

# Multimodal-Conditioned Latent Diffusion Models for Fashion Image Editing

Alberto Baldrati\*, Davide Morelli\*, Marcella Cornia, Marco Bertini, Rita Cucchiara



**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).

**Abstract**—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.

The first two authors equally contributed to this research.

A. Baldrati and M. Bertini are with the Media Integration and Communication Center (MICC), University of Florence, Italy (e-mail: {alberto.baldrati, marco.bertini}@unifi.it).

D. Morelli and R. Cucchiara are with the Department of Engineering “Enzo Ferrari”, University of Modena and Reggio Emilia, Italy (e-mail: {davide.morelli, rita.cucchiara}@unimore.it).

A. Baldrati and D. Morelli are also with the University of Pisa, Italy.

M. Cornia is with the Department of Education and Humanities, University of Modena and Reggio Emilia, Italy (e-mail: marcella.cornia@unimore.it).

Manuscript received March, 2024.

**Index Terms**—Fashion Product Design, Latent Diffusion Models, Textual Inversion, Generative AI, Multimodal Learning.

## I. INTRODUCTION

**I**N 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 [1]–[4], garment and outfit recommendation [5]–[7], and virtual try-on [8]–[13]. While these works have advanced research in the field, limited attention has been paid to text-conditioned fashion image editing, mainly 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 [14]–[16] 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 [17]–[20] 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. The integration of multiple prompts in generative models is a complex computer vision task, especially in the context of garment images. In fact, these images exhibit considerable variation, influenced by factors such as the target gender (*i.e.* whether the garment is designed for men or women), the 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 [20], 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 [21]. 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 [11] and VITON-HD [12], 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 proposed 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>.

## II. RELATED WORK

**Text-Guided Image Generation.** The objective of text-to-image synthesis is to create an image that accurately represents a given textual prompt. Initially, approaches in this domain relied on GANs [22]–[25], while recent advancements have shifted towards the use of diffusion models [20], [26]–[28]. Among them, Nichol *et al.* [26] proposed a text-to-image diffusion model with local editing capabilities to match more complex prompts. On a similar line, while Ramesh *et al.* [27] proposed a two-stage approach involving a prior which produces CLIP image embeddings [29] conditioned on textual captions and a diffusion-based decoder that translates images based on such embeddings, the approach proposed in [30] leverages the T5 language model [31] followed by a cascade of super-resolution diffusion models to improve the generation process. Unlike directly applying the diffusion process to pixel space, the current trend favors the use of latent diffusion models as introduced in [20]. In these models, the forward and reverse processes are defined in the latent space of a pre-trained autoencoder, aiming for improved computational efficiency and the production of high-quality images.

Only a few attempts of text-to-image synthesis have been conducted for the fashion domain [14]–[16]. Notably, Zhu *et al.* [14] 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 [16], where a latent code regularization technique is employed to enhance the GAN inversion process. This involves leveraging CLIP textual embeddings [29] to guide the image generation process. Differently, Jiang *et al.* [15] 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 [32]–[36]. In this context, Choi *et al.* [32] proposed refining the generative mechanism of an unconditional denoising diffusion probabilistic model [19] by aligning each latent variable with a given reference image. Conversely, the approach proposed by Mang *et al.* [33] introduces noise to a stroke-based input and applies the reverse stochastic differential equation to generate images, without additional training. Instead, Wang *et al.* [34] 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 [36] or by training an MLP-based edge predictor to map latent features to spatial maps [37].

Among contemporary works designing novel conditioning strategies for pre-trained latent diffusion models, Zhang *et al.* introduced ControlNet [38] which extends Stable Diffusion [20] with an additional conditioning input. This procedure entails creating two variants of the original model parameters: one remains fixed and unaltered (referred to as the locked copy), while the other can be updated during training (referred to as the trainable copy). The objective is to enable the trainable version to assimilate the newly introduced condition, while the locked version preserves the knowledge of the original model. Conversely, the model proposed in [35] incorporates modality-specific adapter modules facilitating Stable Diffusion conditioning on novel modalities. On a similar line, Ye *et al.* [39] introduces a lightweight adapter that can enable Stable Diffusion conditioning on image prompts, directly using 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.

**Textual Inversion.** Textual inversion, as introduced in the recent work by Gal *et al.* [40], is a novel technique aimed at learning pseudo words within the embedding space of a text encoder to represent visual concepts effectively. Building on this, several promising methods have been developed for personalized image generation and editing [41]–[44]. Among them, Ruiz *et al.* [42] 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.* [45] 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.* [41] decomposed the CLIP [29] 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.

**Virtual Try-On.** Another related research area involves the virtual try-on of a desired garment [8]. This task is usually addressed by first generating the warped version of the input

garment and then synthesizing the final image of a reference person wearing the warped clothing item while preserving human pose and identity. Research efforts in this domain have mainly focused on the improvement of the geometric transformation phase [9], [10], [46] and on the design of additional components to enhance the realism of generated images [47]–[49], also considering the problem in high-resolution settings [11], [12]. Recent attempts follow the latest trends in image generation and apply diffusion models to improve the quality of virtual try-on generated images [13], [50]–[53]. While these approaches can be applied to enhance user experience, they all take the try-on garment as input and, therefore, can not directly be compared with the model proposed in this work, which instead generates the final image leveraging text, human pose, fabric texture, and sketch modalities.

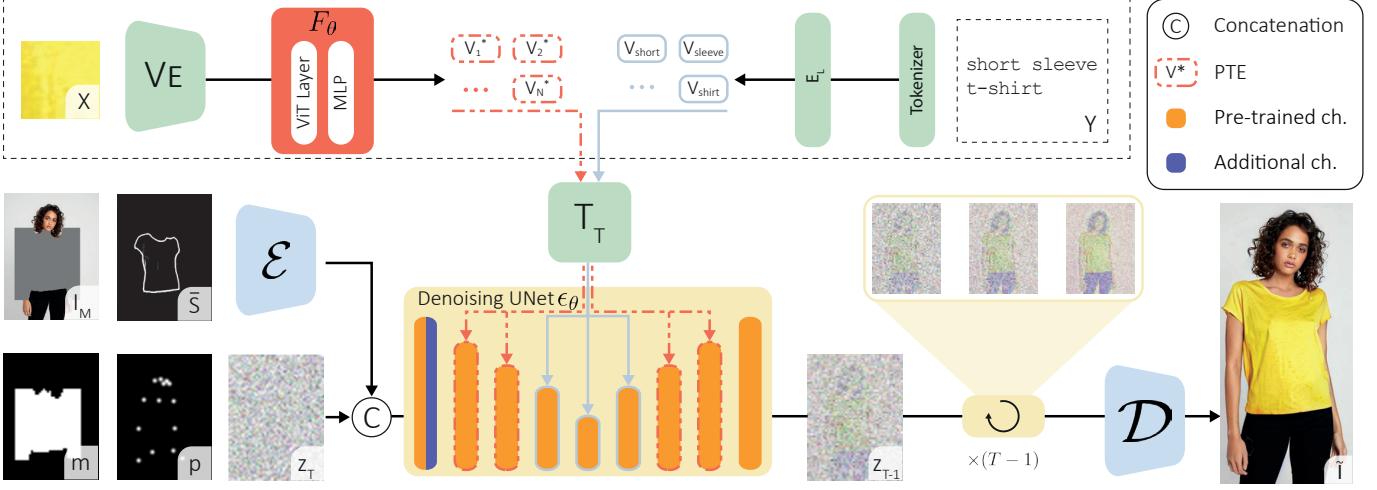
### III. PROPOSED METHOD

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  $X \in \mathbb{R}^{H \times W \times 3}$ , 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.

#### A. Preliminaries

**Stable Diffusion.** While diffusion models [54] are latent variable architectures that work in the same dimensionality of the data (*i.e.* in the pixel space), latent diffusion models (LDMs) [20] 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 [20], 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  $\epsilon_\theta$ , 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  $\epsilon_\theta$  convolutional input (*i.e.*  $z_t$  in this case) as spatial input  $\gamma$ , because of the property of convolutions to preserve the spatial structure, and the attention



**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  $\epsilon_\theta$  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.

conditioning input as  $\psi$ . The denoising network  $\epsilon_\theta$  is trained according to the following loss:

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

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 [29] 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 [55], 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  $\mathcal{W}$ .

### B. 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 rescaled 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  $\gamma$ , while we inject semantic information such as textual descriptions and fabric textures as attention conditioning input  $\psi$ . 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 [56]–[58], 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 [38], 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 [59].

**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  $F_\theta$ . The overall mathematical formulation is as follows:

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

We then use the predicted PTEs  $V^*$  to condition the Stable Diffusion denoising network  $\epsilon_\theta$  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 [60] 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 [40], [42], [45]. Instead of iteratively optimizing pseudo-word token embeddings, our solution trains the adapter  $F_\theta$  to generate these embeddings in a single forward pass.

### C. Training and Inference

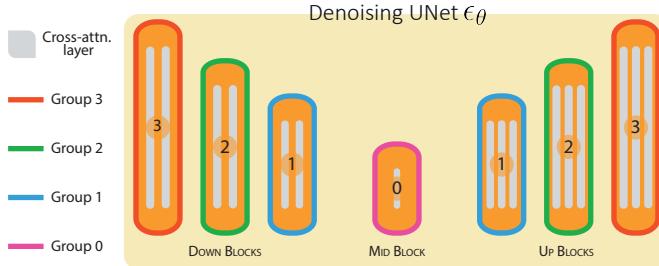
Following the standard LDM approach, the proposed denoising network predicts the noise added stochastically to the encoded input,  $z = \mathcal{E}(I)$ . The objective function can be specified as

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

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:

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



**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.

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  $\alpha$  is a hyperparameter that controls the degree of extrapolation towards the condition.

We use the fast variant multi-condition classifier-free guidance proposed in [61] 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  $\Delta_{\text{joint}}^t = \epsilon_\theta(z_t|\{c_i\}_{i=1}^{i=K}) - \epsilon_\theta(z_t|\emptyset)$ :

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

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  $\gamma$ , that we feed to the denoising network convolutional input, and attention conditioning input  $\psi$ , 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  $\psi$ . 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  $\epsilon_\theta$  are executed  $T$  times during inference, where  $T$  is the number of 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 [60]. 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  $\epsilon_\theta$ , 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.

#### IV. 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 [11] and VITON-HD [12], 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  $1024 \times 768$ , each composed of a garment image and a corresponding model image wearing it. 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 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.

##### A. Dataset Collection and Annotation

**Data Preparation.** We begin annotating the Dress Code dataset, which comprises over 53k model-garment pairs across various categories. The initial task is to annotate each garment with a concise yet detailed textual description, using fashion-specific and non-generic terms suitable for guiding the generation process. Inspired by research indicating that people typically describe fashion items in just a few words [62], we propose to use noun chunks. These are brief textual phrases consisting of a noun grouped together with its modifiers, effectively conveying key information while eliminating superfluous details and words.



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

Given the time-consuming and resource-intensive nature of manual annotation<sup>1</sup>, we propose a semi-automatic framework using noun chunks for annotating the dataset. Initially, we gather domain-specific captions from two existing fashion datasets, FashionIQ [63] and Fashion200k [64]. These captions are then standardized using word lemmatization, with each word reduced to its root form using the NLTK library<sup>2</sup>. Subsequently, we extract noun chunks from these standardized captions, eliminating the article at the beginning of each noun chunk if present and any textual elements that contain or start with special characters. This pre-processing step yields over 60,000 unique noun chunks, categorized into three groups: upper-body clothes, lower-body clothes, and dresses. Table I reports detailed statistics about the number of unique captions and extracted noun chunks from which we start the annotation.

We then use the CLIP model [29] and its open-source counterpart, OpenCLIP [65], to identify the most relevant noun chunks for each garment. For the CLIP model, we choose the ViT-L/14@336px and RN50×64 versions and for OpenCLIP, we use the ViT-L/14, ViT-H/14, and ViT-g/14 models. To improve accuracy, we employ prompt ensembling. For each image, we associate 25 unique noun chunks by selecting the top-5 from each model based on the cosine similarity between the image and text embeddings, avoiding any duplicates.

**Human-Labeled Textual Annotation.** To guarantee the accuracy and diversity of our annotations, we manually label a substantial portion of the Dress Code dataset images. Specifically, we select the three most fitting noun chunks from the 25 automatically associated ones for each garment image. We develop a custom tool to streamline the annotation process, limiting the average annotation time to 60 seconds per item. This tool also allows the manual insertion of noun chunks when the automatic choices are unsuitable. We manually annotate 26,400 garments (8,800 in each category) from the total 53,792 products in the dataset, ensuring the inclusion of all fashion items from the original test set [11].

**Hybrid Textual Annotation.** To finalize the textual annotation process, we first fine-tune the OpenCLIP ViT-B/32 model pre-

<sup>1</sup>Considering the Dress Code size of over 53k items and estimating 5 minutes per annotation, a single annotator working 8 hours a day, 5 days a week, for 260 days a year would need over 2 years to complete the task.

<sup>2</sup><https://www.nltk.org/>

**TABLE I:** Number of unique captions and noun chunks for each category of the FashionIQ and Fashion200k datasets.

Dataset	Unique Captions			Unique Noun Chunks		
	Upper	Lower	Dresses	Upper	Lower	Dresses
FashionIQ [63]	27,339	0	15,101	7,801	0	3,592
Fashion200k [64]	25,959	11,022	16,694	22,898	13,420	15,890

trained on the English portion of the LAION-5B dataset [66], using our newly annotated image-text pairs. Next, we use this updated model and the extracted noun chunks to automatically tag the remaining items in the Dress Code dataset with the three most relevant noun chunks, chosen based on cosine similarity between multimodal embeddings. We apply the same method for automatically annotating all upper-body garment images in the VITON-HD dataset, restricting the noun chunks to those specifically describing upper-body clothing.

**Extracting Sketches.** Introducing garment sketches enhances the design details in our dataset, offering a more accurate and comprehensive view that text alone cannot express. This leads to better quality and control in the generated designs. We use PiDiNet [67], a pre-trained edge detection network, to extract sketches for both the Dress Code and VITON-HD datasets.

The datasets, originally introduced for virtual try-on, include paired and unpaired test sets. In the paired setting, each sample features an in-shop garment and a model wearing it, while in the unpaired setting, the in-shop garment is different from the one the model is wearing. For the paired set, we use the human parsing mask to isolate the garment on the model before processing it with the edge detection network. For the unpaired set, since we do not have a warped reference garment, we first match the in-shop garment to the model’s pose and shape. To this aim, we employ a geometric transformation module that leverages a thin-plate spline transformation [68] and refines the results with a UNet model [69]. This allows us to extract sketches from each warped garment, making our approach also applicable to unpaired settings.

In particular, we follow recent virtual try-on literature [10], [47] and implement a warping module that computes a correlation map using encoded representations of the in-shop garment  $C$  and a cloth-agnostic person representation, combining the pose map  $P \in \mathbb{R}^{H \times W \times 18}$  and the masked model image  $I_M \in \mathbb{R}^{H \times W \times 3}$ . These representations are obtained through two distinct convolutional networks. The correlation map then guides the prediction of spatial transformation parameters for a thin-plate spline geometric transformation. Using these parameters, we generate a coarse warped garment  $\hat{C}$  from  $C$ . A UNet model further refines this output, taking as input the coarse warped garment  $\hat{C}$ , the pose map  $P$ , and the masked model image  $I_M$ , to produce the finely warped garment  $\tilde{C}$ .

**Extracting Textures.** To enable users precise control over garment generation, including the fabric texture samples in the dataset is crucial, as text conditioning cannot convey detailed visual fabric characteristics. 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  $128 \times 128$  pixels, selecting only

**TABLE II:** Comparison of Dress Code Multimodal and VITON-HD Multimodal with other fashion datasets featuring multimodal 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 [12]	X	✓	X	X	27,358	13,679	-	-
Dress Code [11]	X	✓	X	X	107,584	53,792	-	-
Be Your Own Prada [14]	✓	✓	X	X	78,979	N/A	3,972	445
DF-Multimodal [15]	✓	✓	X	X	44,096	N/A	10,253	77
<b>VITON-HD Multimodal</b>	✓	✓	✓	✓	27,358	13,679	5,143	1,613
<b>Dress Code Multimodal</b>	✓	✓	✓	✓	107,584	53,792	25,596	2,995

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.*  $1024 \times 768$  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  $64 \times 64$  pixels to guarantee at least one patch  $X$  for each garment  $C$ .

### B. Comparison with Other Datasets

The only two text-to-image generation datasets in the fashion domain, referenced in [14] and [15], both utilize images from the DeepFashion dataset [2]. The dataset from [14] includes brief textual descriptions, while DeepFashion-Multimodal [15] features attributes (*e.g.* category, color, fabric, etc.) for crafting longer captions. In Table II, 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.

## V. EXPERIMENTAL EVALUATION

### A. Implementation Details and Competitors

**Training and Inference.** All models are trained on the original splits of the Dress Code [11] and VITON-HD [12] 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  $512 \times 384$ . When trained on Dress Code Multimodal, models undergo 200k training steps, while for VITON-HD Multimodal, they are trained for 75k steps. As the latent diffusion model, we use Stable Diffusion inpainting v2<sup>3</sup>. To ensure a fair comparison with other models, we also develop a version of Ti-MGD based on Stable Diffusion inpainting v1<sup>4</sup>. 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 [70] is employed as optimizer, with a weight decay of  $10^{-2}$ . To speed up training and save memory, we use mixed precision [71]. 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  $F_\theta$  comprises a single ViT layer and an MLP projection. The MLP includes three fully connected layers, each separated by GELU non-linearity [72] and a dropout layer [73]. The network outputs  $N = 16$  PTEs. As the visual encoder  $V_E$ , we use OpenCLIP ViT-H/14 [65], pre-trained on the English portion of the LAION-5B dataset [66]. We use the DDIM [74] with 50 steps as our noise scheduler during inference, setting the classifier-free guidance parameter  $\alpha$  to 7.5. To improve the high-frequency details in the region outside the inpainting area, we leverage the EMASC module defined in [13].

**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<sup>4</sup>. In scenarios involving text and pose inputs, Ti-MGD is compared with Stable Diffusion v1.5 integrated with ControlNet [38] for pose conditioning<sup>5</sup>. For inputs of text, pose, and sketch, we set Ti-MGD against an adapted version of SDEdit [33] and Stable Diffusion v1.5 integrated with ControlNet with pose and sketch adapters<sup>6</sup>. Specifically for SDEdit, we follow the approach in [33], 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 [21]. 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 [39]<sup>7</sup> 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.

### B. Evaluation Metrics

To evaluate the realism of generated images, we use the Fréchet Inception Distance (FID) [75] and the Kernel Inception Distance (KID) [76], following the implementation proposed

<sup>3</sup><https://huggingface.co/stabilityai/stable-diffusion-2-inpainting>

<sup>4</sup><https://huggingface.co/runwayml/stable-diffusion-inpainting>

<sup>5</sup>[https://huggingface.co/llyasviel/control\\_v11p\\_sd15\\_openpose](https://huggingface.co/llyasviel/control_v11p_sd15_openpose)

<sup>6</sup>[https://huggingface.co/llyasviel/control\\_v11p\\_sd15\\_softedge](https://huggingface.co/llyasviel/control_v11p_sd15_softedge)

<sup>7</sup><https://huggingface.co/h94/IP-Adapter>

**TABLE III:** Quantitative results of Ti-MGD approach against competitors on the Dress Code Multimodal and VITON-HD Multimodal datasets for both paired and unpaired settings. When considering evaluation metrics, PD stands for Pose Distance, SD for Sketch Distance, and TS for Texture Similarity. Best results are in bold, second best are underlined.

Model	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 ↑
<i>Paired setting</i>																
Stable Diffusion v1.5 [20]	✓				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 [38]	✓	✓			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 [33]	✓	✓	✓		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 [38]	✓	✓	✓		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 [38]+IP-Adapter [39]	✓	✓	✓	✓	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 [21]	✓	✓	✓		5.74	2.11	<b>31.68</b>	4.72	0.188	0.571	10.60	3.26	<b>32.39</b>	5.94	0.243	0.587
<b>Ti-MGD (SDv1)</b>	✓	✓	✓	✓	<b>3.46</b>	<b>0.66</b>	<u>31.26</u>	<b>4.34</b>	<u>0.176</u>	<u>0.578</u>	<u>6.14</u>	<u>0.78</u>	31.16	<b>4.72</b>	<b>0.179</b>	<b>0.616</b>
<b>Ti-MGD (SDv2)</b>	✓	✓	✓	✓	<u>3.79</u>	<u>0.99</u>	31.01	<b>4.34</b>	<u>0.172</u>	<u>0.595</u>	<b>6.04</b>	<u>0.63</u>	<u>31.30</u>	<b>4.67</b>	<u>0.187</u>	<b>0.624</b>
<i>Unpaired setting</i>																
Stable Diffusion v1.5 [20]	✓				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 [38]	✓	✓			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 [33]	✓	✓	✓		8.79	3.67	27.65	<b>6.13</b>	0.354	0.532	15.14	5.99	24.95	7.10	0.446	0.559
ControlNet [38]	✓	✓	✓		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 [38]+IP-Adapter [39]	✓	✓	✓	✓	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 [21]	✓	✓	✓		7.73	2.82	<b>30.04</b>	6.79	0.342	0.554	12.81	3.86	<b>30.75</b>	7.22	0.331	0.578
<b>Ti-MGD (SDv1)</b>	✓	✓	✓	✓	<u>5.69</u>	<u>1.33</u>	29.44	<b>6.19</b>	<u>0.222</u>	<u>0.577</u>	<u>10.18</u>	<u>1.96</u>	28.56	<b>6.59</b>	<b>0.239</b>	<u>0.608</u>
<b>Ti-MGD (SDv2)</b>	✓	✓	✓	✓	<b>5.68</b>	<u>1.32</u>	29.78	6.26	<b>0.218</b>	<u>0.597</u>	<u>9.30</u>	<u>1.26</u>	<u>29.43</u>	<b>6.56</b>	<u>0.247</u>	<b>0.630</b>

in [77]. For assessing how well the images adhere to textual conditioning, we apply the CLIP Score (CLIP-S) [78] from the TorchMetrics library [79], 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  $224 \times 224$  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 the OpenPif-Paf [80] human pose estimation network and compute the  $\ell_2$  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<sup>8</sup>. Then, we paste the segmented garment area onto a white background ( $512 \times 384$ ) and use the PIDInet [67] 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  $64 \times 64$  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.

### C. Experimental Results

**Comparison with Other Methods.** We test our proposed method for each dataset under paired and unpaired settings. In the paired setting, conditions such as text, sketch, and fabric texture correspond to the garment worn by the model in the image. Conversely, in the unpaired setting, the input conditions relate to a different garment. Table III 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 [20] 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 [38] 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 [38] (row 4 in Table III) and SDEdit [33]. 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 conditioned denoising network as the SDEdit backbone. When texture conditioning is added, we compare our Ti-MGD method against ControlNet combined with the

<sup>8</sup><https://github.com/levindabhi/cloth-segmentation>

**TABLE IV:** Category-wise quantitative results of Ti-MGD approach 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.

Model	Modalities				Upper-body						Lower-body						Dresses					
	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 ↑
<i>Paired setting</i>																						
Stable Diffusion v1.5 [20]	✓				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 [38]	✓	✓			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 [33]	✓	✓	✓		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 [38]	✓	✓	✓		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 [38]+IP-Adapter [39]	✓	✓	✓	✓	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 [21]	✓	✓	✓		12.42	3.71	<b>31.90</b>	3.72	0.180	0.547	10.70	2.01	<b>31.10</b>	5.70	0.200	0.567	11.38	1.89	<b>32.02</b>	4.93	0.182	0.592
<b>Ti-MGD (SDv1)</b>	✓	✓	✓	✓	<b>7.92</b>	<b>0.76</b>	<b>31.07</b>	<b>3.38</b>	<b>0.161</b>	<b>0.559</b>	<b>7.22</b>	<b>0.59</b>	<b>30.60</b>	<b>5.28</b>	<b>0.192</b>	<b>0.572</b>	<b>9.14</b>	<b>0.92</b>	<b>32.10</b>	<b>4.52</b>	<b>0.166</b>	<b>0.620</b>
<b>Ti-MGD (SDv2)</b>	✓	✓	✓	✓	<b>8.01</b>	<b>0.97</b>	<b>30.78</b>	<b>3.33</b>	<b>0.160</b>	<b>0.568</b>	<b>7.56</b>	<b>0.95</b>	<b>30.66</b>	<b>5.25</b>	<b>0.191</b>	<b>0.582</b>	<b>9.69</b>	<b>1.63</b>	<b>31.60</b>	<b>4.57</b>	<b>0.164</b>	<b>0.637</b>
<i>Unpaired setting</i>																						
Stable Diffusion v1.5 [20]	✓				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 [38]	✓	✓			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 [33]	✓	✓	✓		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 [38]	✓	✓	✓		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 [38]+IP-Adapter [39]	✓	✓	✓	✓	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 [21]	✓	✓	✓		15.99	4.50	<b>29.76</b>	5.41	0.327	0.532	14.82	2.81	<b>29.96</b>	7.96	0.352	0.561	14.71	3.63	30.41	7.15	0.348	0.568
<b>Ti-MGD (SDv1)</b>	✓	✓	✓	✓	<b>12.33</b>	<b>1.71</b>	<b>28.50</b>	<b>4.59</b>	<b>0.223</b>	<b>0.555</b>	<b>12.93</b>	<b>1.51</b>	<b>29.34</b>	<b>7.52</b>	<b>0.236</b>	<b>0.566</b>	<b>12.65</b>	<b>1.96</b>	<b>30.49</b>	<b>6.62</b>	<b>0.208</b>	<b>0.609</b>
<b>Ti-MGD (SDv2)</b>	✓	✓	✓	✓	<b>12.01</b>	<b>1.32</b>	<b>29.08</b>	<b>4.63</b>	<b>0.220</b>	<b>0.573</b>	<b>13.31</b>	<b>1.90</b>	<b>29.57</b>	<b>7.55</b>	<b>0.231</b>	<b>0.582</b>	<b>12.56</b>	<b>2.02</b>	<b>30.69</b>	<b>6.72</b>	<b>0.203</b>	<b>0.635</b>

IP-Adapter [39]. Notably, ControlNet+IP-Adapter not only boosts texture coherence metrics but also realism, as indicated by improved FID and KID scores. 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 the 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 IV 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, the FID exhibits considerable variance in its results [76]. In contrast, the KID delivers more reliable findings. 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, we show in Fig. 5 a qualitative comparison between our proposed Ti-MGD and ControlNet+IP-Adapter. We report results generated by our method based on SDv1 for a fair comparison. Images in row1-col1 and row2-col2 show respectively the improved adherence of Ti-MGD model compared to the competitor respectively to the texture and sketch input. In row3-col1 and row4-col1, it is possible to see how the proposed method is able to combine text and texture information meaningfully, resulting in better input coherence than its competitor. Finally, images in row1-col2, row2-col1, row3-col2, and row4-col2 demonstrate the ability of Ti-MGD approach to blend the sketch and texture input, generating a realistic garment while following the multimodal input.

**User Study.** To validate our results with human feedback, we

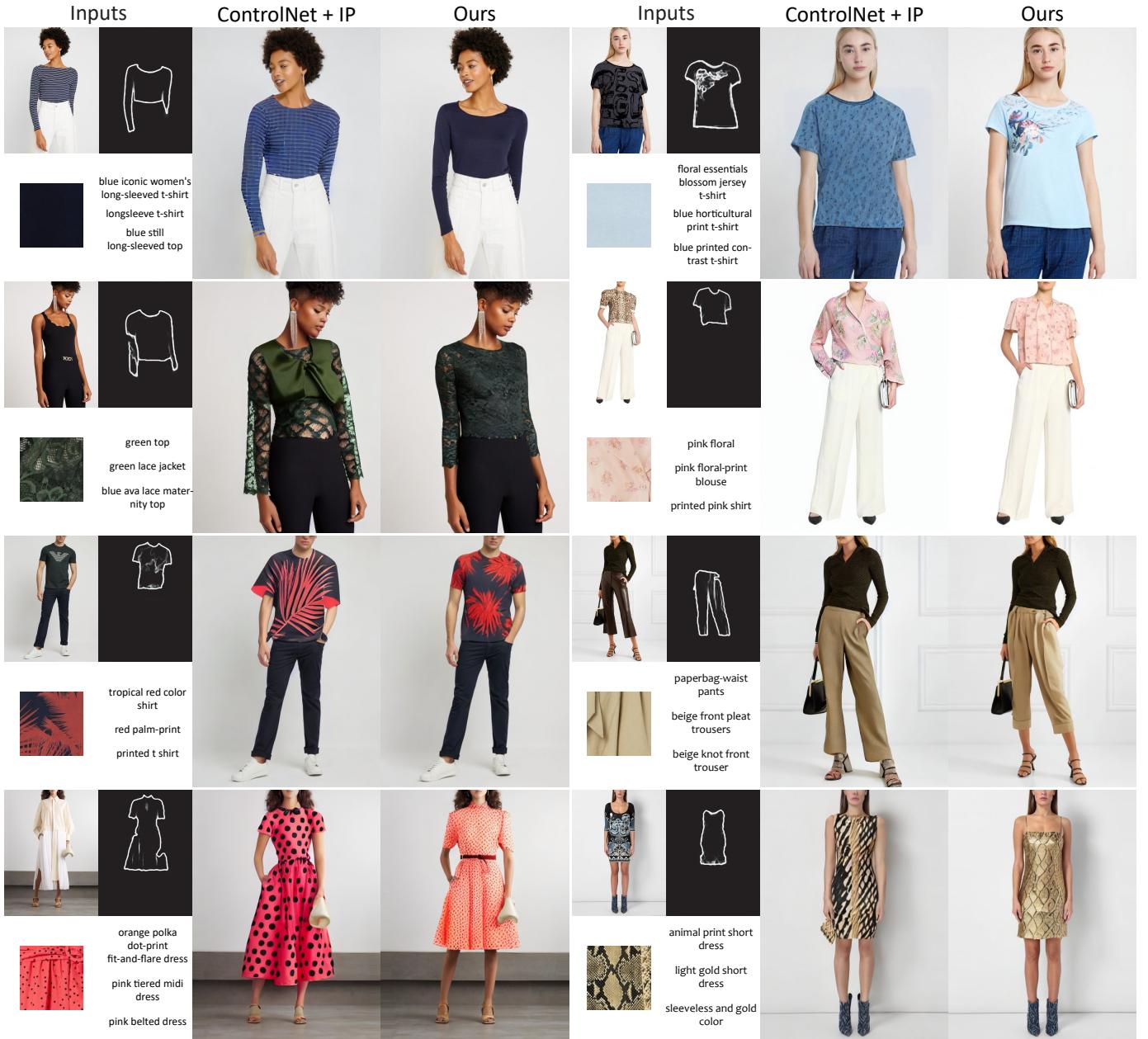
**TABLE V:** User study results on the unpaired setting of both Dress Code Multimodal and VITON-HD Multimodal datasets. We report the percentage of times an image from Ti-MGD is preferred against a competitor. Note that when comparing against ControlNet with all modalities, we employ IP-Adapter to condition 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
VITON HD M.	✓	✓	✓	✓	-	97.67	-	-	84.15	-
	✓				95.83	-	-	84.34	-	-
	✓	✓			.	95.60	-	-	77.38	-
	✓	✓	✓		-	94.25	64.94	-	82.05	78.05
	✓	✓	✓	✓	-	96.62	-	-	89.62	-

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 V reveal that our model consistently outperforms others in both realism and input adherence, confirming the effectiveness of our method. Comparing these results with previously reported quantitative evaluations highlights a correlation between our quantitative metrics and human judgments, both in terms of realism and coherence.

**Varying Input Modalities.** In Table VI, we analyze the behavior of our model with SDv2 as backbone 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). Notice that the text input anchors the CLIP-S metrics of all experiments and makes them comparable in all cases except when incorporating texture input. In this latter case, we observe a slight decrease in the CLIP-S due to the mismatch between the texture and textual information.

Starting from the fully conditioned model (*i.e.* text, pose, sketch, texture), we replace the texture conditioning with text. The decrease in the texture similarity confirms the impact of the texture input on the generation process in both Dress Code Multimodal and VITON-HD Multimodal datasets. This also marginally affects the realism metrics (FID and KID), suggesting that texture information narrows the gap between



**Fig. 5:** Qualitative comparison of images generated using our approach with SDv1 (Ti-MGD) versus ControlNet with IP-Adapter. The smaller images represent the model inputs, while the bigger images depict the generated outputs.

generated and real images. Masking the sketch input leads to a higher sketch distance, underlining the capabilities of our model to incorporate this input effectively. In this case, the pose distance also increases slightly, reflecting the intertwined nature of the sketch and pose information, as the sketch implicitly includes details about the model’s pose. Further masking the pose map input shows a decrease in the pose distance score, while realism metrics remain comparable. These results collectively demonstrate that adding modalities enhances the relative adherence metric, confirming the ability of our model to handle multiple conditions in a distinct manner efficiently. Fig. 6 shows from a qualitative point of view that masking the input modalities can affect the generated image.

**Unconditional Training and Sketch Conditioning.** Table VII 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

**TABLE VI:** 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					
Text	Pose	Sketch	Texture	FID ↓	KID ↓	CLIP-S ↑	PD ↓	SD ↓	TS ↑
✓				6.43	1.35	30.41	7.24	0.404	0.539
✓	✓			6.45	1.50	30.18	6.42	0.374	0.539
✓	✓	✓		6.53	1.87	<b>30.44</b>	6.28	0.225	0.553
✓	✓	✓	✓	<b>5.68</b>	<b>1.32</b>	29.78	<b>6.26</b>	<b>0.218</b>	<b>0.597</b>
Modalities				VITON-HD Multimodal					
Text	Pose	Sketch	Texture	FID ↓	KID ↓	CLIP-S ↑	PD ↓	SD ↓	TS ↑
✓				10.37	1.54	29.37	8.18	0.493	0.559
✓	✓			10.53	1.71	29.31	7.26	0.472	0.560
✓	✓	✓		10.22	1.86	<b>29.62</b>	6.56	0.249	0.581
✓	✓	✓	✓	<b>9.30</b>	<b>1.26</b>	29.43	<b>6.55</b>	<b>0.247</b>	<b>0.630</b>

**TABLE VII:** 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	<b>5.68</b>	<b>1.32</b>	<b>29.78</b>	<b>6.26</b>	<b>0.218</b>	<b>0.597</b>		
0.2	0.0	5.64	1.19	29.30	6.40	0.370	0.584		

a specific texture based on text alone can be challenging. We can clearly see the effect of this asymmetry when analyzing images obtained 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 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 VIII, 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

**TABLE VIII:** Quantitative results of our proposed approach (Ti-MGD) 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-Attn.	Dress Code Multimodal					
	Groups	FID ↓	KID ↓	CLIP-S ↑	PD ↓	SD ↓
<span style="color: blue;">█ █ █ █ █ █</span>	5.63	1.36	27.13	6.23	0.215	0.606
<span style="color: blue;">█ █ █ █ █ █</span> <span style="color: orange;">█</span>	5.63	1.36	27.13	6.25	0.215	0.605
<span style="color: blue;">█ █ █ █ █ █</span> <span style="color: orange;">█ █</span>	<b>5.68</b>	<b>1.32</b>	29.78	<b>6.26</b>	<b>0.218</b>	<b>0.597</b>
<span style="color: blue;">█ █ █ █ █ █</span> <span style="color: orange;">█ █ █</span>	6.31	1.77	30.45	6.25	0.223	0.560
<span style="color: blue;">█ █ █ █ █ █</span> <span style="color: orange;">█ █ █ █</span>	6.53	1.87	30.44	6.28	0.225	0.553
<span style="color: orange;">█ █ █ █ █ █</span> <span style="color: blue;">█</span>	6.53	1.87	30.44	6.27	0.225	0.553
<span style="color: orange;">█ █ █ █ █ █</span> <span style="color: blue;">█ █</span>	6.22	1.77	28.15	6.21	0.217	0.572
<span style="color: orange;">█ █ █ █ █ █</span> <span style="color: blue;">█ █ █</span>	5.62	1.38	27.22	6.22	0.215	0.604

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. It is possible to note that performing texture conditioning on Groups 3 and 2 allows the generation of an image in line with both textual and texture input information.

**Limitations and Failure Cases.** Fig. 9 shows some failure cases of the proposed approach. In the example in the first row and first column, we can see that the body shape of the generated model does not match the original one. We attribute this discrepancy to the inability of keypoints to describe body shape fully. In this sense, a potential future direction could involve augmenting the pose input with dense or 3D information. The image in the second row and first column demonstrates that when the sketch comprises distinct areas, each delineated by an edge, it could happen that only one delimited area of the generated garment is conditioned on texture. To solve this issue, a future direction could involve introducing spatial control for texture conditioning. Examples in the second column highlight the dependence of our model performance on the provided sketch. When the geometric warping module fails to generate a sketch able to fit the model’s shape, the generation task also fails, resulting in unwanted artifacts. The examples presented in the third column reveal that our model sometimes fails to generate hands and text accurately when they occupy a limited area within the source image. This behavior is intrinsic in latent diffusion models [20] and stems from the high compression nature of the latent space.

## VI. CONCLUSION

In this work, we have presented Ti-MGD, a novel approach for multimodal-conditioned fashion image editing in which the image generation process is conditioned with multiple modalities, such as text, body pose, garment sketch, and fabric texture. This is achieved by extending latent diffusion models to incorporate these different modalities and modifying the structure of the denoising network to take multimodal inputs. To effectively incorporate texture information, we have leveraged textual inversion techniques and proposed to combine



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

text and texture features through the cross-attention layers of the denoising network. The proposed texture conditioning method enables fine-grained control over the generated images without adding denoising network parameters. To validate our approach, we have also extended two existing fashion datasets with multimodal annotations using a semi-automatic procedure. Our comprehensive experiments on standard and newly proposed metrics validate the effectiveness of Ti-MGD, outperforming state-of-the-art methods in terms of realism and coherence with multimodal inputs. Conclusively, the proposed approach not only sets a new standard for multimodal-conditioned fashion image editing but also opens avenues for further research at the intersection of computer vision and fashion. The presented results represent one of the first successful attempts to imitate the work of designers in the creative process of fashion design and could be a starting point for the widespread adoption of diffusion models in the creative industries, oversight by human input.

## 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 (grant number 951911 - AI4Media), and by the PRIN project “CREATIVE: CRoss-modal understanding and gEnerATIon of Visual and tExtual content” (CUP B87G22000460001), co-funded by the Italian Ministry of University.

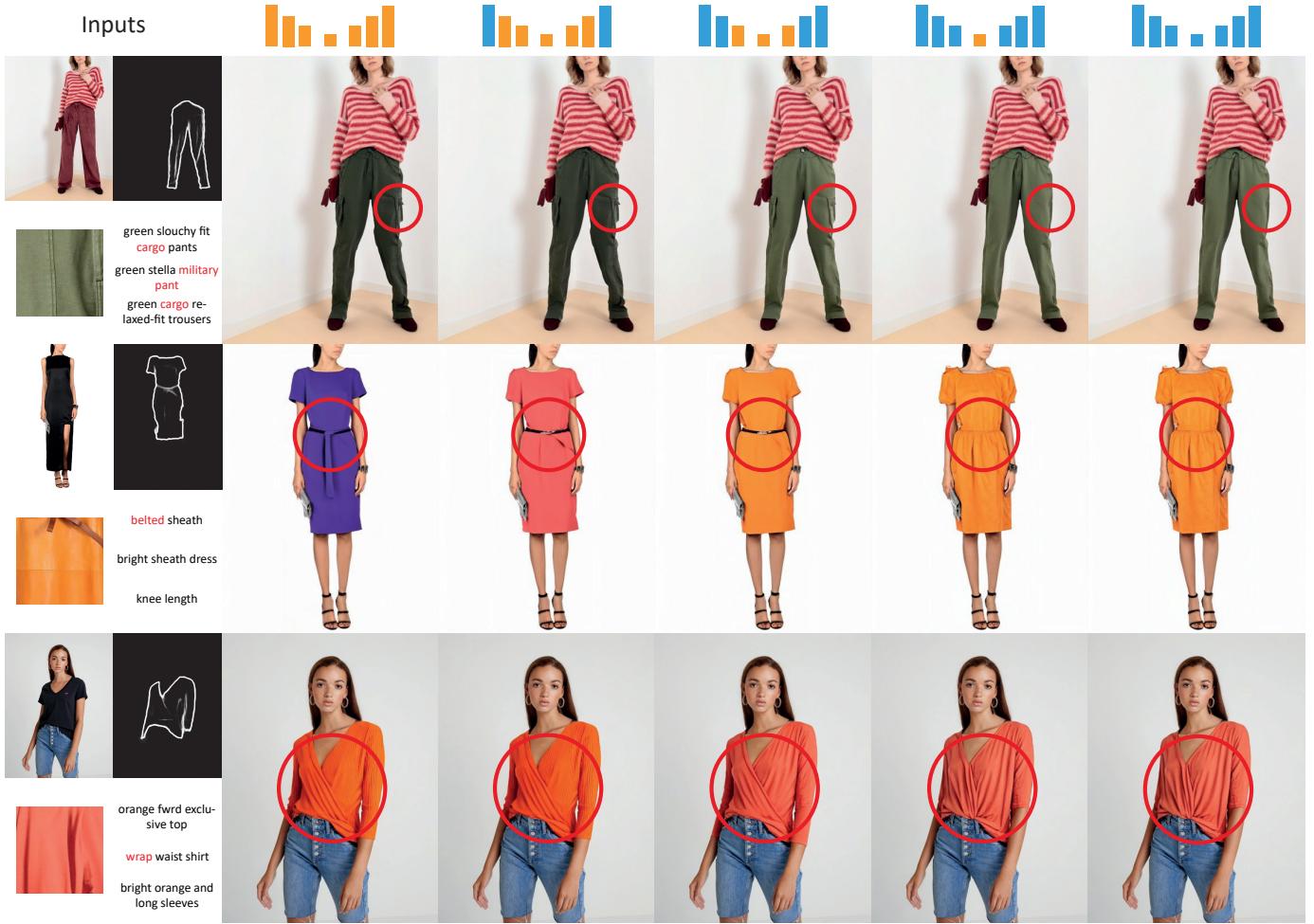
## REFERENCES

- [1] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, “Where to buy it: Matching street clothing photos in online shops,” in *ICCV*, 2015.
- [2] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, “DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations,” in *CVPR*, 2016.
- [3] G. Cartella, A. Baldrati, D. Morelli, M. Cornia, M. Bertini, and R. Cucchiara, “OpenFashionCLIP: Vision-and-Language Contrastive Learning with Open-Source Fashion Data,” in *ICIAP*, 2023.



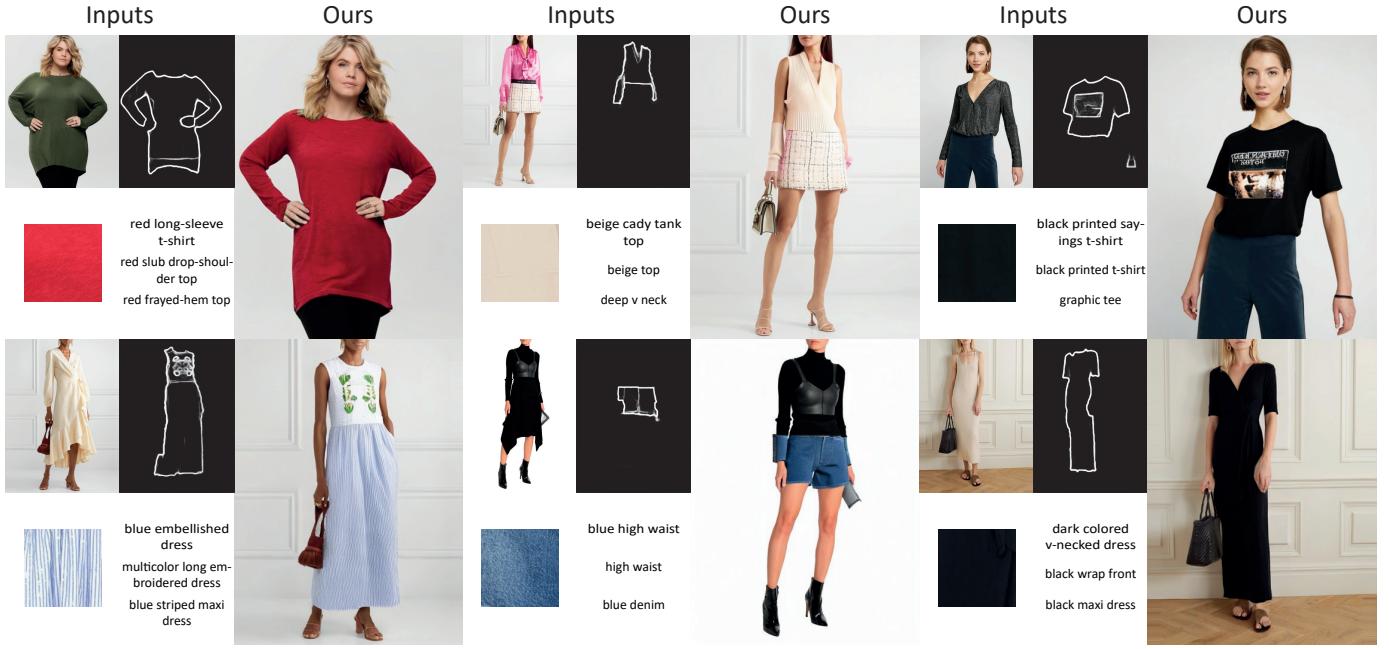
**Fig. 7:** Qualitative examples of images generated by our proposed approach (Ti-MGD) when varying the sketch conditioning rate. We report the such rate on top of each column. Results are reported on sample images from the Dress Code Multimodal dataset.

- [4] D. Morelli, M. Cornia, and R. Cucchiara, “FashionSearch++: Improving consumer-to-shop clothes retrieval with hard negatives,” in *CEUR Workshop Proceedings*, 2021.
- [5] W.-L. Hsiao and K. Grauman, “Creating capsule wardrobes from fashion images,” in *CVPR*, 2018.
- [6] G. Cucurull, P. Taslakian, and D. Vazquez, “Context-aware visual compatibility prediction,” in *CVPR*, 2019.
- [7] R. Sarkar, N. Bodla, M. I. Vasileva, Y.-L. Lin, A. Beniwal, A. Lu, and G. Medioni, “OutfitTransformer: Learning Outfit Representations for Fashion Recommendation,” in *WACV*, 2023.
- [8] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, “VITON: An Image-Based Virtual Try-On Network,” in *CVPR*, 2018.
- [9] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang, “Toward characteristic-preserving image-based virtual try-on network,” in *ECCV*, 2018.
- [10] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, and P. Luo, “Towards photo-realistic virtual try-on by adaptively generating-preserving image content,” in *CVPR*, 2020.
- [11] D. Morelli, M. Fincato, M. Cornia, F. Landi, F. Cesari, and R. Cucchiara, “Dress Code: High-Resolution Multi-Category Virtual Try-On,” in *ECCV*, 2022.
- [12] S. Choi, S. Park, M. Lee, and J. Choo, “VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization,” in *CVPR*, 2021.
- [13] D. Morelli, A. Baldriati, G. Cartella, M. Cornia, M. Bertini, and R. Cucchiara, “LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On,” in *ACM Multimedia*, 2023.
- [14] S. Zhu, R. Urtasun, S. Fidler, D. Lin, and C. Change Loy, “Be Your Own Prada: Fashion Synthesis with Structural Coherence,” in *ICCV*, 2017.
- [15] Y. Jiang, S. Yang, H. Qiu, W. Wu, C. C. Loy, and Z. Liu, “Text2human: Text-driven controllable human image generation,” *ACM Transactions on Graphics*, vol. 41, no. 4, pp. 1–11, 2022.
- [16] M. Pernuš, C. Fookes, V. Štruc, and S. Dobrišek, “FICE: Text-Conditioned Fashion Image Editing With Guided GAN Inversion,” *arXiv preprint arXiv:2301.02110*, 2023.
- [17] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” in *NeurIPS*, 2020.
- [18] P. Dhariwal and A. Nichol, “Diffusion Models Beat GANs on Image Synthesis,” in *NeurIPS*, 2021.
- [19] A. Q. Nichol and P. Dhariwal, “Improved denoising diffusion probabilistic models,” in *ICML*, 2021.
- [20] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis With Latent Diffusion Models,” in *CVPR*, 2022.
- [21] A. Baldriati, D. Morelli, G. Cartella, M. Cornia, M. Bertini, and R. Cucchiara, “Multimodal Garment Designer: Human-Centric Latent Diffusion Models for Fashion Image Editing,” in *ICCV*, 2023.
- [22] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, “AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks,” in *CVPR*, 2018.
- [23] M. Zhu, P. Pan, W. Chen, and Y. Yang, “DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-to-Image Synthesis,” in *CVPR*, 2019.
- [24] H. Zhang, J. Y. Koh, J. Baldridge, H. Lee, and Y. Yang, “Cross-modal contrastive learning for text-to-image generation,” in *CVPR*, 2021.
- [25] M. Tao, H. Tang, F. Wu, X.-Y. Jing, B.-K. Bao, and C. Xu, “DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis,” in *CVPR*, 2022.
- [26] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, “GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models,” in *ICML*, 2022.
- [27] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical Text-Conditional Image Generation with CLIP Latents,” *arXiv preprint arXiv:2204.06125*, 2022.
- [28] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, “SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis,” *arXiv preprint arXiv:2307.01952*, 2023.
- [29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning Transferable Visual Models From Natural Language Supervision,” in *ICML*, 2021.
- [30] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, R. Gontijo-Lopes, B. K. Ayan, T. Salimans *et al.*, “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding,” in *NeurIPS*, 2022.



**Fig. 8:** Qualitative examples of our proposed approach (Ti-MGD) on Dress Code Multimodal and VITON-HD Multimodal datasets 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. 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.

- [31] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *JMLR*, vol. 21, no. 1, pp. 5485–5551, 2020.
- [32] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, “ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models,” in *ICCV*, 2021.
- [33] C. Meng, Y. H. adnd Yang Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, “SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations,” in *ICLR*, 2022.
- [34] T. Wang, T. Zhang, B. Zhang, H. Ouyang, D. Chen, Q. Chen, and F. Wen, “Pretraining is All You Need for Image-to-Image Translation,” *arXiv preprint arXiv:2205.12952*, 2022.
- [35] C. Mou, X. Wang, L. Xie, J. Zhang, Z. Qi, Y. Shan, and X. Qie, “T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models,” *arXiv preprint arXiv:2302.08453*, 2023.
- [36] S.-I. Cheng, Y.-J. Chen, W.-C. Chiu, H.-Y. Tseng, and H.-Y. Lee, “Adaptively-Realistic Image Generation from Stroke and Sketch with Diffusion Model,” in *WACV*, 2023.
- [37] A. Voynov, K. Aberman, and D. Cohen-Or, “Sketch-Guided Text-to-Image Diffusion Models,” in *ACM SIGGRAPH*, 2023.
- [38] L. Zhang and M. Agrawala, “Adding Conditional Control to Text-to-Image Diffusion Models,” in *ICCV*, 2023.
- [39] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, “Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models,” *arXiv preprint arXiv:2308.06721*, 2023.
- [40] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, “An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion,” in *ICLR*, 2023.
- [41] I. Han, S. Yang, T. Kwon, and J. C. Ye, “Highly Personalized Text Embedding for Image Manipulation by Stable Diffusion,” *arXiv preprint arXiv:2303.08767*, 2023.
- [42] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, “DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation,” in *CVPR*, 2023.
- [43] G. Daras and A. G. Dimakis, “Multiresolution Textual Inversion,” in *NeurIPS Workshops*, 2022.
- [44] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, “Null-text Inversion for Editing Real Images using Guided Diffusion Models,” in *CVPR*, 2023.
- [45] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, “Multi-Concept Customization of Text-to-Image Diffusion,” in *CVPR*, 2023.
- [46] M. Fincato, F. Landi, M. Cornia, F. Cesari, and R. Cucchiara, “VITON-GT: An Image-based Virtual Try-On Model with Geometric Transformations,” in *ICPR*, 2020.
- [47] T. Issenhuth, J. Mary, and C. Calauzenes, “Do Not Mask What You Do



**Fig. 9:** Failure cases of our proposed approach (Ti-MGD) on the Dress Code Multimodal and VITON-HD Multimodal datasets.

- Not Need to Mask: A Parser-Free Virtual Try-On,” in *ECCV*, 2020.
- [48] Y. Ge, Y. Song, R. Zhang, C. Ge, W. Liu, and P. Luo, “Parser-free virtual try-on via distilling appearance flows,” in *CVPR*, 2021.
- [49] B. Fele, A. Lampe, P. Peer, and V. Struc, “C-VTON: Context-Driven Image-Based Virtual Try-On Network,” in *WACV*, 2022.
- [50] X. Li, M. Kampffmeyer, X. Dong, Z. Xie, F. Zhu, H. Dong, X. Liang *et al.*, “WarpDiffusion: Efficient Diffusion Model for High-Fidelity Virtual Try-on,” *arXiv preprint arXiv:2312.03667*, 2023.
- [51] J. Gou, S. Sun, J. Zhang, J. Si, C. Qian, and L. Zhang, “Taming the Power of Diffusion Models for High-Quality Virtual Try-On with Appearance Flow,” in *ACM Multimedia*, 2023.
- [52] J. Zeng, D. Song, W. Nie, H. Tian, T. Wang, and A. Liu, “CAT-DM: Controllable Accelerated Virtual Try-on with Diffusion Model,” *arXiv preprint arXiv:2311.18405*, 2023.
- [53] J. Kim, G. Gu, M. Park, S. Park, and J. Choo, “StableVITON: Learning Semantic Correspondence with Latent Diffusion Model for Virtual Try-On,” *arXiv preprint arXiv:2312.01725*, 2023.
- [54] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *ICML*, 2015.
- [55] A. Baldrati, L. Agnolucci, M. Bertini, and A. Del Bimbo, “Zero-Shot Composed Image Retrieval with Textual Inversion,” *ICCV*, 2023.
- [56] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, “Resolution-Robust Large Mask Inpainting With Fourier Convolutions,” in *WACV*, 2022.
- [57] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, “RePaint: Inpainting Using Denoising Diffusion Probabilistic Models,” in *CVPR*, 2022.
- [58] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, “MAT: Mask-Aware Transformer for Large Hole Image Inpainting,” in *CVPR*, 2022.
- [59] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, K. Kreis, M. Aittala, T. Aila, S. Laine, B. Catanzaro *et al.*, “eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers,” *arXiv preprint arXiv:2211.01324*, 2022.
- [60] A. Voynov, Q. Chu, D. Cohen-Or, and K. Aberman, “P+: Extended Textual Conditioning in Text-to-Image Generation,” *arXiv preprint arXiv:2303.09522*, 2023.
- [61] O. Avrahami, T. Hayes, O. Gafni, S. Gupta, Y. Taigman, D. Parikh, D. Lischinski, O. Fried, and X. Yin, “SpaText: Spatio-Textual Representation for Controllable Image Generation,” in *CVPR*, 2023.
- [62] F. Bianchi, J. Tagliabue, and B. Yu, “Query2Prod2Vec: Grounded word embeddings for eCommerce,” in *NAACL*, 2021.
- [63] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, and R. Feris, “Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback,” in *CVPR*, 2021.
- [64] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis, “Automatic Spatially-Aware Fashion Concept Discovery,” in *ICCV*, 2017.
- [65] M. Wortsman, G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, R. G. Lopes, H. Hajishirzi, A. Farhadi, H. Namkoong *et al.*, “Robust fine-tuning of zero-shot models,” in *CVPR*, 2022.
- [66] C. Schuhmann, R. Beaumont, R. Vencu, C. W. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. R. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmareczyk, and J. Jitsev, “LAION-5B: An open large-scale dataset for training next generation image-text models,” in *NeurIPS*, 2022.
- [67] Z. Su, W. Liu, Z. Yu, D. Hu, Q. Liao, Q. Tian, M. Pietikäinen, and L. Liu, “Pixel difference networks for efficient edge detection,” in *ICCV*, 2021.
- [68] I. Rocco, R. Arandjelovic, and J. Sivic, “Convolutional neural network architecture for geometric matching,” in *CVPR*, 2017.
- [69] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *MICCAI*, 2015.
- [70] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” in *ICLR*, 2019.
- [71] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu, “Mixed Precision Training,” in *ICLR*, 2018.
- [72] D. Hendrycks and K. Gimpel, “Gaussian Error Linear Units (GELUs),” *arXiv preprint arXiv:1606.08415*, 2016.
- [73] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *JMLR*, vol. 15, no. 56, pp. 1929–1958, 2014.
- [74] J. Song, C. Meng, and S. Ermon, “Denoising Diffusion Implicit Models,” in *ICLR*, 2021.
- [75] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a Nash equilibrium,” in *NeurIPS*, 2017.
- [76] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, “Demystifying MMD GANs,” in *ICLR*, 2018.
- [77] G. Parmar, R. Zhang, and J.-Y. Zhu, “On Aliased Resizing and Surprising Subtleties in GAN Evaluation,” in *CVPR*, 2022.
- [78] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, “CLIPScore: A Reference-free Evaluation Metric for Image Captioning,” in *EMNLP*, 2021.
- [79] N. S. Detlefsen, J. Borovec, J. Schock, A. H. Jha, T. Koker, L. Di Liello, D. Stanci, C. Quan, M. Grechkin, and W. Falcon, “TorchMetrics: Measuring Reproducibility in PyTorch,” *J. of Open Source Software*, vol. 7, no. 70, p. 4101, 2022.
- [80] S. Kreiss, L. Bertoni, and A. Alahi, “OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association,” *IEEE Trans. on ITS*, vol. 23, no. 8, pp. 13 498–13 511, 2021.