Controllable Garment Generation with Multi-Modal Diffusion Guidance

Sanhita Pathak[1]
Sanhita.Pathak@dbst.iitd.ac.in
Vinay Kaushik[2]
vkaushik@iiitsonepat.ac.in
Brejesh Lall[3]
brejesh@ee.iitd.ac.in

[1] BSTTM, IIT Delhi
New Delhi, India
[2] Dept. of CSE, IIIT Sonepat
Sonepat, India
[3] Dept. of EE, IIT Delhi
New Delhi, India

Abstract

We address the task of controllable garment image generation, where the goal is to synthesize realistic clothing images from incomplete and interpretable inputs such as sketches, color cues, text prompts, and fabric textures. This task poses unique challenges in aligning high-level semantics with spatial structure while preserving visual fidelity. We propose **GenWear**, a unified conditional diffusion framework built atop a frozen Paint-by-Example (PBE) denoising UNet, augmented with learnable modules for modality-specific control. We introduce SketchCtrl, a ControlNet based module which injects multi-scale spatial features from sketches via zero-convolution, ensuring structural fidelity. ColorFiLM employs feature-wise linear modulation to steer global color tone without disrupting pretrained activations. A BLIP-2-based text adapter biases the sketch encoder with semantic priors to resolve ambiguities, while a cross-attentive texture adapter injects local fabric cues into the decoder for material realism. These modules operate synergistically, enabling disentangled control without modifying the core diffusion backbone. Unlike prior works, GenWear does not require a full reference image at test time, treating the sketch as both spatial and semantic guide. Experiments on VITON-HD dataset demonstrate that our approach achieves state-of-the-art quality and controllability across diverse garment types and modalities.

1 Introduction

Fine-grained garment generation plays a vital role in fashion synthesis [23], enabling precise control over styling and appearance. From a sustainability perspective, traditional garment design often incurs significant material waste due to extensive physical prototyping prior to finalization. In contrast, realistic digital stylization offers an efficient and eco-friendly alternative, allowing designers to visualize modifications without sample production. This not only enhances creative flexibility for designers but also enables greater personalization for end users. While garment generation is inherently tied to virtual try-on systems[9, 31], dedicated frameworks that focus on detailed garment synthesis remain largely unexplored.

Garment generation framework can significantly enrich virtual try-on systems by producing realistic and diverse garments, thereby improving downstream applications such as e-commerce visualization, fashion design, and personalized styling. Garments often differ in subtle but important ways—e.g., neckline shape, sleeve length, button placement, pocket types[9]. Capturing this diversity requires a model that interprets and synthesizes fine structural cues with high fidelity. They also have many interdependent properties (e.g., color influences perceived texture, texture modifies structure). Integrating sketch, color, texture, and semantic text in a way that avoids mode collapse like GANs or conflicting signals is non-trivial. Often achieving global coherence (e.g., symmetrical sleeves, centered prints) while allowing local variability (e.g., fabric texture, stitch details) is difficult in garment generation and is further challenging when done in a single stage pipeline. Many prior models (PBE[32], Ti-MGD[2]) rely on a complete garment reference image. This poses a dependency on the availability of complete image. Garment datasets are often curated for try-on, not generation. They lack garment categories, edge annotations, or semantic labels needed by conventional conditional generators. Furthermore, garment generation often lacks flexibility in terms of the provided conditioning.

To ensure high fidelity, we train our model on the publicly available VITON-HD dataset [9], which offers high-resolution images of real garments. Leveraging this dataset, we automatically extract key guidance modalities: garment color vectors using segmentation masks, sketch representations via informative edge drawings, and textual descriptions using BLIP2 [15], including both captions and corresponding embeddings. Our pipeline is fully automated and requires no human annotations, making it applicable to any garment dataset equipped with segmentation masks (which can also be generated using standard segmentation techniques). We adopt a multimodal, single-stage framework that integrates color, texture, sketch, and text, to guide high-quality garment synthesis, capturing both structural and stylistic nuances.

In this work, we address the challenge of fine-grained garment generation by proposing a unified, multimodal framework. Structural details such as buttons, collars, and pockets are primarily dictated by sketch inputs, which serve as the spatial control signal via a ControlNet-style module. This allows for direct manipulation of garment shape through designer-drawn outlines, maintaining spatial fidelity within a single-stage generation process. To integrate semantic and visual cues, we introduce modality-specific adapter pathways: sketch control via ControlNet, global color conditioning through a Color Adapter employing Feature-wise Linear Modulation (FiLM), and a BLIP-2 embedding to embed textual descriptions. Textual descriptions further enhance semantic alignment during generation. Local garment texture patch is utilized to provide local fine-grained visual detail aligned with the global structure, and is processed through a dedicated Texture Adapter. All modalities are injected concurrently into a single-stage diffusion PBE UNet, enabling end-to-end multimodal conditioning without requiring multiple generation stages. We repurpose the VITON-HD try-on dataset by auto-generating sketches, masks, and texture crops, avoiding the need for garment class labels or pose annotations. Our formulation as an inpainting problem enables flexible spatial control and supports scalable training on real-world data. Our key contributions are summarized as follows:

- We present **GenWear**, a unified conditional diffusion architecture for garment image generation, which reformulates multi-modal control, utilizing sketch, color, texture, and language, as modular conditioning streams injected into a single frozen PBE UNet, eliminating the need for reference images at test time.
- We propose **SketchCtrl**, a dual-path sketch conditioning module that combines ControlNet style spatial feature injection with CLIP-based semantic alignment, enabling fine-grained structural control from design sketches alone.
- We introduce **ColorFiLM** and **TexAdapter**, two separate yet disentangled control

mechanisms for global appearance and local texture detail, implemented via FiLM modulation and cross-attention texture injection.
- We demonstrate strong controllability and visual quality on VITON-HD multimodal through extensive ablations and benchmarks, outperforming baselines while offering interpretable and disentangled control.

2 Related works

2.1 GAN based methods

Some approaches [5, 6, 8, 17] for garment generation utilise GANS which would translate sketch to image, where structural maps would lead the spatial garment mapping. DeepPortraitDrawing [29] extends this strategy to full-body synthesis using a local-to-global pipeline. Similarly, Interactive frameworks [11] introduce gating mechanisms to support user-driven editing, while DALColor [10] combines WGAN-GP with line-art colorization for refined appearance control. However, these methods [5, 6, 11] are limited to single-condition generation and are difficult to apply in real-world fashion scenarios that demand multi-conditional controllability. Moreover, GAN-based approaches[20, 21, 26] are prone to adversarial training instability and image artifacts, often compromising realism.

2.2 Diffusion based methods

Diffusion models [12] have made significant strides in conditional image generation, due to their iterative denoising mechanism and flexible control capacity. Plugin-based methods such as IP-Adapter [33], ControlNet [34], and BLIP-Diffusion [14] inject external conditions into pre-trained diffusion models via lightweight modules and enhance controllability. Parallel architectures [18, 31] further improve control precision by incorporating reference guided features during generation. Magic Clothing[7] utilised additional garment generation diffusion to preserve the garment details and maintain faithfulness to the text prompts. In the fashion domain, DiffCloth [16] enables localized garment editing via part-specific textual prompts, allowing independent control over different regions (e.g., sleeves, collars). ARMANI[36], a garment generation method converts images into tokens using a cross-modal codebook and models their distribution with a Transformer based on control signals. It incorporates textual tokens to enable image generation from inputs like text, sketches, or partial images. GarmentAligner[35], a text-to-garment diffusion model trained with retrieval-augmented multi-level corrections. It is a single modality based method which takes text inputs to generate the garments. HieraFashDiff[30], is a fashion editing based method that uses two successive stages and supports fashion design generation and fine-grained local editing. The method utilises text as a conditioning modality for fashion editing. DreamPaint[25], is an inpainting based framework that can inpaint any e-commerce product on any user-provided context image. These methods condition the garment. The methods with garment generation framework propose various methods to provide uni/bi modal conditioning control while garment generation which introduces lack in accuracy and persistent artefacts. Which motivated us to propose a multimodal framework to provide garment generation control with improved realism.

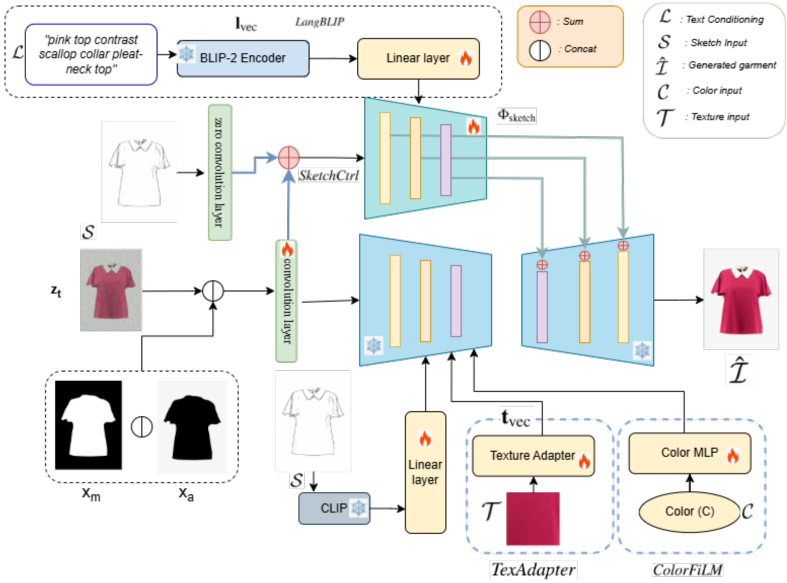

Figure 1: Architecture diagram of our proposed methodology shows the multimodal conditioning based garment generation leveraging multiple adapters with Diffusion UNet.

3 Methodology

3.1 Problem Formulation and Approach Overview

We address the task of controllable garment generation from designer-specified cues by introducing *GenWear*, a single-stage, multi-modal diffusion framework capable of synthesizing high-quality garment images with interpretable and flexible controls. Given a garment sketch $\mathcal{S} \in \mathbb{R}^{1\times H\times W}$, a global color vector $\mathcal{C} \in \mathbb{R}^3$, a local texture crop $\mathcal{T} \in \mathbb{R}^{3\times h\times w}$, and a textual garment description $\mathcal{L}$, the model aims to generate a photo-realistic garment image $\hat{\mathcal{I}} \in \mathbb{R}^{3\times H\times W}$ that faithfully reflects structural intent, fabric appearance, and semantic style.

At the core of GenWear is a frozen Paint-by-Example (PBE) UNet [32], originally designed for reference-based inpainting. Instead of relying on a full garment reference image, we adapt this backbone to synthesize garments using a sketch as the primary structural cue, augmented with optional color, texture, and text inputs. Each module injects complementary information into the generative process: the sketch $\mathcal{S}$ provides detailed structural cues such as seams, collars, and silhouettes; the color vector $\mathcal{C}$ governs the garment's global chromatic properties; the texture crop $\mathcal{T}$ supplies localized fabric information; and the text $\mathcal{L}$ introduces semantic concepts like garment category, design, or other context. This design choice eliminates the need for complete reference images during inference and supports flexible garment ideation workflows.

These conditioning signals are injected through different modules for each modality. Specifically, we employ a ControlNet based sketch encoder *SketchCtrl*, to extract hierarchical spatial features from sketch $\mathcal{S}$ and injects them into the UNet decoder via zero-convolution adapters at multiple scales. To modulate visual appearance, the color vector $\mathcal{C}$ is transformed via a Feature-wise Linear Modulation module *ColorFiLM*, into affine parameters (γ, β), which adjust the scale and shift of intermediate feature maps across the network. For higher-level control, we encode texture $\mathcal{T}$ using a texture adapter *TexAdapter*, motivated from IP-adapter, and embed garment textual description $\mathcal{L}$ using a BLIP-2 encoder [15] embedding *LangBLIP*, along with *ColorFiLM* embedding, yielding a joint style embedding that is injected into the PBE's decoder via cross-attention layers for efficient conditioning.

The diffusion process follows a standard denoising objective [12], where a Gaussian-noised image $\mathcal{I}_t$ is progressively refined into garment image $\hat{\mathcal{I}}$, through a series of conditional updates, similar to PBE [32]. The diffusion model takes the noisy latent z_t, in timestep t and concatenated garment mask x_m and garment agnostic x_a as its inputs, similar to PBE [32]. *GenWear* enables joint reasoning over local and global signals in a single-stage pipeline, eliminating the need for traditional sequential garment layout and texture synthesis. By disentangling conditioning pathways for structure, color, texture, and semantics, GenWear offers an interpretable and modular pipeline for garment synthesis. Designers can iteratively adjust individual aspects of a garment—such as exploring fabric variants or colorways—without retraining the model. The following sections describe in detail the design and integration of each conditioning module depicted in Figure 1.

3.2 Sketch-Based Structural Control

We introduce *SketchCtrl*, a dual-path conditioning module that leverages design sketches $\mathcal{S} \in \mathbb{R}^{1\times H\times W}$ to guide garment generation both spatially and semantically. Sketches provide fine-grained structural cues, such as collar, sleeve, and hemline details, while remaining lightweight and editable, making them an ideal interface for controllable synthesis. For spatial control, we employ a ControlNet encoder [34] to extract multi-scale features $\{\mathcal{F}^l_{\text{sketch}}\}$ aligned with the SDXL 1.5 encoder. These features are injected via zero-convolution adapters into the frozen PBE UNet's decoder [32], influencing the denoising process at each layer to enforce structural fidelity.

To complement spatial alignment, we encode $\mathcal{S}$ with a frozen CLIP image encoder [24], project the global embedding utilizing a learnable linear layer, and inject it into the UNet's cross-attention layers as a semantic substitute for the reference image. This allows our model to operate without requiring any reference garment image during inference. *SketchCtrl* enables structurally and semantically grounded generation using only a sketch, serving as the primary control signal for garment generation.

3.3 Color-Based Feature Modulation

To enable explicit color control in garment synthesis, we introduce *ColorFiLM*, a lightweight module that injects global color semantics via feature-wise linear modulation (FiLM)[3]. Standard diffusion models often fail to adhere to color constraints in inference, especially when dealing with diverse textures and materials[13, 28]. Our proposed approach provides direct modulation without altering the frozen PBE backbone. Given a binary garment mask $\mathcal{M} \in \{0,1\}^{H\times W}$ and ground truth image $\mathcal{I}$, we compute the color vector $\mathcal{C} \in \mathbb{R}^3$ as the average RGB value over the masked region:

$$\mathcal{C} = \frac{1}{|\mathcal{M}|} \sum_{(i,j):\mathcal{M}_{i,j}=1} \mathcal{I}_{i,j} \tag{1}$$

At inference, $\mathcal{C}$ may be user-defined or sampled from a reference patch. We project $\mathcal{C}$, through an MLP to generate affine parameters $\gamma(\mathcal{C})$ and $\beta(\mathcal{C})$, applied to intermediate PBE encoder's features F as:

$$\text{FiLM}(F) = \gamma(\mathcal{C}) \cdot F + \beta(\mathcal{C}) \tag{2}$$

Figure 2: Preprocessing inputs to our proposed method obtained from VITON-HD Dataset.

These are injected at last 3 decoder blocks, where global attributes like tone and shading are formed. Compared to palette-based attention[19] or conditional normalization[27], ColorFiLM introduces no spatial overhead and maintains full compatibility with frozen weights. ColorFiLM enables smooth interpolation and disentangled control, supporting tasks such as color variant generation and interactive design with minimal computational cost.

3.4 Textual Semantics Adapter

Sketches may lack semantic specificity, making it difficult to infer garment type/style solely from structure. To resolve this ambiguity, we introduce *LangBLIP*, a text-based adapter to inject semantic priors into the generation process without overriding spatial conditioning. We use a pretrained BLIP-2 encoder [15] to extract a fixed-length embedding $f_{\text{text}} \in \mathbb{R}^d$ from a free-form prompt $\mathcal{L}$ (e.g., "red cotton blouse with puffed sleeves"), provided by MGD[1]. Rather than conditioning the image synthesis directly, we inject f_{text} into the ControlNet sketch encoder by introducing a learnable linear projection layer and adding it as a conditioning to the ControlNet's encoder blocks. This modulates spatial features $\{\mathcal{F}^l_{\text{sketch}}\}$ with high-level textual semantics during feature extraction. LangBLIP operates in a lightweight, supportive capacity, guiding structure-aware generation without dominating it. Its modular design preserves compatibility with the frozen PBE backbone and avoids reliance on prompt tuning or cross-attention. This enables controllable semantic refinement over sketches, capturing distinctions such as garment category, formality, or seasonal intent.

3.5 Local Texture Injection via Adapter

To capture fine-grained material properties in garment synthesis, we introduce *TexAdapter*, a lightweight module inspired by the adapter-based conditioning paradigm of T2I-Adapter [22] and IP-Adapter [33]. Similar to these works, TexAdapter injects texture in the form of auxiliary visual information into the denoising Unet. Given a reference patch $\mathcal{T} \in \mathbb{R}^{3\times h\times w}$, we extract a texture embedding $f_{\text{tex}} \in \mathbb{R}^d$ using a 3 layer CNN encoder (3x3,3x3, 1x1) with residual blocks followed by a linear layer for feature projection. This embedding encodes localized fabric cues such as weave, gloss, or print. Following IP-Adapter [33], we inject f_{tex} via multi-head cross-attention layers into selected blocks of the frozen PBE UNet's encoder, where fine spatial details are reconstructed. UNet features attend to the texture embedding using a key-query-value mechanism, enabling texture-aware modulation without altering structural or color guidance. Unlike CLIP-guided or prompt-based texture hallucination, *TexAdapter* enables explicit and disentangled control using visual patches. It requires no semantic alignment between sketch and texture, allowing flexible material transfer across structures (e.g., lace on a blazer sketch). This approach maintains full compatibility with the frozen denoising backbone while enhancing the realism and diversity of garment textures under user control.

4 Experiments

4.1 Implementation Details and Dataset

To evaluate the effectiveness of our unified conditional diffusion framework, we conduct experiments on the VITON-HD dataset [9], a high-resolution benchmark for garment image synthesis and virtual try-on. VITON-HD provides paired images of upper body garments with diverse styles, fabrics, and structural variations, making it well-suited for controlled garment generation. For each garment image, we extract the following conditioning inputs: Sketch $\mathcal{S}$ is obtained by applying InformativeDrawings[4] to garment images. These sketches capture structural outlines such as collars, sleeves, and hemlines in detail. Color vector $\mathcal{C}$ is computed as the average RGB value over the segmented garment region. Texture crop $\mathcal{T}$ is generated by generating by picking random 64×64 patches, sampled from the garment region to serve as local appearance cues. Text description $\mathcal{L}$ is taken from MGD[1].

Table 1: Quantitative results of our approach compared with other methods on VITON-HD Dataset on different modalities.

	Modalities						
Model	**Text**	**Colour**	**Sketch**	**Texture**	**FID ↓**	**LPIPS ↓**	**CLIP-S ↑**
PBE[32]	×	×	✓	×	17.1	0.212	0.42
BLIP-Diffusion[14]	✓	×	✓	×	13.0	0.206	0.49
ControlNet[34]	✓	×	✓	×	14.3	0.198	0.47
IP-Adapter[33]	✓	×	✓	✓	12.8	0.186	0.51
Ours	✓	✓	✓	✓	11.9	0.175	0.55

We train our model on garments and parsed data from VITON-HD at a spatial resolution of 512×384. Training is performed on a single NVIDIA A100 GPU with 80GB memory using the AdamW optimizer. The learning rate is set to 1×10^{-4} with a linear schedule similar to the PBE model. The model is trained for 250K steps with a batch size of 16. Following DDPM [12], we use a noise prediction loss. All conditioning modules *SketchCtrl*, *ColorFiLM*, *LangBLIP*, and *TexAdapter* are jointly trained end-to-end while keeping the PBE UNet backbone frozen.

4.2 Quantitative results

We evaluate our method on the VITON-HD test set using FID, LPIPS, and CLIP-S metrics. As shown in Table 1, our model achieves the best performance across all metrics, with a FID of 11.9, LPIPS of 0.175, and CLIP-S of 0.55, indicating superior image realism, perceptual quality, and semantic alignment. Compared to BLIP-Diffusion [14] and IP-Adapter [33], which lack full multimodal conditioning, our approach benefits from the integration of text, color, sketch, and texture modalities. Notably, models that do not incorporate texture cues (e.g., PBE [32], ControlNet [34]) show higher FID and lower semantic consistency. These results validate the effectiveness of our joint conditioning framework in generating photo-realistic and semantically grounded garments.

4.3 Qualitative results

Figure 3 compares our method against Paint-by-Example (PBE) [32], Gemini [28], and ChatGPT-4o [13] across various garment inputs composed of sketch, texture crop, and textual caption. Our method consistently achieves superior visual fidelity and alignment with conditioning signals. Unlike PBE, which largely ignores texture and semantic inputs, and

Figure 3: Qualitative comparison of our proposed method on VITON-HD dataset.

Figure 4: Ablation study of various proposed modules on garment generation.

Gemini/ChatGPT-4o, which often hallucinate textures or fail to preserve structure, our model effectively integrates all modalities. For instance, in row 1, the jacquard weave and rich color are faithfully rendered with the correct mock-neck silhouette. Rows 2 and 3 highlight fine lace and sleeve details synthesized from both text and texture, which are often missing or oversimplified in baselines. In simpler cases, our model accurately reflects subtle material properties such as ribbing, pleats, or sheen while adhering to the sketch and preserving color intent, demonstrating robustness across garment types and complexity levels. Overall, GenWear provides more faithful and interpretable garment synthesis through its unified conditioning of sketch, texture, color, and text.

4.4 Ablation Study

To assess the contribution of each control modality in our framework, we perform an ablation study by incrementally removing individual conditioning components while keeping the rest of the model unchanged. We evaluate the resulting generations on four key metrics: Fréchet Inception Distance (FID), LPIPS perceptual similarity, Color-Structure Similarity (CSS), and CLIP Score (CLIP-S). These metrics collectively assess visual quality, perceptual realism,

Figure 5: Limitations of our proposed garment generation model

color-layout consistency, and semantic alignment.

Figure 4 illustrates the incremental effect of each modality-specific module in our architecture. Starting from the Sketch Only baseline, which captures the structure but fails to convey accurate color or texture, we observe progressive improvements as modalities are added. Integrating LangBLIP introduces semantic alignment with text, enabling shape and garment type refinement (e.g., sleeves and neckline in row 1). Adding ColorFiLM corrects color mismatches (row 1) and enhances hue consistency. Finally, the TexAdapter module contributes fine-grained texture realism and material properties, as seen in the pleated patterns (row 2) and ribbing fidelity (row 3). These results demonstrate the modular and complementary nature of our design, with each component contributing distinct visual attributes critical to high-fidelity synthesis.

Table 2: Ablation study on the VITON-HD test set.

Configuration	**FID** ↓	**LPIPS** ↓	**CSS** ↑	**CLIP-S** ↑
Full Model (Ours)	**11.9**	**0.175**	**0.77**	**0.55**
w/o Texture Adapter	14.8	0.196	0.72	0.54
w/o ColorFiLM	13.6	0.188	0.63	0.54
w/o Text (LangBLIP)	13.3	0.182	0.70	0.50
Sketch Only	17.1	0.212	0.61	0.42

Table 2 presents a quantitative ablation on the VITON-HD test set. Removing any module leads to measurable performance degradation, confirming their complementary roles. Excluding the Texture Adapter results in the most significant drop across FID and LPIPS, demonstrating its importance in generating realistic textures. Removing ColorFiLM leads to a notable reduction in CSS, indicating its critical role in preserving color-structure consistency. Omitting LangBLIP (Text) primarily affects CLIP-S, reflecting the impact of textual grounding on semantic alignment. The sketch-only baseline performs worst across all metrics, highlighting the necessity of multimodal guidance for high-quality generation.

4.5 Limitations and Future Work

While our method generates realistic garments from multimodal inputs, it has limitations. As shown in Figure 5, the model may over-prioritize textual cues, e.g., generating pink garments due to color words in the caption, despite neutral texture patches (top-left, bottom-left). In cases where the texture patch contains dominant elements like buttons (bottom-right), structural boundaries may appear blurry or distorted. Additionally, semantically ambiguous captions can impair fine-grained control (top-right). Future work could enhance modality alignment through better cross-attention mechanisms, and leverage 3D-aware or physically grounded models for improved structure fidelity.

5 Conclusion

We introduce a unified framework for garment synthesis conditioned on sketch, texture patch, color and textual description. Our approach leverages sketch control via a ControlNet-inspired module to preserve fine structural details. Central to our approach is the TexAdapter module, which enables fine-grained texture modulation, and ColorFiLM module, which aligns color semantics with sketch features. Additionally, a language-guided encoder (Lang-BLIP) effectively incorporates rich semantic priors from BLIP2 text embedding. Extensive experiments and ablation on VITON-HD dataset demonstrate that our approach achieves state-of-the-art performance across multiple perceptual and semantic metrics.

References

[1] Alberto Baldrati, Davide Morelli, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Multimodal garment designer: Human-centric latent diffusion models for fashion image editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 23393–23402, 2023.

[2] Alberto Baldrati, Davide Morelli, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Multimodal-conditioned latent diffusion models for fashion image editing, 2024. URL https://arxiv.org/abs/2403.14828.

[3] Marc Brockschmidt. Gnn-film: Graph neural networks with feature-wise linear modulation. In *International Conference on Machine Learning*, pages 1144–1152. PMLR, 2020.

[4] Caroline Chan, Frédo Durand, and Phillip Isola. Learning to generate line drawings that convey geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7915–7925, 2022.

[5] Shu-Yu Chen, Wanchao Su, Lin Gao, Shihong Xia, and Hongbo Fu. Deep generation of face images from sketches. *arXiv preprint arXiv:2006.01047*, 2020.

[6] Shu-Yu Chen, Feng-Lin Liu, Yu-Kun Lai, Paul L Rosin, Chunpeng Li, Hongbo Fu, and Lin Gao. Deepfaceediting: Deep face generation and editing with disentangled geometry and appearance control. *arXiv preprint arXiv:2105.08935*, 2021.

[7] Weifeng Chen, Tao Gu, Yuhao Xu, and Arlene Chen. Magic clothing: Controllable garment-driven image synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6939–6948, 2024.

[8] Wengling Chen and James Hays. Sketchygan: Towards diverse and realistic sketch to image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9416–9425, 2018.

[9] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14131–14140, 2021.

[10] Yuanzheng Ci, Xinzhu Ma, Zhihui Wang, Haojie Li, and Zhongxuan Luo. User-guided deep anime line art colorization with conditional adversarial networks. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1536–1544, 2018.

[11] Arnab Ghosh, Richard Zhang, Puneet K Dokania, Oliver Wang, Alexei A Efros, Philip HS Torr, and Eli Shechtman. Interactive sketch & fill: Multiclass sketch-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1171–1180, 2019.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[13] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[14] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 2023.

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[16] Yifei Li, Tao Du, Kui Wu, Jie Xu, and Wojciech Matusik. Diffcloth: Differentiable cloth simulation with dry frictional contact. *ACM Transactions on Graphics (TOG)*, 42(1):1–20, 2022.

[17] Zeyu Li, Cheng Deng, Erkun Yang, and Dacheng Tao. Staged sketch-to-image synthesis via semi-supervised generative adversarial networks. *IEEE Transactions on Multimedia*, 23:2694–2705, 2020.

[18] Ente Lin, Xujie Zhang, Fuwei Zhao, Yuxuan Luo, Xin Dong, Long Zeng, and Xiaodan Liang. Dreamfit: Garment-centric human generation via a lightweight anything-dressing encoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5218–5226, 2025.

[19] Shuqi Liu, Mingtian Tao, Yifei Huang, Changbo Wang, and Chenhui Li. Image-driven harmonious color palette generation for diverse information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[20] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *Advances in neural information processing systems*, 30, 2017.

[21] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5084–5093, 2020.

[22] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 4296–4304, 2024.

[23] Sanhita Pathak, Vinay Kaushik, and Brejesh Lall. Mac-vton: Multi-modal attention conditioning for virtual try-on with diffusion-based inpainting. In *2024 39th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE, 2024.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[25] Mehmet Saygin Seyfioglu, Karim Bouyarmane, Suren Kumar, Amir Tavanaei, and Ismail B Tutar. Dreampaint: Few-shot inpainting of e-commerce items for virtual try-on without 3d modeling. *arXiv preprint arXiv:2305.01257*, 2023.

[26] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3408–3416, 2018.

[27] Peng Su, Kun Wang, Xingyu Zeng, Shixiang Tang, Dapeng Chen, Di Qiu, and Xiaogang Wang. Adapting object detectors with conditional domain normalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 403–419. Springer, 2020.

[28] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[29] Xian Wu, Chen Wang, Hongbo Fu, Ariel Shamir, and Song-Hai Zhang. Deepportraitdrawing: Generating human body images from freehand sketches. *Computers & Graphics*, 116:73–81, 2023.

[30] Zhifeng Xie, Hao Li, Huiming Ding, Mengtian Li, Xinhan Di, and Ying Cao. Hierafashdiff: Hierarchical fashion design with multi-stage diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8762–8770, 2025.

[31] Yuhao Xu, Tao Gu, Weifeng Chen, and Arlene Chen. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8996–9004, 2025.

[32] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. *arXiv preprint arXiv:2211.13227*, 2022.

[33] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[34] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models.

[35] Shiyue Zhang, Zheng Chong, Xujie Zhang, Hanhui Li, Yuhao Cheng, Yiqiang Yan, and Xiaodan Liang. Garmentaligner: Text-to-garment generation via retrieval-augmented multi-level corrections. In *European Conference on Computer Vision*, pages 148–164. Springer, 2024.

[36] Xujie Zhang, Yu Sha, Michael C. Kampffmeyer, Zhenyu Xie, Zequn Jie, Chengwen Huang, Jianqing Peng, and Xiaodan Liang. Armani: Part-level garment-text alignment for unified cross-modal fashion design. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 4525–4535, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392037. doi: 10.1145/3503161.3548230. URL https://doi.org/10.1145/3503161.3548230.