CloDe: A Multimodal Tuning-Free Clothing Designer

Ajda Lampe[1], Ana Cvetković[1], Dong Seog Han[3], Satish Kumar Singh[4], Vitomir Štruc[2], Peter Peer[1,3]
[1] University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia
[2] University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia
[3] Kyungpook National University, Daegu, South Korea
[4] Indian Institute of Information Technology Allahabad, India
ajda.lampe@fri.uni-lj.si

***Abstract*—Recent diffusion-based image generation models have enabled new directions in virtual try-on and garment synthesis. However, existing approaches often trade off between controllability and accessibility: tuning-based methods typically require additional training or specialized inputs, while tuning-free methods may lack structural control and fine-grained appearance guidance. In this work, we propose CloDe, a multimodal, tuning-free garment synthesis framework built on a pretrained diffusion inpainting model. Our method enhances controllability without finetuning the backbone by introducing three components: (i) a category-aware masking strategy that improves edit locality and reduces garment leakage, (ii) pose conditioning via ControlNet to better preserve human structure, and (iii) optional fabric conditioning using IP-Adapter to incorporate fine-grained appearance cues. Built on DiCTI, our approach maintains a fully zero-shot pipeline while improving synthesis quality across realism, pose consistency, and texture fidelity. We evaluate CloDe on VITON-HD using quantitative metrics and a user study, showing consistent improvements over DiCTI in both objective measures and perceptual quality.**

***Keywords*—Generative AI; Diffusion models; Garment synthesis**

I. Introduction

Recent progress in generative image models, particularly diffusion-based approaches, has created new opportunities for fashion applications such as virtual try-on and interactive garment design. While much of the existing work focuses on transferring existing garments onto a person [1], [2], comparatively less attention has been given to rapid garment design exploration using generative models [3], [4]. Such systems could support early-stage fashion ideation, communication between consumers and designers, or image-based retrieval of similar garments. Beyond fashion design, they may also enable systematic generation of synthetic, richly labeled training data for downstream tasks such as virtual try-on and garment attribute recognition.

Early text-guided garment synthesis methods relied primarily on Generative Adversarial Networks (GANs) [5], often through latent-space manipulation [3], [6], [7]. Although these approaches demonstrated that natural language can guide garment editing, latent manipulation frequently produces out-of-distribution representations, leading to limited controllability and visual artifacts. More recently, diffusion models [8], [9] have shifted garment synthesis toward a more stable and flexible paradigm with improved realism and higher-quality generation. Some approaches extend pretrained diffusion models through additional trainable modules and conditioning inputs [4], [10], while others favor tuning-free pipelines that leverage the broad generalization capabilities of large pretrained diffusion backbones [11].

Despite recent progress, important limitations remain. Tuning-based methods often require specialized inputs such as garment sketches and additional model training, limiting their accessibility, while tuning-free approaches offer greater flexibility, but may suffer from weaker controllability, inaccurate editing regions, and pose drift.

In this work, we propose CloDe, a multimodal tuning-free garment synthesis method building on DiCTI [11]. We address three key failure modes of tuning-free garment synthesis: garment leakage, pose drift, and weak appearance control through (i) improved masking, (ii) ControlNet-based pose conditioning, and (iii) optional IP-Adapter fabric conditioning. We evaluate our method against DiCTI and show improvements in both quantitative metrics and visual quality, supported by a user study.

II. Related work

Early garment synthesis methods primarily relied on GANs [5]. FashionGAN [6], FICE [3], and FashionTex [7] explored text-guided garment editing using conditional synthesis, latent optimization, and texture conditioning. However, GAN-based methods often suffer from entangled latent representations, limiting edit locality and causing unintended changes outside the target garment region.

More recently, diffusion models [8] have become the dominant paradigm due to improved realism and controllability. DiCTI [11] introduced tuning-free garment synthesis using Stable Diffusion inpainting [9]. Recent multimodal approaches such as MGD [4] and TI-MGD [10] further improve controllability using pose, sketch, and texture conditioning.

Motivated by these developments, our work builds on DiCTI [11] and focuses on multimodal tuning-free garment synthesis. In contrast to methods requiring task-specific training or architectural changes [4], we integrate pretrained ControlNet [12] for pose guidance and an image adapter for

979-8-3195-4357-8/26/$31.00

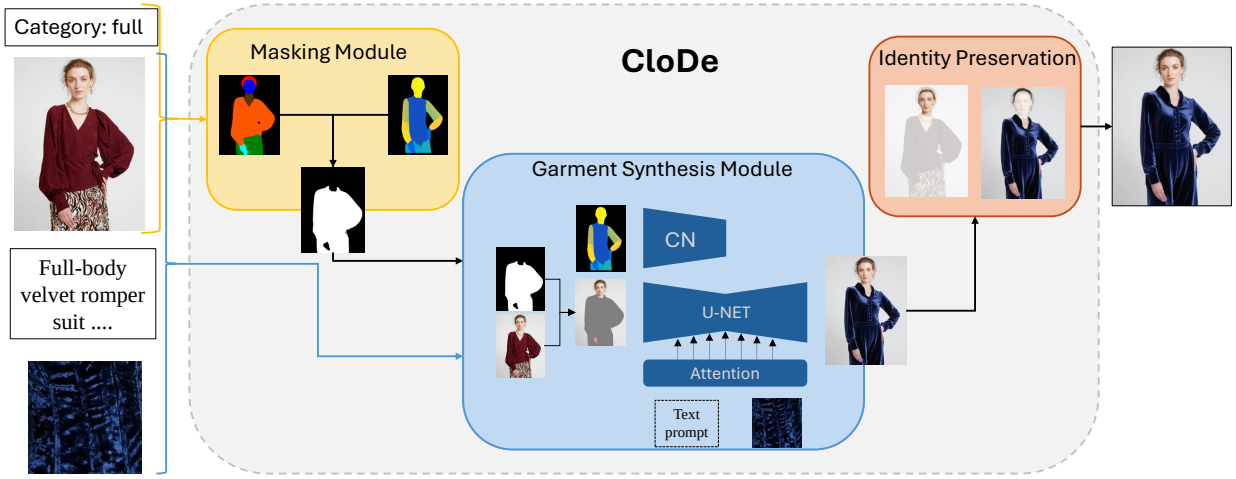

Figure 1. A high-level overview of the proposed CloDe. An input image and garment editing category determine a mask that defines editing and preservation regions for garment synthesis, which is further guided by textual and image prompts describing the target garment and a pose representation. Finally, facial features are restored to preserve identity.

texture conditioning, while keeping the diffusion backbone unchanged. We further introduce an improved masking strategy that enables selective garment-category editing with improved locality and reduced leakage.

III. Methodology

In this section, we propose an improved tuning-free garment synthesis method CloDe (**Clo**thing **De**signer). Given an input *person image* I representing a target person, a garment class label y_c, a text description y_t, and optionally a texture patch I_p representing the desired fabric and pattern, the goal of our method is to generate an image of the person in I wearing a novel garment of class y_c that corresponds to the text conditioning, while respecting the identity and pose of the person. We group different garment categories into three main classes: *top*, *bottom* and *full*, representing upper, lower, and full-body garments, respectively.

A. CloDe Overview

Fig. 1 shows a high-level overview of our proposed method. Inputs consist of an image depicting a person and a textual description of the target garment, along with the class label of the garment to be replaced – either *top*, *bottom* or *full*. Optionally, a fabric reference image may be supplied to improve guidance and constrain synthesis.

We approach garment synthesis as an inpainting problem – masking out a part of the image and replacing it based on control signals. Thus, an inpainting mask is first determined in a Mask Generation Module (Section III-B). The mask is then supplied to the Garment Synthesis Module (Section III-C), together with the textual description and the optional fabric reference image. The resulting image is then passed through an Identity Preservation Module to restore facial features, following ideas from prior work [11].

B. Determining the editing area

Inpainting models require a binary mask that defines the editing area while preserving regions that should remain unchanged (i.e., face, hands, and non-target garments). Ideally, the mask should remove all parts of the target garment while excluding preservation areas and avoiding enforcing a strict shape derived from the original garment.

We combine two types of masks: (1) a *clothing-agnostic mask* and (2) a *clothing-aware mask*. The clothing-agnostic mask is determined by selecting appropriate body regions from a DensePose [13] body part segmentation. The clothing-aware mask is constructed from garment parsing categories. The 18 classes produced by an off-the-shelf SegFormer [14]-based clothing segmentation model are mapped into three general categories: *top*, *bottom* or *full*, and a binary mask is generated to match the target category. The first removes coarse garment regions, while the second removes remaining garment-specific details such as neckline or loose sleeves that may otherwise remain visible during synthesis. Both masks are dilated using a circle-shaped structuring element of size 70 to better accommodate loose-fitting garments. We combine both masks by taking their union. Finally, the preservation areas containing identity features (i.e. face, hands and feet) and non-target garments are removed from the mask. This formulation enables editing of a single garment category rather than requiring full-outfit replacement, improving controllability and enabling more flexible combinations across garment types.

C. Garment Synthesis

Once the mask is prepared, it is used to mask out the editing region and construct the input to a pre-trained latent diffusion inpainting model. Since the mask covers a large portion of the person's body, especially in full-body edits, it can reduce structural cues and lead to pose drift. We address this by augmenting the diffusion pipeline with a ControlNet [12] conditioned on pose information derived from DensePose [13]

body segmentation. This provides explicit structural guidance and helps preserve body configuration during synthesis. Furthermore, we enable optional image-prompt conditioning through an IP-adapter to complement the textual prompt that may inaccurately describe desired texture and patterns.

D. Implementation details

Our synthesis pipeline is implemented using Stable Diffusion v1.5 inpainting model with pretrained ControlNet (controlnet-densepose-1.5) and IP-Adapter (ip-adapter-plus sd15) modules. ControlNet conditioning strength is set to 0.8, while IP-Adapter with a strength of 0.75 is enabled only when a fabric reference image is provided. We use segformer_b2_clothes checkpoint for clothing-aware segmentation. Text prompts consist of garment descriptions written in natural language, followed by generic quality-enhancing commands such as "photograph", "detailed hands", and "realistic lighting". Negative prompts are provided as comma-separated commands to discourage blurry, unrealistic, or anatomically implausible outputs. Fabric reference images are extracted by center cropping garment regions from existing worn garment images and resizing them to 224×224 pixels.

IV. Experiments and Results

We evaluate CloDe through quantitative metrics, qualitative comparisons, and a user study. We first describe the dataset and evaluation setup, followed by the metrics used, and finally present numerical and visual results.

A. Dataset

We evaluate CloDe on VITON-HD [1], a standard virtual try-on dataset containing 13,679 high-resolution images of human subjects in near-front-facing poses. We select a subset of 150 images with variation in body shape, pose, and clothing style. Each image is paired with five manually crafted textual prompts for each garment category (top, bottom, full), resulting in 2,250 image–prompt pairs.

B. Performance metrics

We evaluate CloDe using four metrics measuring realism, text alignment, pose preservation, and texture fidelity. **KID** measures distribution similarity between generated and real images (lower is better). **CLIP Score (CLIP-S)** measures text-image alignment in CLIP embedding space (higher is better).

Pose Distance (PD) [10] measures pose preservation as the weighted Euclidean distance between pose keypoints, with higher weights for torso joints (lower is better). **Texture Score (TS)** [10] measures similarity between the reference texture and the generated garment region using CLIP embeddings (higher is better).

C. Results

Since DiCTI [11] is the closest tuning-free baseline and exhibits several limitations targeted by our method, we focus our evaluation on comparison against DiCTI. We start with visual analysis, followed by quantitative comparison and a user study.

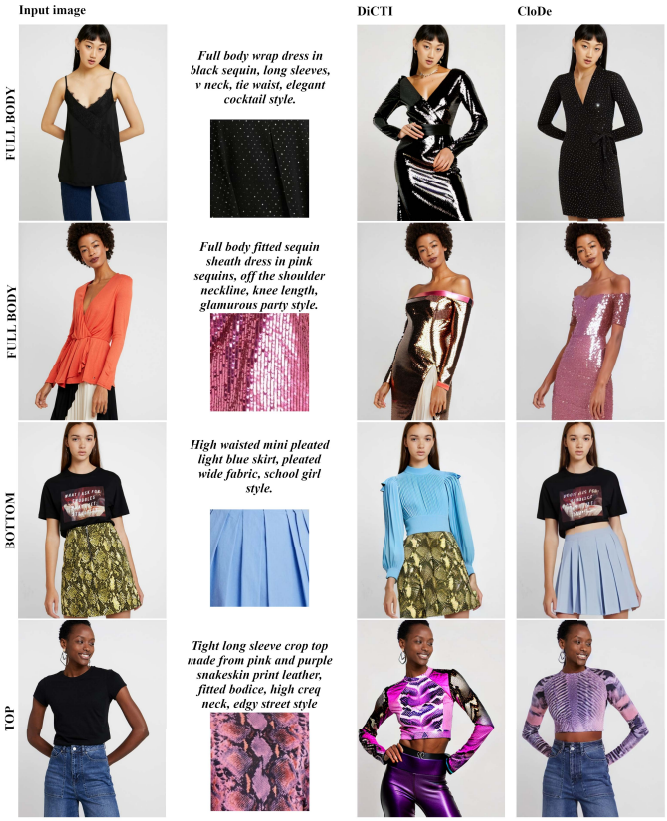

Figure 2. Comparison of DiCTI and CloDe across different garment types. The CloDe uses both text prompts and fabric reference images, while DiCTI uses text-only conditioning.

1) Visual analysis: Fig. 2 shows qualitative comparisons between DiCTI and CloDe. DiCTI relies solely on text and lacks explicit structural and appearance guidance, which can lead to pose inconsistencies and less faithful fabric rendering. In contrast, CloDe improves pose preservation through ControlNet conditioning and better garment locality through category-aware masking. The optional fabric reference further improves texture fidelity and reduces color and pattern ambiguity compared to text-only guidance.

2) Quantitative comparison: Table I shows quantitative results across all garment categories. CloDe consistently improves KID, Pose Distance (PD), and Texture Score (TS), indicating higher realism, better pose preservation, and improved texture fidelity. The largest gains are observed in PD and KID, where improved masking and pose conditioning significantly reduce structural artifacts. TS also improves, particularly for full-body garments, where multimodal conditioning has the strongest effect. CLIP Score decreases slightly for top and bottom categories, likely due to reduced modification of non-target regions. This suggests that CLIP-S may reward broader image changes that improve global prompt alignment, even when those modifications reduce garment locality or realism.

3) User study: We conducted a user study with 30 participants to evaluate perceptual quality, as standard metrics do not fully capture human judgment in garment synthesis.

TABLE I. PER-CATEGORY COMPARISON (750 IMAGES PER CATEGORY) OF DICTI AND CLODE. THE ARROWS NEXT TO THE METRIC LABELS INDICATE THE DIRECTION OF IMPROVEMENT.

Metric	Top		Bottom		Full		Overall	
	DiCTI	CloDe	DiCTI	CloDe	DiCTI	CloDe	DiCTI	CloDe
KID ↓	0.040	**0.011**	0.037	**0.015**	0.073	**0.069**	0.037	**0.016**
CLIP-S ↑	**29.27**	26.89	**25.44**	22.11	28.81	**30.43**	**27.84**	26.48
PD ↓	76.70	**29.67**	70.96	**27.46**	77.27	**48.19**	74.97	**35.11**
TS ↑	0.571	**0.591**	0.365	**0.368**	0.573	**0.653**	0.593	**0.627**

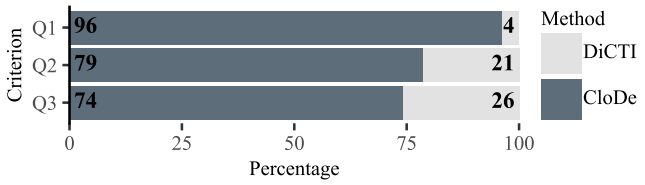

Figure 3. The ratio between votes for CloDe and DiCTI [11] in the user study for each of the criteria: Q1 - pose preservation, Q2 - fabric accuracy and Q3 - garment structure.

Participants were shown the input image together with outputs from DiCTI and CloDe and asked to select the preferred output for a given, disregarding other factors. To reduce presentation bias, outputs were shown in randomized left-right order, and method names were hidden from participants. We evaluate: pose preservation (Q1), fabric accuracy (Q2), and garment structure (Q3). The evaluated image pairs were sampled across all garment categories to ensure diverse examples. Each participant evaluated 12 pairs per criterion, resulting in 360 votes per criterion.

Fig. 3 shows the ratio of votes for each method. CloDe is preferred across all criteria, with the largest margin in pose preservation (Q1), consistent with quantitative PD results. The strongest preference in Q1 suggests that structural consistency is the most immediately noticeable improvement to human observers. Strong preferences are also observed for fabric accuracy (Q2) and garment structure (Q3), suggesting that multimodal conditioning improves both texture fidelity and perceived garment realism. Despite lower CLIP-S scores in some categories, participants consistently preferred CloDe in terms of fabric fidelity and garment structure. This suggests that improved edit locality and structural consistency may better align with human perception than global text-image similarity metrics alone.

V. CONCLUSION

In this work, we presented CloDe, a multimodal tuning-free garment synthesis method based on diffusion inpainting. By combining category-aware masking, ControlNet pose guidance, and optional IP-Adapter conditioning, our approach improves realism, pose preservation, and texture fidelity over DiCTI while preserving tuning-free flexibility. A current limitation is texture conditioning for highly repetitive patterns such as checkered fabrics. Future work will explore stronger garment-specific visual conditioning mechanisms for improved fine-grained texture synthesis.

ACKNOWLEDGMENT

Supported by ARIS Research Programmes P2-0250 Metrology and Biometric Systems and P2-0214 Computer Vision.

REFERENCES

[1] S. Choi, S. Park, M. Lee, and J. Choo, "VITON-HD: High-resolution virtual try-on via misalignment-aware normalization," in *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2021.

[2] B. Fele, A. Lampe, P. Peer, and V. Struc, "C-VTON: Context-driven image-based virtual try-on network," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2022.

[3] M. Pernuš, C. Fookes, V. Štruc, and S. Dobrišek, "FICE: Text-conditioned fashion-image editing with guided gan inversion," *Pattern Recogn.*, vol. 158, Feb. 2025.

[4] A. Baldrati, D. Morelli, G. Cartella, M. Cornia, M. Bertini, and R. Cucchiara, "Multimodal garment designer: Human-centric latent diffusion models for fashion image editing," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23336–23345, 2023.

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, (Cambridge, MA, USA), p. 2672–2680, MIT Press, 2014.

[6] S. Zhu, S. Fidler, R. Urtasun, D. Lin, and C. C. Loy, "Be your own prada: Fashion synthesis with structural coherence," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1689–1697, 2017.

[7] A. Lin, N. Zhao, S. Ning, Y. Qiu, B. Wang, and X. Han, "Fashiontex: Controllable virtual try-on with text and texture," in *Proceedings of the ACM Special Interest Group on Computer Graphics and Interactive Techniques Conference (SIGGRAPH)*, 2023.

[8] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6840–6851, 2020.

[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

[10] A. Baldrati, D. Morelli, M. Cornia, M. Bertini, and R. Cucchiara, "Multimodal-conditioned latent diffusion models for fashion image editing," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

[11] A. Lampe, J. Stopar, D. K. Jain, S. Omachi, P. Peer, and V. Štruc, "DiCTI: Diffusion-based clothing designer via text-guided input," in *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–9, 2024.

[12] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *IEEE International Conference on Computer Vision (ICCV)*, 2023.

[13] R. A. Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7297–7306, 2018.

[14] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12077–12090, 2021.