CONTROLLABLE GARMENT GENERATION WITH MULTI-MODAL DIFFUSION GUIDANCE

Sanhita Pathak[1], Vinay Kaushik[2], Brejesh Lall[1]
IIT Delhi[1], IIIT Sonepat[2], India

IIT Delhi

IIIT Sonepat

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.

Contributions

- 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.

Implementation Details and Dataset

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, 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.

Architecture Diagram

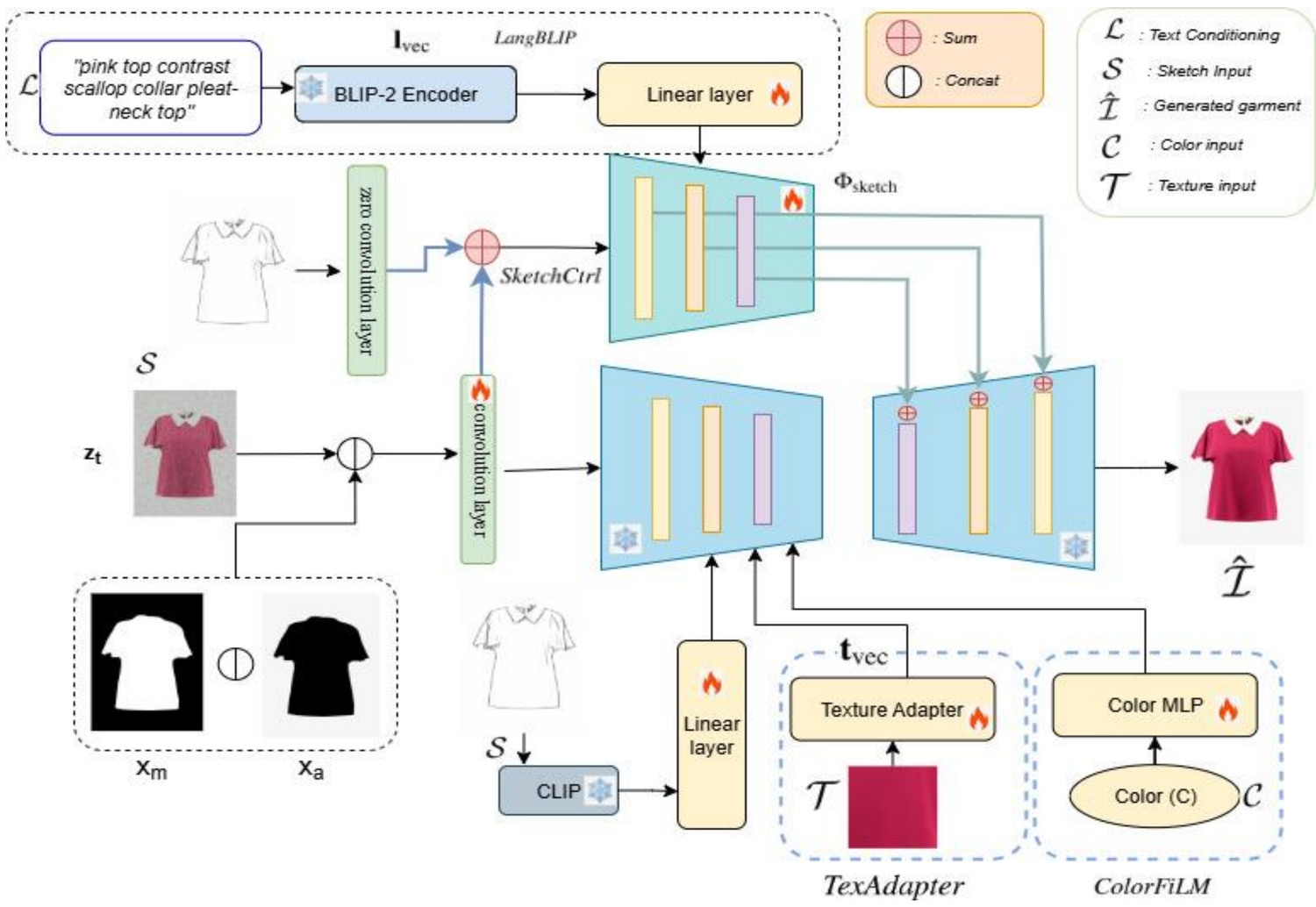

Figure 1: Architecture diagram of our proposed methodology shows the multimodal conditioning based garment generation leveraging multiple adapters with Diffusion UNet.

Ablation Results

Figure 2: Ablation study of various proposed modules on garment generation. 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.

Quantitative and Qualitative Results

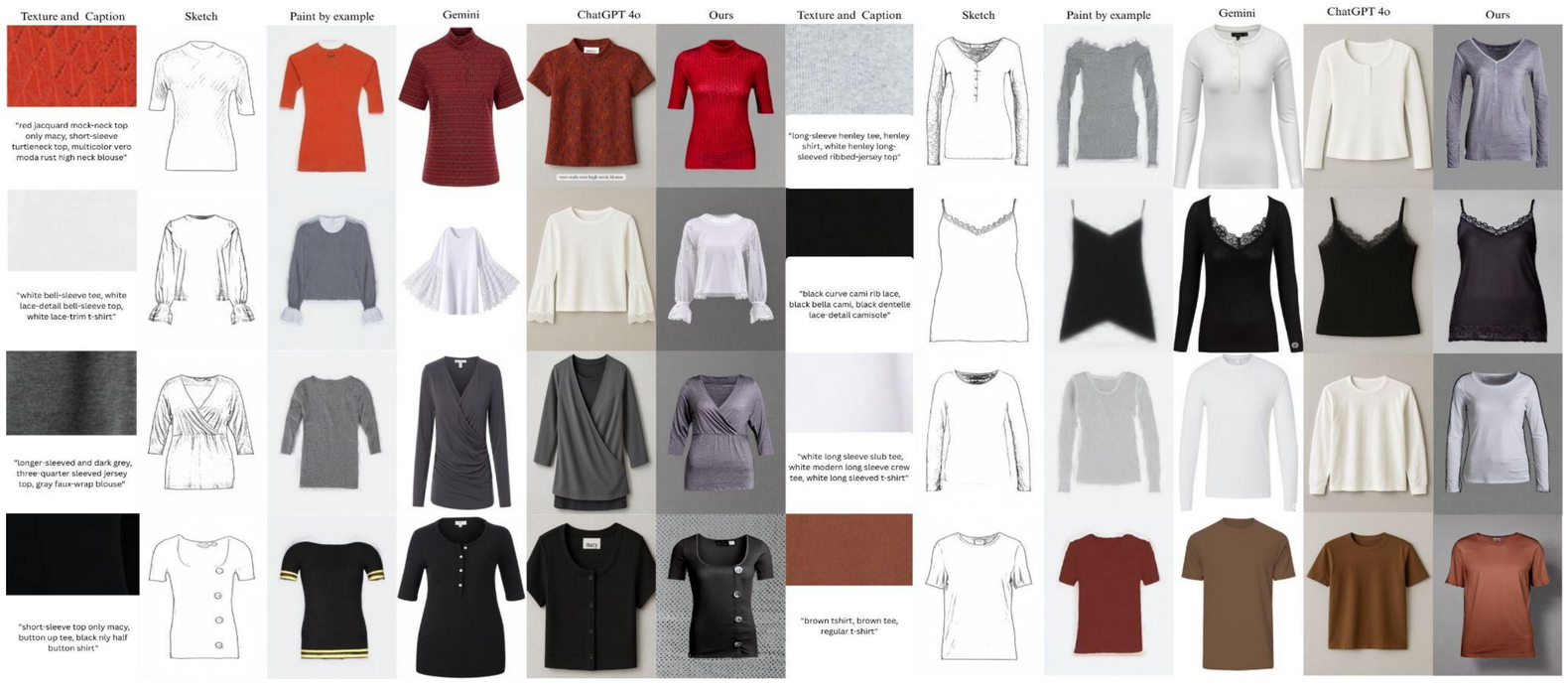

Figure 3: Qualitative comparison of our proposed method on VITON-HD dataset. The results obtained with Genwear show better results compared to other state-of-the-art techniques.

Model	Modalities: 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

Table 1: Quantitative results of our approach compared with other methods on VITON-HD Dataset on different modalities. The methods compared utilise some specific modalities but not all unlike our proposed method.

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 (LangBLIP) 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

- 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.
- 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.
- 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.
- 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.
- 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.
- 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.