Clothe and Pose

Nakul Sharma Aayush Bansal Minh Vo

Figure 1. **(a)** Our approach inputs a single image of the user and casual captures of garments (both front and back views), and **(b)** outputs desired poses of the same user wearing the given garments.

Abstract

We introduce Clothe and Pose, an image generation and editing task that enables users to try on garments while simultaneously adopting any desired pose. Our method takes a single user image, a set of garment images, and a reference pose as input, and outputs the user wearing the target garment in the specified pose. We also propose an evaluation framework for clothing and re-posing tasks. Our study encompasses a wide variety of garments—including athletic wear, bottoms, dresses, innerwear, and swimwear—across diverse poses. Finally, we demonstrate the utility of Clothe and Pose for human-centric editing on both real-world captures and synthetic imagery.

1. Introduction

Consider a user, Bob, who wishes to try on a new outfit that he found online. Is a single, static perspective sufficient for him to evaluate the garment? In a physical retail environment, a customer naturally moves, turns, and adopts various poses to observe how a garment drapes and moves with the body. To bridge this gap in digital fashion, we introduce *Clothe and Pose*. Our proposed framework for the task takes a single image of the user, a target pose and garment images as input, generating high-fidelity outputs of the user in multiple desired poses, as shown in Figure 1.

The existing *virtual try-on technology* [5, 18, 37] limits the user to one perspective. One way to overcome this challenge is to use the state-of-the-art *reposing* modules [34, 66] that can generate arbitrary poses of the user from a given single image. A modular approach (CatVTON [7] + Leffa Repose [66]) leads to error propagation, as shown in Figure 2. We also observe that reposing modules struggle to preserve the garment and human identity. They hallucinate garment details that are occluded or not visible in the user image provided. We posit that clothing and posing are interdependent – the way one poses depends on how they clothe. We introduce a simple baseline for Clothe and Pose task.

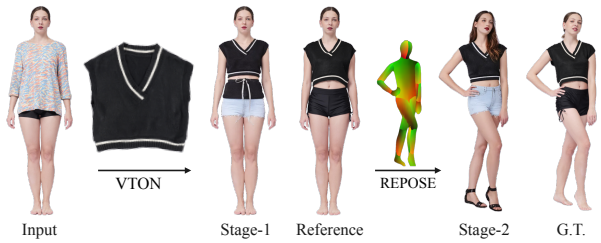

Figure 2. **Error Propagation in Modular Clothe and Pose**. We demonstrate cumulative error propagation in a two-stage pipeline using a sleeveless top as the target garment. In Stage-1, CatVTON [7] performs virtual try-on (VTON) but hallucinates the bottom garment, inconsistent with the reference image. In Stage-2, Leffa [66], a SOTA pose transfer model transforms the Stage-1 output into the target pose; however, it fails to preserve the person's identity, as evidenced by the visual discrepancy between the Stage-2 output and the ground truth (G.T.).

(a) Original garment influence the final output (b) Existing Virtual try-on evaluation (c) Our evaluation setup

Figure 3. (a) Alice wants to try out a new t-shirt. Ideally, it should look the same irrespective of any garment she is currently wearing. State-of-the-art CatVTON [7] reveals that the original garment influences the output. This behavior is consistent in most VTON methods [6, 55, 66]. We highlight two popular scenarios — the left is the current academic evaluation setup compared against the other two practical scenarios; (b) Current VTON benchmarks [5, 37] contain tuples of (G_A, P_AG_A) and the success of a VTON method is evaluated by masking the region around G_A in P_AG_A and then reconstructing P_AG_A using G_A. This setup doesn't allow us to quantify the performance in practical scenarios; (c) We introduce an evaluation setup that allows us to contrast generated output with a ground truth.

In this work, we also observe fundamental limitations that have restricted the practical and safe deployment of the technology for try-on. One limitation is the influence of the original garment worn by a user, as shown in Figure 3(a). This arises from the current evaluation [5, 37], which contains tuples of (G_A, P_AG_A), where G_A is an image capture of the garment and P_AG_A is an image of a person wearing G_A. This setup does not allow for proper unpaired evaluation due to missing ground truth. Training and evaluation, both are performed by masking pixels as shown in Figure 3(b) and optimizing on these benchmarks leads to input dependency bias — enforced during training and optimization for benchmark performance. Ideally, we should have an image of P_A wearing another garment G_B but it is highly unlikely that one will obtain a perfect pixel alignment in P_AG_A and P_AG_B. Our setup (of simultaneous clothing and posing) overcomes this limitation. We introduce an evaluation setup that contains triplets of (P_AG_A, G_B, P_AG_B) as shown in Figure 3(c). Our evaluation setup allows us to contrast the quality of virtual try-on with a physical try-on and quantify the influence of target poses.

The existing benchmarks [5, 37] focus majorly on tops and dresses, and only consider the front view of these garments. In our evaluation, we study a wide range of clothing including tops, bottoms, dresses, innerwear, athletic wear, and swimwear, as shown in Figure 5 and Figure 6. Our goal is to minimize hallucination and thus, we provide front and back view of the top and bottom garments to maximize information. The human subjects in our evaluation set are captured in a controlled setting with minimal accessories and tied hair, with no background distractions. This allows us to carefully focus on the preservation of human and garment details under varying clothing and pose conditions. Inadvertently, our evaluation data allows us to study the problem of *Reposing*. We observe that explicit garment information helps to better repose, as shown in Figure 7, verifying our hypothesis that clothing and pose are indeed interdependent. Finally, it is non-trivial to generate and capture images with desired clothing and pose. Our approach allows us to edit garments and pose of an individual in both real and generated images.

Contributions: (1) We introduce the Clothe and Pose task, to try-on garments and pose, (2) we introduce a new comprehensive evaluation setup to study the impact of different factors on clothing and posing tasks; and (3) we demonstrate the usefulness of our framework in human-centric editing tasks.

2. A brief history of virtual try-on and reposing

Virtual Try-on. After the initial proposal of image-based virtual try-on task [5, 23] and exploration of related tasks [19, 42, 45], several datasets have been proposed to train and evaluate VTON systems [5, 9, 37]. Initial approaches adopted a two-stage framework [5, 11, 15, 17, 18, 20, 22, 27–29, 32, 37, 38, 51, 53, 56–58], wherein the first stage network warps the garment and the second stage network synthesizes the try-on — earlier methods relied on GANs [16], while later and current methods utilize LDMs [43] as generative backbones. More recent methods produce try-ons in a single stage, often exploiting large pre-trained latent diffusion backbones and their attention mechanism [2, 6, 6, 7, 24, 30, 55, 59, 62, 66–68]. Liu

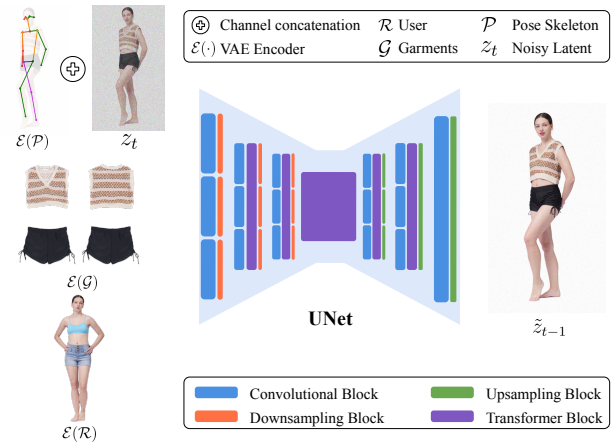

Figure 4. **Input:** a user image, a spatial concatenation of front and back views of the top and bottom garments, and a noisy latent image channel-wise concatenated with the latents of the skeleton of the target pose. **Output:** a denoised image latent. The three inputs are treated as three separate streams in the model without text conditioning, as described in the text.

et al. [33] use SMPL-based conditioning to simulate try-ons in different poses. This method is specifically helpful for frontal poses, but limits generating try-ons in arbitrary poses, whereas our method allows synthesis in arbitrary poses as demonstrated in Figure 5. Current Virtual Try-On limits the pose of the user: the way a user poses in an image is a function of garment, our goal is to provide the flexibility to change the pose, as well as a way to edit clothing and pose in previously captured images.

Pose Transfer or **Reposing** refers to synthesizing humans in different poses from a single reference image. Representative methods [3, 13, 34–36, 47, 48] propose solving pose transfer in isolation, while [1, 8, 46, 66] aim to develop frameworks that can be trained for both pose transfer and virtual try-on, separately. These methods do not support conditioning on garments, which limits their performance for non-frontal poses as shown in Figure 7. Our method is jointly trained for clothing and posing simultaneously, and supports garment conditions, allowing it to perform pose transfer to arbitrary poses.

Several recent methods [12, 61, 64, 65] propose pipelines for trying on garments in arbitrary poses. They lack a unified approach and a proper evaluation setup — the two critical aspects we focus on in this work.

Editing Images. InstructPix2Pix [4] demonstrates the utilization of synthetic data and large pre-trained latent diffusion backbones for image editing using text. Proprietary models [26, 39] allow editing images using text and reference images, making the editing process more controllable. Qwen-Image-Edit [52] allows editing images using a mix of text, multiple reference images with support for conditioning on depth map, pose keypoints, among others. Our method enables better editing as shown in Figures 6 and 11.

3. Method

Given a user image, the target clothing images of top and bottoms, our model learns to generate the try-on in a single target pose. The remainder of this section is organized as follows: in Section 3.1, we outline our task and notations, in Section 3.2 we describe our proposed conditional model for the multi-pose try-on, and finally in Section 3.3 we describe the training details.

3.1. Overview

Formally, our single training sample for clothe and pose consists of a reference user image $\mathcal{R}$, a target image $\mathcal{T}$ of the same user and its corresponding pose $\mathcal{P}$, front-view and back-view of the upper garment worn in the target image $\mathcal{U}_f$ and $\mathcal{U}_b$, respectively and front-view and back-view of the lower garment worn in the target frame $\mathcal{B}_f$ and $\mathcal{B}_f$, respectively. Pose $\mathcal{P}$ is represented as an image by creating a skeleton image of keypoints obtained using ViTPose [54]. We spatially concatenate all the top and bottom garment conditions and refer to them as $\mathcal{G}$ in the rest of the paper.

We adopt the latent diffusion training paradigm [43] for modeling the problem. Specifically, we build upon the SDXL [40] to propose a new architecture with improved conditioning for the generation of the target image conditioned on user image, garment images, and the target pose. It is worth noting here that we use the SDXL autoencoder to obtain a latent representation of each of our conditions and these latent representations are used for training the model.

3.2. Model

Virtual try-on methods have popularized self-attention in the denoising UNet [44] for garment-conditioned generation, converging on two architectural choices: (i) spatially concatenating garment latents with noisy target-image latents along the width [7, 60], and (ii) running a separate branch to extract garment features, then spatially concatenating them to the generation latents before the self-attention module [6, 50, 55, 66]. While these designs have substantially improved generation quality for virtual try-on, both come with their own caveats to be extended to multi-condition settings. Maintaining a full, separate, time-dependent branch per condition multiplies inference and training cost, as each branch independently performs cross-attention and self-attention operations. Spatial concatenation within a single network avoids this cost, but forces the model to simultaneously solve two unrelated tasks—generating the try-on image and reproducing the garment (or any conditioning image) in the output—which is suboptimal, since a significant fraction of network capacity goes toward copying the condition to the output rather than toward generation quality.

To address these limitations, we propose a multi-stream architecture inspired by recent advances in text-to-image

Figure 5. **Clothe and Pose**. We study the performance of different methods for joint clothing and posing We observe that our method performs better across a wide range of clothing and poses, including extreme pose changes (row 3) from a frontal reference image of the user to the target back pose, effectively utilizing both the front and back views of the garment, as shown in the second column.

generative modeling [14, 25]. Figure 4 provides an overview of our architecture and training flow. The remainder of this subsection details our key modifications to the standard SDXL UNet architecture.

Removing Text cross-attention. SDXL integrates text cross-attention blocks across the architecture that interfaces the UNet with CLIP [41] text embeddings, adding roughly 0.5 B parameters. Recent VTON systems typically keep this branch alive, either by injecting a generic template/empty prompt, by synthesizing garment descriptions with LLMs (this causes significant additional compute or monetary cost for unseen garments during inference), or by describing every image with a mixture of pose and appearance tokens. We aim to guide our model explicitly using target pose and target garment images and thus remove the textual control from the model.

Multi-stream blocks. We use channel-wise concatenation for all conditions that demonstrate pixel-alignment with the target — in our task this is only limited to the pose condition. For conditions that are not pixel-aligned with the output (in our case, garment conditions and the user reference image), we propose to process them in parallel to the noisy latent inside each network block with a separate set of learnable weights followed by the interaction of repre-

Figure 6. **Innerwear and Swimsuits**. We contrast our approach with the best performing baseline, Qwen-Image-Edit.

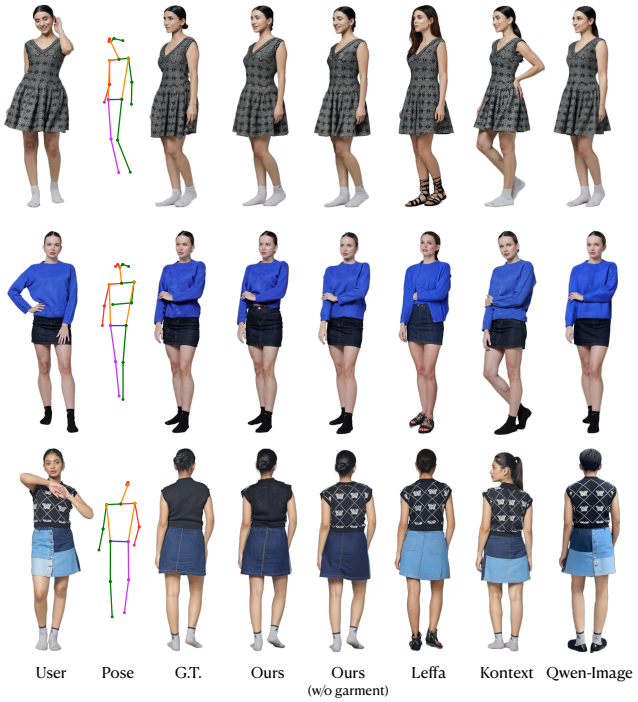

Figure 7. **Reposing**. We observe that the front and back views of garments improve the quality of results, and faithfulness to the original garment, specifically demonstrated in the last row.

sentations of all the conditions, and of the noisy latent in a joint self-attention module. This ensures each condition representation and the noisy latents, get their own neural weights, unlike the spatial concatenation design.

Joint Attention. The joint attention for the interaction of conditions and noisy latents is inspired by recent successes in text-to-image modeling [14, 25]. After the initial parallel processing of each condition, we add position embeddings to embed spatial positional information, and a learnable "stream embedding" vector to each of the conditions to embed the info of their origin in the self-attention mechanism. In our early experiments, these techniques help the model learn faster. Once position and stream embeddings are added, the representations for each condition are passed through QKV projection layers. Following [10, 14], we apply QK-normalization to stabilize our training. After normalization, the Q, K and V matrices of each stream are concatenated along sequence length. Finally, the self-attention operation [49] is applied to these matrices. This joint attention mechanism allows the freedom to each condition's representation to interact with other representations, and thus allows the model to learn their interaction internally. After the joint attention operation, the representations for each condition are separated and processed independently in the next neural block.

3.3. Training

Training this model requires a paired tuple $(\mathcal{R}, \mathcal{G}, \mathcal{T})$ where $\mathcal{R}$ is the reference user image, $\mathcal{G}$ represents the garment conditions, and $\mathcal{T}$ is the target image of the same user in different clothing and pose. However, such paired training data are not readily available abundantly.

We design a two-stage training strategy that leverages the complementary strengths of different data sources. In the first stage, we train on large-scale pose transfer data, augmented with garment images, to learn robust identity preservation across different poses. In the second stage, we fine-tune on a balanced mixture of our multi-pose virtual try-on data and pose transfer data, enabling the model to learn both garment transfer and identity preservation simultaneously. We provide the details of training data for both the stages in Appendix C.2.

Stage 1: Identity Preservation Pre-training. In the first stage, we leverage large-scale pose transfer data augmented with garment conditions where the same person appears in different poses but maintains consistent clothing. Given a reference user image $\mathcal{R}$ and target pose $\mathcal{P}$, the model learns to generate the person in the target pose while preserving their identity. Crucially, we also supply partial garment information during this stage — providing either the top garments ($\mathcal{U}_f$, $\mathcal{U}_b$) or bottom garments ($\mathcal{B}_f$, $\mathcal{B}_b$) but not both, owing to the nature of the available data.

To enforce robust garment conditioning, we randomly

Table 1. Quantitative evaluation on Clothe and Pose. Best results in **bold**.

Method	Front→Back			Front→Front			Front→Left			Front→Right		
	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑
IDM-VTON+Leffa	0.278	16.782	79.963	0.274	16.757	80.390	0.265	17.094	81.375	0.263	16.962	81.673
OOTDiffusion+Leffa	0.287	16.250	79.001	0.280	16.227	79.682	0.272	16.618	80.720	0.269	16.508	81.053
CatVTON+Leffa	0.277	16.995	80.118	0.272	16.903	80.607	0.262	17.233	81.596	0.264	17.072	81.682
Leffa+Leffa	0.278	16.703	79.454	0.276	16.619	79.842	0.264	16.984	81.027	0.264	16.802	81.233
CatVTON+Kontext	0.324	16.265	79.719	0.293	16.949	80.854	0.306	16.644	81.600	0.291	16.947	81.801
Leffa+Kontext	0.317	16.092	79.075	0.290	16.575	80.012	0.302	16.407	80.488	0.283	16.720	81.229
Qwen-Image-Edit	0.340	15.247	74.631	0.187	17.523	83.771	0.207	17.063	83.279	0.186	17.432	84.518
Ours	**0.166**	**18.599**	**84.380**	**0.155**	**18.785**	**85.296**	**0.153**	**18.984**	**85.999**	**0.151**	**19.028**	**86.191**

dropout patches of the garment in user images.This forces the model to rely on the explicit garment conditions $\mathcal{G}$ rather than copying garment appearance from the reference image.

Stage 2: Multi-Pose Virtual Try-On Fine-tuning. In the second stage, we fine-tune the pre-trained model on a mixture of our multi-pose virtual try-on data and pose transfer data. We sample the multi-pose try-on data with a probability of 0.6. For the multi-pose virtual try-on samples, the model has access to complete garment conditions — both top garments ($\mathcal{U}_f, \mathcal{U}_b$) and bottom garments ($\mathcal{B}_f, \mathcal{B}_b$) — enabling the model to learn full outfit transfer.

This mixing serves multiple purposes: (1) the virtual try-on samples teach the model to perform complete garment transfer with both top and bottom replacements, (2) the continued presence of pose transfer samples prevents catastrophic forgetting of identity preservation capabilities learned in Stage 1, and (3) the diversity in conditioning (partial garments for pose transfer, complete garments for try-on) improves the model's robustness — this enables our model to work for reposing as well.

During these stages, we replace each condition image with a gray pixel image as a null condition with a probability of 0.15 during training to utilize classifier-free guidance [21] during inference. As a consequence of this dropout and stage-1 training, our model can operate with any combination of input garments absent.

4. Experiments

Our evaluation data consists of ethnically diverse users wearing a total of 360 unique clothing items accompanied with their image captures in multiple poses for each garment associated with them. This data allows us to study the Clothe and Pose effectively. The distribution of the garments in our evaluation data is visualized in Figure 8.

We evaluate Clothe and Pose in four target pose configurations: Front→Front, Front→Left, Front→Right, and Front→Back, where the arrow indicates the transformation from source to target pose. Each of these configurations contain 600 evaluation pairs — totaling 2400 evaluation pairs for Clothe and Pose. The performance is measured

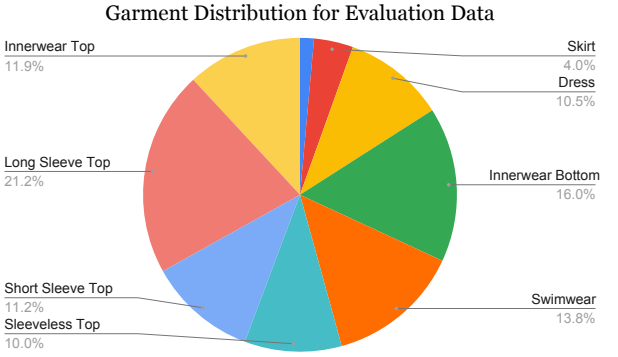

Figure 8. Our evaluation dataset covers a wide range of distribution of the garments for robust and comprehensive evaluation for clothing and posing, as well as reposing.

using LPIPS [63], PSNR and SSIM metrics because of the availability of ground truth data. For all our experiments, across tasks, the model remains the same.

Baselines for Clothe and Pose. Except for qwen-image-edit-2509 [52] (also known as Qwen–Image in figures, and trained on internet-scale data), no other methods support Clothe and Pose task. We construct baselines as pipelines in which a virtual try-on model performs the try-on, followed by its reposing by a pose transfer model. For virtual try-on, we consider the latest models, namely, IDM-VTON [6], OOTDiffusion [55], CatVTON [7] and Leffa [66]. For reposing of these try-ons, we use CFLD [34], Leffa [66] and text-guided Flux.1-Kontext-dev [26] image editing model trained using large-scale dataset.

Clothe and Pose Analysis. The main results for Clothe and Pose reported in Table 1 indicate that our method achieves substantial improvements over all the baselines. Our approach also outperforms Qwen-Image model which is significantly larger in parameter count than our model (20B vs 5B ours). Figures 5 and 6 highlight the performance of our model across various garments and posing scenarios. The performance gain is even more pronounced in the lateral pose transformations

Table 2. Quantitative comparison for Reposing. Best results in **bold**.

Method	Front→Back			Front→Front			Front→Left			Front→Right		
	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑
CFLD	0.258	15.472	81.304	0.243	15.662	82.052	0.249	15.742	82.712	0.236	15.694	83.044
Leffa	0.251	17.769	80.679	0.238	17.795	81.187	0.244	17.928	81.856	0.239	17.960	82.346
Flux.1-Kontext-dev	0.287	17.409	80.588	0.251	18.295	81.538	0.132	20.992	86.699	0.258	18.135	82.261
Qwen-Image	0.279	17.856	77.656	0.129	20.871	86.381	0.117	21.430	87.115	0.119	21.450	87.597
Ours (w/o garment)	0.138	19.972	85.054	0.112	21.207	86.526	0.113	21.183	86.982	0.112	21.348	87.324
Ours	**0.125**	**20.801**	**86.159**	**0.108**	**22.439**	**86.981**	**0.110**	**22.198**	**87.849**	**0.108**	**22.178**	**88.458**

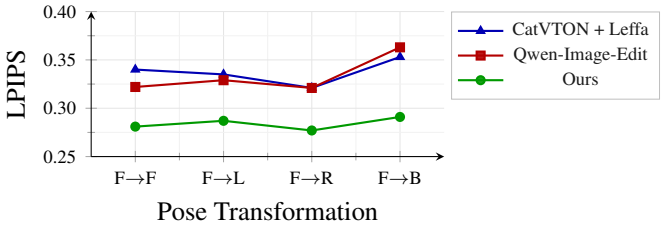

Figure 9. Garment region faithfulness with varying pose transformations in Clothe and Pose evaluation.

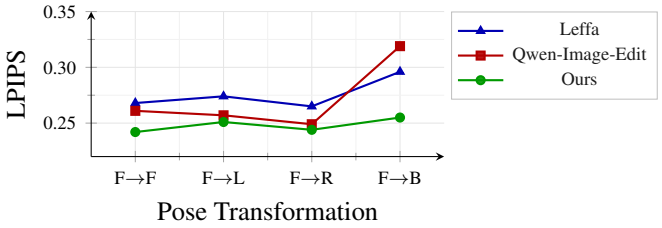

Figure 10. Garment region faithfulness with varying pose transformations in evaluation of Reposing.

(Front→Left and Front→Right) and the extreme pose transformations (Front→Back), because of the ability of our method to condition the try-on image on front and back views of the target garments. On the other hand, the table also suggests that pipeline-based methods, `IDM-VTON+Leffa`, `OOTDiffusion+Leffa`, `CatVTON+Leffa`, `Leffa`, `CatVTON+Kontext`, `Leffa+Kontext` struggle. This significant performance gap highlights the fundamental limitation of sequential processing: error accumulation from the virtual try-on stage propagates and amplifies during pose transformation — which is also evident from their visual results in Figure 5, where the pipeline-based approaches fail to preserve user identity and garment details accurately. In the Appendix D, weprovide more comparisons to include closed-sourced models in pipeline-based methods.

Setup and Baselines for Reposing. Similar to Clothe and Pose evaluation, we consider the same four pose configurations for Reposing evaluation, and utilize 600 pairs per configuration. We use CFLD [34] and Leffa [66] trained on DeepFashion Pose Transfer dataset, text-guided image editing model Flux.1 Kontext-dev, and Qwen-Image-Edit (or simply Qwen Image) that directly injects target pose in form of keypoints along with a reference image. We contrast these baselines with two configurations (but same checkpoint) of our method: one that utilizes the garment input, and the one without it.

Reposing. The evaluation of the Reposing task is reported in Table 2. Our method achieves superior performance across all metrics. The comparison between our full model and the ablated version where no garment is provided at the inference time, reveals interesting insights about the role of conditioning the generation on both front and back views of the garments — which are highly effective for better pose transfer performance. The accompanying visuals in Figure 7 suggest that methods trained and optimized on the DeepFashion Pose Transfer dataset and benchmark, i.e. Leffa and CFLD struggle with user identity preservation — this occurs primarily because the DeepFashion dataset has overlapping training and test set identities (more analysis in Appendix B). Thus, optimizing for DeepFashion Pose Transfer benchmark by excessively training on it makes these models overfitted in user identity aspect and hence they fail to generalize, making the benchmark unsuitable to measure real progress in pose transfer — this is also a major failure mode for pipeline-based methods for Clothe and Pose. Flux.1 Kontext is a text-guided image editing model and its lower performance as reported in the Table 2 combined with visual results Figure 7, and in conjunction with try-on methods in Figure 5 indicate that while the model is good at identity preservation, it struggles to edit pose. In addition to this standard evaluation, we also evaluate the similarity of the garment region of the generated image and the ground truth image, which is critical in assessing the performance of models for accurate garment transfer. We use a SCHP-based [31] parser to extract a mask over the garment regions in the ground truth. This mask is used to capture the garment regions of the generated image and the ground truth to measure their LPIPS score. We record this garment region faithfulness for Clothe and Pose in Figure 9 and in

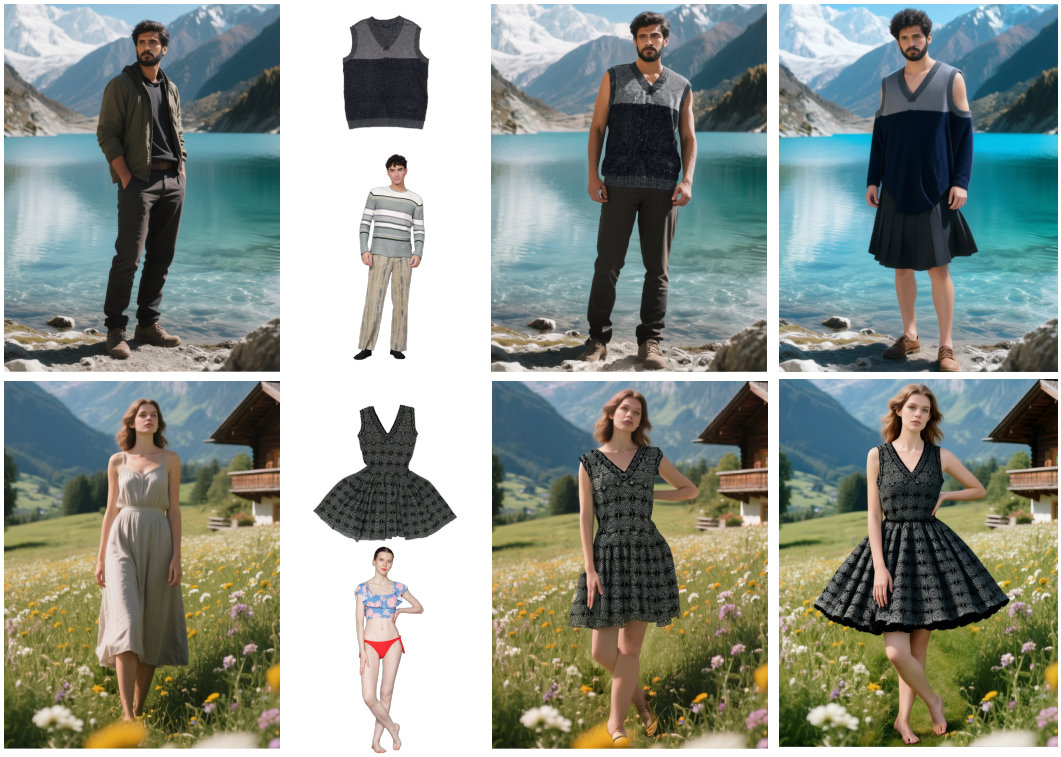

Figure 11. **Human-Centric Image Editing**: Our model is capable of changing clothes and pose in complex surrounding conditions, even for AI-generated images, outperforming Qwen-Image in garment fidelity and identity preservation.

Figure 10 for reposing, for the top-3 performing models. The analysis reveals that our method preserves the garment regions better than the baselines for both tasks.

Editing Images. An interesting direct application of the proposed system is editing previously captured photos to re-imagine the person in a new look, new clothing, but the same old nostalgic place. It could also be helpful in editing the outputs of image generation models to have better post-processing control over their pose or attire, since these things are not as intuitive to express in text prompts. In Figure 11, we qualitatively demonstrate that our method is capable of editing clothes and poses in AI generated images as well. The preservation of complex background, while trained on a limited amount of such data, is enabled by preserving the background info in the latents at the start of the generation process (Appendix C.1).

Our method demonstrates substantial improvements over pipeline-based approaches and Qwen-Image, we recognize this as an early contribution to unified garment-pose synthesis and discuss limitations of our method in Appendix E.

5. Discussion

Fashion is 50% *clothing, and* 50% *posing* – we introduce a computational framework of Clothe and Pose that allows a user to change their clothes and/or pose. Our method is surprisingly simple and outperforms existing methods.

Our work sets a strong foundation for composite clothing and pose transfer, and opens up new avenues to explore better and efficient training schemes. It also lays the groundwork for the exploration of extreme pose transfers in image editing, and provides a proper setup for evaluation.

References

[1] Badour AlBahar, Jingwan Lu, Jimei Yang, Zhixin Shu, Eli Shechtman, and Jia-Bin Huang. Pose with Style: Detail-preserving pose-guided image synthesis with conditional stylegan. *ACM TOG*, 2021. 3

[2] Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. Single stage virtual try-on via deformable attention flows. In *ECCV*, 2022. 2

[3] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *CVPR*, 2018. 3

[4] Tim Brooks, Aleksander Holynski, and Alexei A. Efros.

Learning to follow image editing instructions. In *CVPR*, 2023. 3

[5] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *CVPR*, 2021. 1, 2

[6] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *ECCV*, 2024. 2, 3, 6

[7] Zheng Chong, Xiao Dong, Haoxiang Li, shiyue Zhang, Wenqing Zhang, Hanqing Zhao, xujie zhang, Dongmei Jiang, and Xiaodan Liang. CatVTON: Concatenation is all you need for virtual try-on with diffusion models. In *ICLR*, 2025. 1, 2, 3, 6

[8] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In *ICCV*, 2021. 3

[9] Aiyu Cui, Jay Mahajan, Viraj Shah, Preeti Gomathinayagam, Chang Liu, and Svetlana Lazebnik. Street tryon: Learning in-the-wild virtual try-on from unpaired person images. In *CVPR*, 2024. 2

[10] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *ICML*, 2023. 5

[11] Xin Dong, Fuwei Zhao, Zhenyu Xie, Xijin Zhang, Daniel K Du, Min Zheng, Xiang Long, Xiaodan Liang, and Jianchao Yang. Dressing in the wild by watching dance videos. In *CVPR*, 2022. 2

[12] Chenghu Du, Peiliang Zhang, Junyin Wang, and Shengwu Xiong. Agff: Attention-gated feature fusion for multi-pose virtual try-on. *IEEE Transactions on Consumer Electronics*, 2025. 3

[13] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *CVPR*, 2018. 3

[14] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 4, 5

[15] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *CVPR*, 2021. 2

[16] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[17] Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In *ACM MM*, 2023. 2

[18] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *CVPR*, 2018. 1, 2

[19] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *ICCV*, 2019. 2

[20] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *CVPR*, 2022. 2

[21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[22] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzenes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *ECCV*, 2020. 2

[23] Nikolay Jetchev and Urs M. Bergmann. The conditional analogy gan: Swapping fashion articles on people images. In *ICCVW*, 2017. 2

[24] Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *CVPR*, 2024. 2

[25] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 4, 5

[26] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 3, 6

[27] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *ECCV*, 2022. 2

[28] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Tryongan: Body-aware try-on via layered interpolation. *ACM TOG*, 2021.

[29] Kedan Li, Jeffrey Zhang, and David Forsyth. Povnet: Image-based virtual try-on through accurate warping and residual. *IEEE TPAMI*, 2023. 2

[30] Kedan Li, Jeffrey Zhang, Shao-Yu Chang, and David Forsyth. Controlling virtual try-on pipeline through rendering policies. In *WACV*, 2024. 2

[31] Peike Li, Yunqiu Xu, Yunchao Wei, and Yang Yang. Self-correction for human parsing. *IEEE TPAMI*, 2020. 7

[32] Zhi Li, Pengfei Wei, Xiang Yin, Zejun Ma, and Alex C Kot. Virtual try-on with pose-garment keypoints guided inpainting. In *ICCV*, 2023. 2

[33] Jinxi Liu, Zijian He, Guangrun Wang, Guanbin Li, and Liang Lin. One model for all: Partial diffusion for unified try-on and try-off in any pose. *arXiv preprint arXiv:2508.04559*, 2025. 3

[34] Yanzuo Lu, Manlin Zhang, Andy J Ma, Xiaohua Xie, and Jian-Huang Lai. Coarse-to-fine latent diffusion for pose-guided person image synthesis. In *CVPR*, 2024. 1, 3, 6, 7

[35] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *NeurIPS*, 2017.

[36] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *CVPR*, 2020. 3

[37] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *CVPR*, 2022. 1, 2

[38] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *ACM MM*, 2023. 2

[39] OpenAI. gpt-image-1 api, 2025. 3

[40] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 3

[41] 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 *ICML*, 2021. 4

[42] Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu. Swapnet: Image based garment transfer. In *ECCV*, 2018. 2

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI 2015*, 2015. 3

[45] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural re-rendering of humans from a single image. In *ECCV*, 2020. 2

[46] Kripasindhu Sarkar, Vladislav Golyanik, Lingjie Liu, and Christian Theobalt. Style and pose control for image synthesis of humans from a single monocular view. *arXiv preprint arXiv:2102.11263*, 2021. 3

[47] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *CVPR*, 2018. 3

[48] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. Xinggan for person image generation. In *ECCV*, 2020. 3

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 5

[50] Siqi Wan, Jingwen Chen, Yingwei Pan, Ting Yao, and Tao Mei. Incorporating visual correspondence into diffusion model for virtual try-on. In *ICLR*, 2025. 3

[51] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, 2018. 2

[52] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. 3, 6

[53] Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *CVPR*, 2023. 2

[54] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *NeurIPS*, 2022. 3

[55] Yuhao Xu, Tao Gu, Weifeng Chen, and Arlene Chen. Ootd-iffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In *AAAI*, 2025. 2, 3, 6

[56] Keyu Yan, Tingwei Gao, Hui Zhang, and Chengjun Xie. Linking garment with person via semantically associated landmarks for virtual try-on. In *CVPR*, 2023. 2

[57] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *CVPR*, 2020.

[58] Han Yang, Xinrui Yu, and Ziwei Liu. Full-range virtual try-on with recurrent tri-level transform. In *CVPR*, 2022. 2

[59] Xu Yang, Changxing Ding, Zhibin Hong, Junhao Huang, Jin Tao, and Xiangmin Xu. Texture-preserving diffusion models for high-fidelity virtual try-on. In *CVPR*, 2024. 2

[60] Xu Yang, Changxing Ding, Zhibin Hong, Junhao Huang, Jin Tao, and Xiangmin Xu. Texture-preserving diffusion models for high-fidelity virtual try-on. In *CVPR*, 2024. 3

[61] Feng Yu, Ailing Hua, Chenghu Du, Minghua Jiang, Xiong Wei, Tao Peng, Lijun Xu, and Xinrong Hu. Vton-mp: Multi-pose virtual try-on via appearance flow and feature filtering. *IEEE Transactions on Consumer Electronics*, 2023. 3

[62] Jeffrey Zhang, Kedan Li, Shao-Yu Chang, and David Forsyth. Acdg-vton: Accurate and contained diffusion generation for virtual try-on, 2024. 2

[63] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[64] Na Zheng, Xuemeng Song, Zhaozheng Chen, Linmei Hu, Da Cao, and Liqiang Nie. Virtually trying on new clothing with arbitrary poses. In *ACM MM*, 2019. 3

[65] Na Zheng, Xuemeng Song, Zhaozheng Chen, Linmei Hu, Da Cao, and Liqiang Nie. Virtually trying on new clothing with arbitrary poses. In *ACM MM*, 2019. 3

[66] Zijian Zhou, Shikun Liu, Xiao Han, Haozhe Liu, Kam Woh Ng, Tian Xie, Yuren Cong, Hang Li, Mengmeng Xu, Juan-Manuel Pérez-Rúa, Aditya Patel, Tao Xiang, Miaojing Shi, and Sen He. Learning flow fields in attention for controllable person image generation. *arXiv preprint arXiv:2412.08486*, 2024. 1, 2, 3, 6, 7

[67] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *CVPR*, 2023.

[68] Luyang Zhu, Yingwei Li, Nan Liu, Hao Peng, Dawei Yang, and Ira Kemelmacher-Shlizerman. M&m vto: Multi-garment virtual try-on and editing. In *CVPR*, 2024. 2