to a misalignment with the textual description, as evidenced by a lower CLIP-S. In other words, we argue that generating a garment that simultaneously maximizes CLIP-S and texture similarity is unfeasible since these metrics are correlated with the high-level semantic information while competing with the low-level visual details. In this scenario, we look for a sweet spot between the CLIP-S and texture similarity metrics.

We can observe this phenomenon in Table 6, which shows the performance of our network when varying the textual and texture conditioning across the cross-attention groups. It is worth noting that the CLIP-S and texture similarity scores depend on the number and position of cross-attention layer groups conditioned on text or texture, respectively. The more groups are conditioned on the texture, the higher the texture similarity, while the more groups are conditioned on the text, the higher the CLIP-S. However, if we consider experiments conditioned on the same number of groups on texture (*i.e.* rows 2 vs. 8, 3 vs. 7, or 4 vs. 6), we obtain higher texture similarities when we condition the outer groups on the texture image. The best trade-off is obtained when the texture conditioning is applied to Groups 3 and 2 (*i.e.* row 3), which corresponds to a CLIP-S comparable to the only-text version and texture similarity comparable with the texture-only one. In Fig. 8, we report some qualitative examples of images generated when performing texture conditioning across different cross-attention groups. As it can be seen, performing texture conditioning on Groups 3 and 2 allows the generation of an image in line with both textual and texture input information.

5.4 Limitations and Future Work

Failure Cases. Fig. 9 illustrates representative failure cases of the proposed framework. In the first example (first row, first column), the generated person does not accurately reflect the body shape of the input model. This limitation arises from the reduced expressiveness of sparse keypoints, which do not fully capture body geometry. Incorporating denser or 3D pose representations may help alleviate this issue. In the second example (second row, first column), the sketch contains multiple disjoint regions, each delimited by clear contours; in such cases, only one region may be influenced by the texture input, suggesting the need for more explicit spatial control in texture conditioning. The examples in the second column show that our method is sensitive to inaccuracies in the predicted sketch: when the geometric warping step fails to produce a sketch aligned with the target body

Fig. 6. Qualitative examples of images generated by our proposed approach when varying the input modalities, where *T* represents text, *P* pose, *S* sketch, and *F* fabric texture.

shape, the generated output exhibits structural artifacts. Finally, the examples in the third column indicate that small, high-frequency details (such as hands or textual elements) may not be faithfully reproduced. This behavior is common among latent diffusion models [59] and is linked to the high compression ratio of the latent space.

Future Work. Although Ti-MGD is built upon UNet-based latent diffusion models, recent architectures such as SD v3 and DiT-based diffusion models [16] have demonstrated significant improvements in generative visual quality. Extending our multimodal conditioning strategy to these emerging backbones is a natural direction for future research. In particular, investigating how pose, sketch, and texture guidance can be incorporated within Transformer-based diffusion frameworks may further enhance the controllability and visual quality of synthesized garments, broadening the applicability of Ti-MGD in fashion image generation.

6 Conclusion

In this work, we introduced Ti-MGD, a novel framework for multimodal-conditioned fashion image editing that leverages multiple inputs (such as text, body pose, garment sketch, and fabric texture) to guide the generation process. Our approach extends latent diffusion models to handle these diverse modalities by adapting the denoising network for multimodal input. To incorporate texture effectively, we leverage textual inversion techniques and fuse text and texture features through the cross-attention layers of the denoising network, enabling fine-grained control without increasing model parameters. We also extend two existing fashion datasets with multimodal annotations using a semi-automatic pipeline. Extensive experiments, evaluated with standard and newly proposed metrics,

Fig. 7. Qualitative examples of images generated by our proposed approach when varying the sketch conditioning rate. We report the such rate on top of each column. Results are reported on sample images from the Dress Code Multimodal dataset.

demonstrate that Ti-MGD outperforms state-of-the-art methods in realism and coherence with multimodal inputs. Overall, our approach sets a new benchmark for fashion image editing and opens up new possibilities at the intersection of computer vision and fashion. These results mark one of the first successful attempts to emulate a designer's creative workflow and may enable broader adoption of diffusion models in fashion and other creative industries.

Acknowledgments

This work has been supported by the European Commission under the PNRR-M4C2 project "FAIR - Future Artificial Intelligence Research" and the European Horizon 2020 Programme with the projects "AI4Media - A European Excellence Centre for Media, Society and Democracy" (GA No. 951911) and "ELLIOT - European Large Open Multi-Modal Foundation Models For Robust Generalization On Arbitrary Data Streams" (GA No. 101214398).

References

[1] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. 2023. SpaText: Spatio-Textual Representation for Controllable Image Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. 2022. eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. *arXiv preprint arXiv:2211.01324* (2022).

Fig. 8. Qualitative examples of our proposed approach on Dress Code Multimodal and VITON-HD Multimodal when using different cross-attention groups for conditioning on texture and text input. At the top of the figure, we illustrate the conditioning modality of each cross-attention group. Where, ■ stands for the cross-attention groups conditioned on the texture, while ■ are cross-attention groups conditioned on the text. The red circles in the generated images highlight details derived from the textual input (also highlighted in red in the text). These details are visible when all cross-attention groups of the model are conditioned with textual information, as shown in the leftmost generated image in the figure. Conversely, they disappear in models fully conditioned on texture, as the textual information is no longer transmitted. On the contrary, the coherence of fabric texture is higher in the rightmost images and absent in the leftmost ones. Conditioning on Groups 3 and 2 (central column) emerges as a good trade-off between textual and texture fidelity, as evident from the images.

[3] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Zero-Shot Composed Image Retrieval with Textual Inversion. *Proceedings of the International Conference on Computer Vision* (2023).

[4] Alberto Baldrati, Davide Morelli, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. 2023. Multimodal Garment Designer: Human-Centric Latent Diffusion Models for Fashion Image Editing. In *Proceedings of the International Conference on Computer Vision*.

[5] Federico Bianchi, Jacopo Tagliabue, and Bingqing Yu. 2021. Query2Prod2Vec: Grounded Word Embeddings for eCommerce. In *Proceedings of the Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*.

[6] Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying MMD GANs. In *Proceedings of the International Conference on Learning Representations*.

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

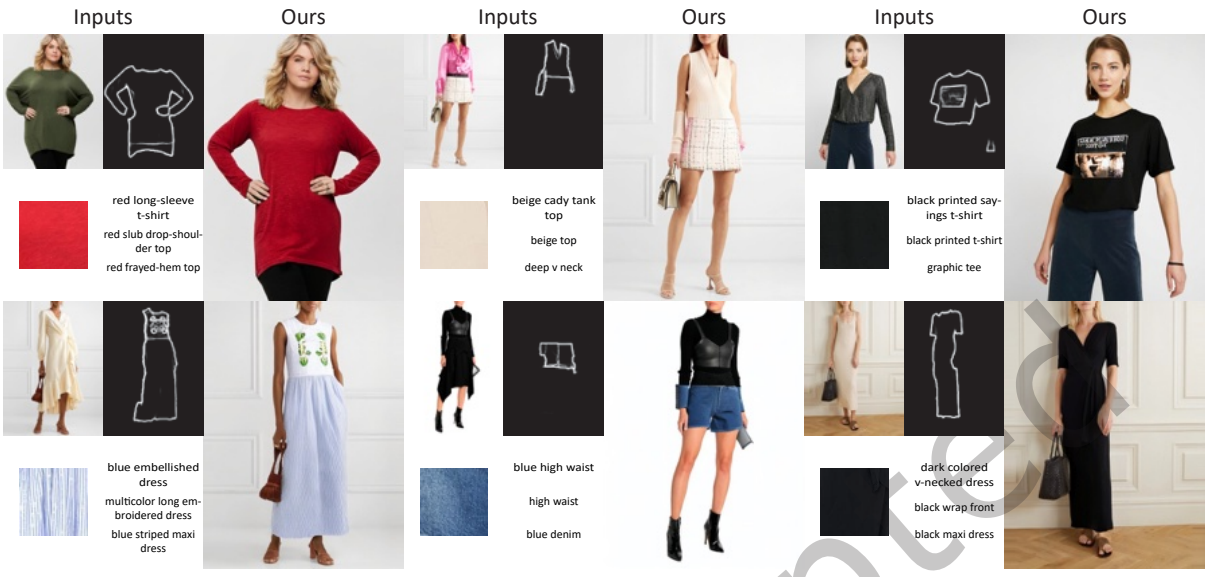

Fig. 9. Failure cases of our proposed approach on Dress Code Multimodal and VITON-HD Multimodal.

[8] Giuseppe Cartella, Alberto Baldrati, Davide Morelli, Marcella Cornia, Marco Bertini, and Rita Cucchiara. 2023. OpenFashionCLIP: Vision-and-Language Contrastive Learning with Open-Source Fashion Data. In *Proceedings of the International Conference on Image Analysis and Processing*.

[9] Shin-I Cheng, Yu-Jie Chen, Wei-Chen Chiu, Hung-Yu Tseng, and Hsin-Ying Lee. 2023. Adaptively-Realistic Image Generation from Stroke and Sketch with Diffusion Model. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.

[10] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. 2021. ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. In *Proceedings of the International Conference on Computer Vision*.

[11] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. 2021. VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[12] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. 2024. Improving Diffusion Models for Authentic Virtual Try-on in the Wild. In *Proceedings of the European Conference on Computer Vision*.

[13] Zheng Chong, Xiao Dong, Haoxiang Li, Shiyue Zhang, Wenqing Zhang, Xujie Zhang, Hanqing Zhao, Dongmei Jiang, and Xiaodan Liang. 2025. CatVTON: Concatenation Is All You Need for Virtual Try-On with Diffusion Models. In *Proceedings of the International Conference on Learning Representations*.

[14] Giannis Daras and Alexandros G Dimakis. 2022. Multiresolution Textual Inversion. In *Advances in Neural Information Processing Systems Workshops*.

[15] Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh Jha, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon. 2022. TorchMetrics - Measuring Reproducibility in PyTorch. *Journal of Open Source Software* 7, 70 (2022), 4101.

[16] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *Proceedings of the International Conference on Machine Learning*.

[17] Matteo Fincato, Federico Landi, Marcella Cornia, Fabio Cesari, and Rita Cucchiara. 2020. VITON-GT: An Image-based Virtual Try-On Model with Geometric Transformations. In *Proceedings of the International Conference on Pattern Recognition*.

[18] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *Proceedings of the International Conference on Learning Representations*.

[19] Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. 2023. Taming the Power of Diffusion Models for High-Quality Virtual Try-On with Appearance Flow. In *Proceedings of the ACM International Conference on Multimedia*.

[20] Junyi Guo, Jingxuan Zhang, Fangyu Wu, Huanda Lu, Qiufeng Wang, Wenmian Yang, Eng Gee Lim, and Dongming Lu. 2025. HiGarment: Cross-modal Harmony Based Diffusion Model for Flat Sketch to Realistic Garment Image. *arXiv preprint arXiv:2505.23186* (2025).

[21] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. 2015. Where to buy it: Matching street clothing photos in online shops. In *Proceedings of the International Conference on Computer Vision*.
[22] Inhwa Han, Serin Yang, Taesung Kwon, and Jong Chul Ye. 2023. Highly Personalized Text Embedding for Image Manipulation by Stable Diffusion. *arXiv preprint arXiv:2303.08767* (2023).
[23] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. 2017. Automatic Spatially-Aware Fashion Concept Discovery. In *Proceedings of the International Conference on Computer Vision*.
[24] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. 2018. VITON: An Image-Based Virtual Try-On Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[25] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415* (2016).
[26] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
[27] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a Nash equilibrium. In *Advances in Neural Information Processing Systems*.
[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*.
[29] Wei-Lin Hsiao and Kristen Grauman. 2018. Creating capsule wardrobes from fashion images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[30] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzenes. 2020. Do Not Mask What You Do Not Need to Mask: A Parser-Free Virtual Try-On. In *Proceedings of the European Conference on Computer Vision*.
[31] Boyuan Jiang, Xiaobin Hu, Donghao Luo, Qingdong He, Chengming Xu, Jinlong Peng, Jiangning Zhang, Chengjie Wang, Yunsheng Wu, and Yanwei Fu. 2024. FitDiT: Advancing the Authentic Garment Details for High-fidelity Virtual Try-on. *arXiv preprint arXiv:2411.10499* (2024).
[32] Yuming Jiang, Shuai Yang, Haonan Qju, Wayne Wu, Chen Change Loy, and Ziwei Liu. 2022. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics* 41, 4 (2022), 1–11.
[33] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. 2023. HumanSD: A Native Skeleton-Guided Diffusion Model for Human Image Generation. In *Proceedings of the International Conference on Computer Vision*.
[34] Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. 2024. StableVITON: Learning Semantic Correspondence with Latent Diffusion Model for Virtual Try-On. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[35] Jeongho Kim, Gyojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. 2024. StableVITON: Learning Semantic Correspondence with Latent Diffusion Model for Virtual Try-On. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[36] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. 2021. OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association. *IEEE Transactions on Intelligent Transportation Systems* 23, 8 (2021), 13498–13511.
[37] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-Concept Customization of Text-to-Image Diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[38] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. 2022. MAT: Mask-Aware Transformer for Large Hole Image Inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[39] Yuhan Li, Hao Zhou, Wenxiang Shang, Ran Lin, Xuanhong Chen, and Bingbing Ni. 2024. AnyFit: Controllable Virtual Try-on for Any Combination of Attire Across Any Scenario. In *Advances in Neural Information Processing Systems*.
[40] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[41] Davide Lobba, Fulvio Sanguigni, Bin Ren, Marcella Cornia, Rita Cucchiara, and Nicu Sebe. 2025. Inverse Virtual Try-On: Generating Multi-Category Product-Style Images from Clothed Individuals. *arXiv preprint arXiv:2505.21062* (2025).
[42] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations*.
[43] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[44] Chenlin Meng, Yutong He adnd Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *Proceedings of the International Conference on Learning Representations*.
[45] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed Precision Training. In *Proceedings of the International Conference on Learning Representations*.
[46] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text Inversion for Editing Real Images using Guided Diffusion Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[47] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. 2023. LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On. In *Proceedings of the ACM International Conference on Multimedia*.

[48] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. 2022. Dress Code: High-Resolution Multi-Category Virtual Try-On. In *Proceedings of the European Conference on Computer Vision*.
[49] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. 2024. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. In *Proceedings of the Conference on Artificial Intelligence*.
[50] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *Proceedings of the International Conference on Machine Learning*.
[51] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proceedings of the International Conference on Machine Learning*.
[52] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. 2022. On Aliased Resizing and Surprising Subtleties in GAN Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[53] Martin Pernuš, Clinton Fookes, Vitomir Štruc, and Simon Dobrišek. 2025. FICE: Text-Conditioned Fashion Image Editing With Guided GAN Inversion. *Pattern Recognition* 158 (2025), 111022.
[54] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv preprint arXiv:2307.01952* (2023).
[55] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning*.
[56] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.
[57] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125* (2022).
[58] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. 2017. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[59] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[60] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*.
[61] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[62] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems*.
[63] Fulvio Sanguigni, Davide Morelli, Marcella Cornia, and Rita Cucchiara. 2025. Fashion-RAG: Multimodal Fashion Image Editing via Retrieval-Augmented Generation. In *Proceedings of the International Joint Conference on Neural Networks*.
[64] Rohan Sarkar, Navaneeth Bodla, Mariya I Vasileva, Yen-Liang Lin, Anurag Beniwal, Alan Lu, and Gerard Medioni. 2023. OutfitTransformer: Learning Outfit Representations for Fashion Recommendation. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.
[65] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, et al. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*.
[66] Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinhui Tang. 2025. IMAGDressing-v1: Customizable Virtual Dressing. In *Proceedings of the Conference on Artificial Intelligence*.
[67] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning*.
[68] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *Proceedings of the International Conference on Learning Representations*.
[69] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15, 56 (2014), 1929–1958.
[70] Zhuo Su, Wenzhe Liu, Zitong Yu, Dewen Hu, Qing Liao, Qi Tian, Matti Pietikäinen, and Li Liu. 2021. Pixel difference networks for efficient edge detection. In *Proceedings of the International Conference on Computer Vision*.
[71] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2022. Resolution-Robust Large Mask Inpainting With Fourier Convolutions. In

Proceedings of the IEEE Winter Conference on Applications of Computer Vision.

[72] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, et al. 2022. DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[73] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. 2023. Sketch-Guided Text-to-Image Diffusion Models. In *ACM SIGGRAPH Conference Proceedings.*

[74] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. 2023. *P+*: Extended Textual Conditioning in Text-to-Image Generation. *arXiv preprint arXiv:2303.09522* (2023).

[75] Siqi Wan, Jingwen Chen, Yingwei Pan, Ting Yao, and Tao Mei. 2025. Incorporating Visual Correspondence into Diffusion Model for Virtual Try-On. In *Proceedings of the International Conference on Learning Representations.*

[76] Siqi Wan, Yehao Li, Jingwen Chen, Yingwei Pan, Ting Yao, Yang Cao, and Tao Mei. 2024. Improving Virtual Try-On with Garment-focused Diffusion Models. In *Proceedings of the European Conference on Computer Vision.*

[77] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. 2018. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision.*

[78] Rui Wang, Hailong Guo, Jiaming Liu, Huaxia Li, Haibo Zhao, Xu Tang, Yao Hu, Hao Tang, and Peipei Li. 2024. StableGarment: Garment-Centric Generation via Stable Diffusion. *arXiv preprint arXiv:2403.10783* (2024).

[79] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. 2022. Pretraining is All You Need for Image-to-Image Translation. *arXiv preprint arXiv:2205.12952* (2022).

[80] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. 2022. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[81] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[82] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[83] Yuhao Xu, Tao Gu, Weifeng Chen, and Arlene Chen. 2025. OOTDiffusion: Outfitting Fusion Based Latent Diffusion for Controllable Virtual Try-On. In *Proceedings of the Conference on Artificial Intelligence.*

[84] Yiyan Xu, Wenjie Wang, Fuli Feng, Yunshan Ma, Jizhi Zhang, and Xiangnan He. 2024. Diffusion Models for Generative Outfit Recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval.*

[85] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. 2020. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[86] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *arXiv preprint arXiv:2308.06721* (2023).

[87] Jianhao Zeng, Dan Song, Weizhi Nie, Hongshuo Tian, Tongtong Wang, and An-An Liu. 2024. CAT-DM: Controllable Accelerated Virtual Try-on with Diffusion Model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[88] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[89] Lvmin Zhang and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the International Conference on Computer Vision.*

[90] Shiyue Zhang, Zheng Chong, Xujie Zhang, Hanhui Li, Yuhao Cheng, Yiqiang Yan, and Xiaodan Liang. 2024. GarmentAligner: Text-to-Garment Generation via Retrieval-Augmented Multi-level Corrections. In *Proceedings of the European Conference on Computer Vision.*

[91] Xiangyu Zhao, Yuehan Zhang, Wenlong Zhang, and Xiao-Ming Wu. 2024. UniFashion: A Unified Vision-Language Model for Multimodal Fashion Retrieval and Generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

[92] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. 2023. TryOnDiffusion: A Tale of Two UNets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[93] Shizhan Zhu, Raquel Urtasun, Sanja Fidler, Dahua Lin, and Chen Change Loy. 2017. Be Your Own Prada: Fashion Synthesis with Structural Coherence. In *Proceedings of the International Conference on Computer Vision.*

A Annotation Pipeline Details

Semi-Automatic Noun Chunk Extraction. To construct our phrase vocabulary, we standardize captions from FashionIQ [81] and Fashion200k [23] via lemmatization (NLTK[7]) and parse them into noun chunks. We remove initial articles and filter phrases containing special characters, resulting in 60,284 unique chunks categorized as upper-body, lower-body, or dresses.

CLIP-Based Matching. We use five vision-language models (*i.e.* CLIP ViT-L/14@336px, CLIP RN50x64, OpenCLIP ViT-L/14, OpenCLIP ViT-H/14, and OpenCLIP ViT-g/14) to embed both images and text. For each garment, we retrieve the top-5 most similar noun chunks from each model using cosine similarity and ensemble the results, yielding 25 unique candidates per item.

Manual Annotation Interface. Annotators use a custom tool to select the top-3 matching phrases from the 25 auto-generated candidates or input a new chunk. The interface is optimized for speed (avg. 60 seconds/item) and ensures coverage of the original Dress Code test set [48].

Hybrid Annotation with Fine-Tuned OpenCLIP. We fine-tune OpenCLIP ViT-B/32 (pre-trained on LAION-5B [65]) using our human-labeled image-chunk pairs. This model is then used to annotate the remainder of Dress Code and VITON-HD's upper-body garments, selecting the top-3 relevant phrases per item.

Sketch Extraction via Warped Garment Alignment. We adopt PiDiNet [70] for edge extraction. For unpaired samples, we warp the in-shop garment using a thin-plate spline (TPS) transformation guided by a correlation map between the encoded garment C and cloth-agnostic person features (pose map P and masked model image I_M). A UNet refines the warped result $\hat{C}$ into $\tilde{C}$, enabling sketch extraction aligned to the target model (similar to [30, 85]).

B Additional Results

User Study. To validate our results with human feedback, we conducted a user study evaluating the realism and multimodal input adherence of generated images. Involving more than 100 participants, the study collected around 5,000 evaluations. Results displayed in Table 7 reveal that our model consistently outperforms others in both realism and input adherence, confirming the effectiveness of our method. Comparing these results with quantitative evaluations reported in the main paper highlights a correlation between our quantitative metrics and human judgments, both in terms of realism and coherence.

[7]https://www.nltk.org

Table 7. User study results on the unpaired setting of both Dress Code Multimodal and VITON-HD Multimodal. We report the percentage of times an image from Ti-MGD is preferred against a competitor. Note that when comparing against ControlNet with all modalities, we employ IP-Adapter to condition on texture.

		Modalities				Realism			Multimodal Coherence		
		T	P	S	F	SD	ControlNet	SDEdit	SD	ControlNet	SDEdit
Dress	Code M.	✓				94.54	-	-	77.92	-	-
		✓	✓			-	91.89	-	-	79.07	-
		✓	✓	✓		.	96.10	80.21	-	94.44	67.65
		✓	✓	✓	✓	-	97.67	-	-	84.15	-
VITON	HD M.	✓				95.83	-	-	84.34	-	-
		✓	✓			.	95.60	-	-	77.38	-
		✓	✓	✓		-	94.25	64.94	-	82.05	78.05
		✓	✓	✓	✓	-	96.62	-	-	89.62	-

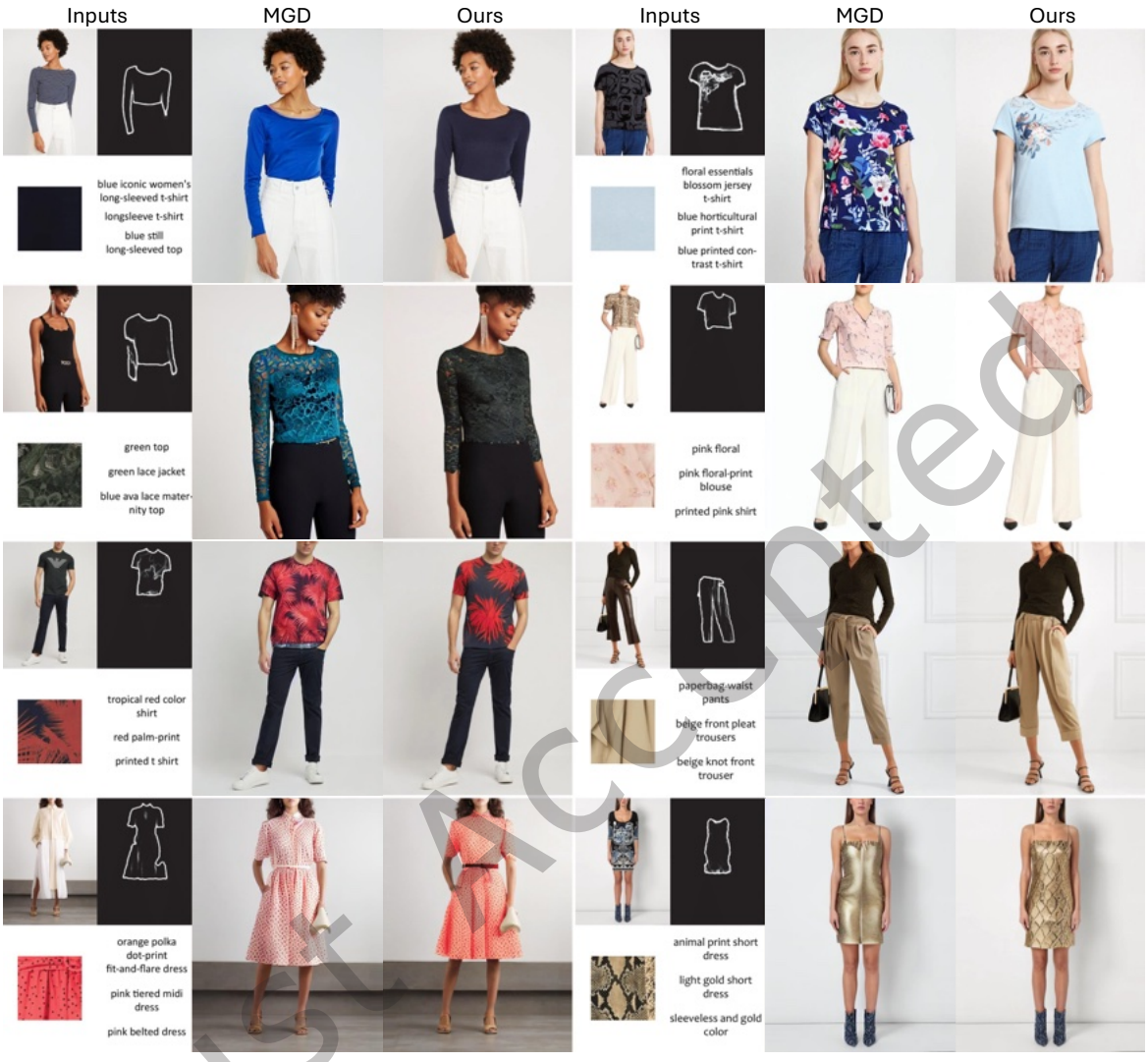

Fig. 10. Qualitative comparison of images generated using our approach versus MGD [4]. The smaller images represent the model inputs, while the bigger images depict the generated outputs.

Qualitative Comparison with MGD. To complement the quantitative evaluation reported in the main paper, Fig. 10 provides a qualitative comparison between MGD and Ti-MGD. Since MGD does not incorporate texture as an input modality, the comparison is conducted by providing both methods with the shared conditioning signals (*i.e.*, text, pose, and sketch), while Ti-MGD additionally receives the texture input. The results highlight how Ti-MGD integrates texture information to produce garments with more faithful appearance details, while maintaining the structural improvements of MGD in terms of pose and sketch consistency.

Received 4 September 2025; revised 2 December 2025; accepted 3 January 2026