Supplementary material for Clothe and Pose

A. The Input Dependency in VTON methods

In Figure 1, we observe that the garment worn by a user in the given image also influences the quality, snugness, and the overall fit of virtual try-on outputs across a variety of VTON methods. This occurs primarily because the current virtual try-on evaluation setup [1, 7] is incomplete by design. The current datasets contain tuples of $(G_A, P_A G_A)$ that are used to train and evaluate systems for virtual try-on, where G_A is an image capture of the garment and $P_A G_A$ is an image of a person wearing G_A. The standard virtual try-on evaluation setup consists of two axes for the evaluation: i) the paired setting: a masked version of $P_A G_A$, along with G_A are used to reconstruct the original $P_A G_A$ image, and ii) the unpaired setting: where a masked version of $P_A G_A$ and another random garment image G_X are used to synthesize a virtual try-on. Performance in the paired setting can be easily quantified due to the availability of ground truth $P_A G_A$ and thus the standard metrics such as PSNR, SSIM and LPIPS can help assess the fit, and overall quality of the try-on. It is also worth noting that the training setup is exactly same as the paired evaluation setup. However, performance on the unpaired set is measured using KID and FID metrics due to the unavailability of the ground truth, and hence can only measure the closeness of the generations to the training distribution. *As a result, training in this setup and optimizing for these standard benchmarks implicitly encourages models to exploit correlations between the user's existing clothing and the target garment, limiting our understanding of any virtual try-on method and making them less robust for practical deployment*.

Now, these drawbacks can be resolved with two techniques: i) using a larger agnostic mask during training, as in [13], but this type of masking limits the transfer of the original user characteristics like tattoos, earrings, and accessories to the try-on, or ii) by utilizing a "reference image" of another person in the target garment during generation to determine the style and fit of the garment, but these require access to the reference image for each garment that user wants to try. Even though these techniques can be helpful in avoiding dependence on inputs with a few caveats, they do not address the fundamental problem of flawed evaluation in try-on systems.

B. Analysis of DeepFashion Pose Transfer Dataset

We present a comprehensive analysis of identity leakage in the widely-used DeepFashion pose transfer dataset [14], revealing fundamental issues that compromise its validity as an evaluation benchmark, as well as a desirable training dataset. Our investigation examines the dataset's train-test split, which consists of 48K unique images in training set and 4K unique images in test set — these translate to 11,939 unique clothing items in the training set and 998 unique clothing items in the test set.

B.1. Overlapping user identities in test-train set

To assess the extent of identity overlap between training and test splits, we extracted facial embeddings using ArcFace [2] and computed pairwise similarities across splits. Our analysis reveals that 913 of the 998 test clothing items (91.5%) are worn by individuals whose facial embeddings match training set identities with a cosine similarity exceeding 0.7. This conservative estimate excludes cases where facial features could not be reliably extracted, such as images containing only lower body views with leggings, skirts, or denim, suggesting the actual overlap may be even higher. To further characterize the severity of this contamination, we performed DBSCAN clustering [3] on the extracted facial embeddings. The training set yielded 129 unique identity clusters, while the test set produced only 40 clusters. Analysis of the cluster centers reveals a striking pattern: all 40 test cluster centroids exhibit at least 60% similarity with some training cluster centroid, with 25 of these (62.5%) demonstrating similarities exceeding 70%. This systematic overlap indicates that the test set essentially represents a subset of training identities rather than a genuinely held-out evaluation set, despite the non-overlapping identities claim stated in Zhu et al. [14].

B.2. Implications for Model Development and Evaluation

The pervasive train-test identity overlap fundamentally undermines the dataset's utility for developing and evaluating pose transfer models. Modern diffusion architectures possess remarkable capacity for memorizing fine-grained facial features, particularly when trained on datasets with

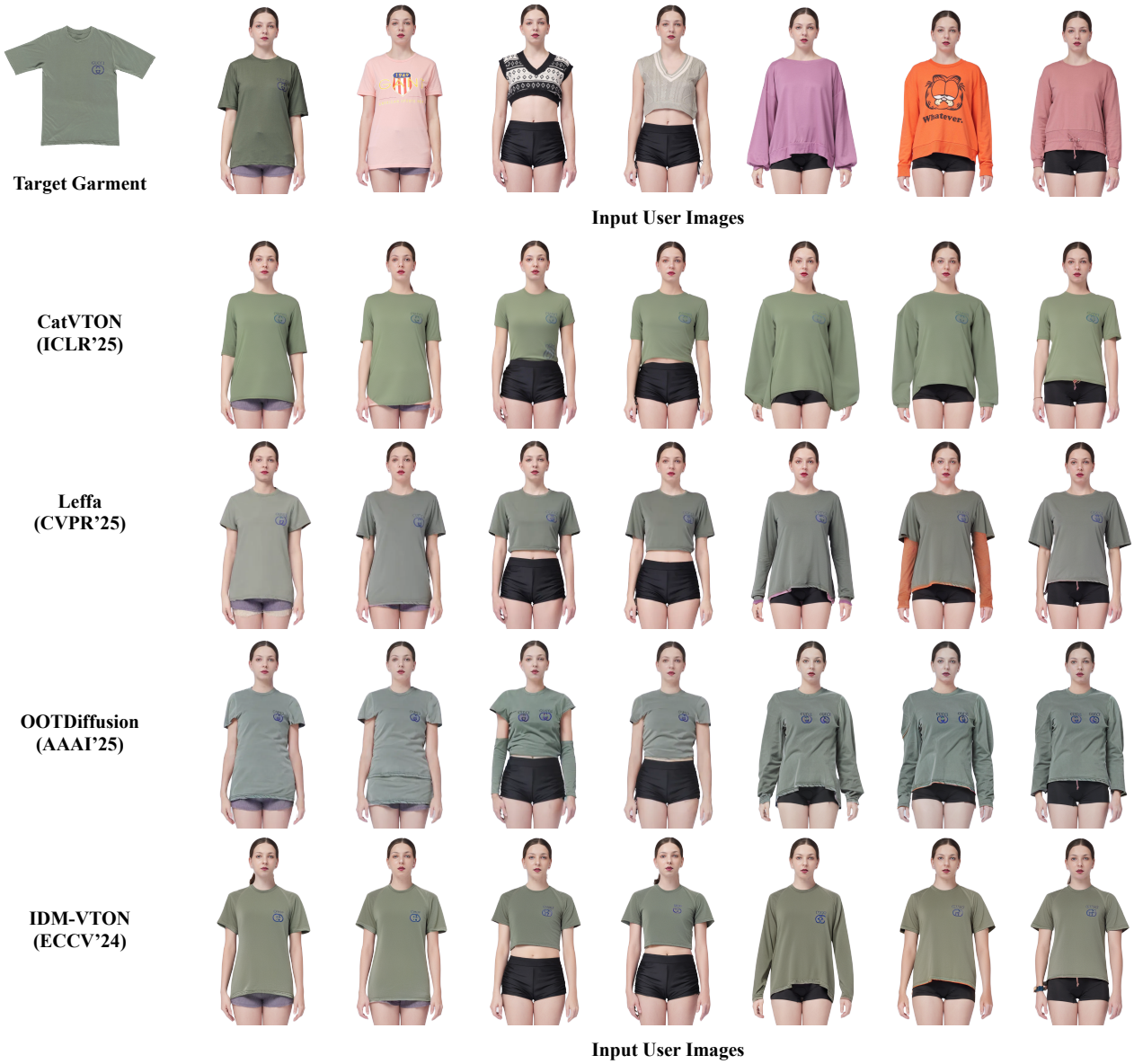

Figure 1. **Input Garment Dependency in VTON methods**. We showcase that the virtual try-ons obtained using different state-of-the-art methods show dependence on the original garment, caused by limited evaluation setup of traditional VTON systems.

sub-hundred unique identities. This memorization capability, combined with the observed identity overlap, creates a scenario where models can achieve artificially inflated performance metrics by leveraging memorized facial and body features rather than learning robust identity transfer mechanisms. The implications extend beyond mere metric inflation. Since identity preservation constitutes a primary evaluation criterion for pose transfer tasks, the dataset effectively provides this information "for free" through data leakage. Models evaluated on this benchmark may appear to excel at identity preservation when they are merely retrieving memorized faces from the training distribution. This creates a false sense of progress in the field and masks the true challenge of generalizing to novel identities. Consequently, the limited diversity of identities renders models trained on this dataset unsuitable for real-world deployment. Commercial applications require robust generalization to entirely unseen users, a capability that cannot be reliably assessed or developed using a dataset with such extensive train-test contamination. The current evaluation protocol essentially tests interpolation within a known identity space rather than extrapolation to novel identities, fundamentally misaligning benchmark performance with real-world requirements. Our reposing setup successfully deals

with these problems with an entirely new set of garments as well as user identities in the test set.

C. Additional Method & Experimental Details

C.1. Test-Time improvements

Improving inference initialization. Since SDXL's noise schedule does not completely destroy low-frequency image signals at $T = 999$, the training process never truly learns to generate images from pure noise $\mathcal{N}(0, 1)$. Consequently, initializing the latent $z_T \sim \mathcal{N}(0, 1)$ during inference creates a distribution mismatch between training and inference, resulting in inconsistent backgrounds and artifacts. To address this, we initialize z_T by partially preserving the background structure from the reference image:

$$\tilde{z}_T = \sqrt{\alpha_T}\mathcal{E}(\mathcal{R}) + \sqrt{1-\alpha_T}\epsilon, \quad \epsilon \sim \mathcal{N}(0, 1), \quad (1)$$

$$z_T = (1 - m) \odot \tilde{z}_T \ + \ m \odot \epsilon, \quad (2)$$

where α_T denotes the coefficient of forward diffusion process at T=999, $\mathcal{E}(\cdot)$ is the VAE encoder, and m is a binary mask indicating the human body region in $\mathcal{E}(\mathcal{R})$. This formulation: (1) aligns with the training process where the signal is not completely destroyed, (2) retains background structure information from the user reference image, and importantly, (3) conforms to the variance-preserving nature of the diffusion process at T=999 for the masked region.

Stabilizing conflicting conditions. During Stage 2 training, we sample pose transfer data 40% of the time to maintain identity preservation capabilities. However, we observed that despite applying patch-based dropout to garment regions in the reference image $\mathcal{R}$, the model occasionally exhibits undesirable behavior where it confuses and mixes appearance features from both the original garment (visible in $\mathcal{R}$) and the target garment conditions $\mathcal{G}$ during generation. This results in generated try-on images with hybrid garment appearances — for instance, combining patterns from the original clothing with colors from the target garment. To combat these artifacts, we mask the garment region in user images during the self-attention mechanism such that the latents corresponding to the resulting try-on do not attend to the garment region in the user image. This modification significantly reduces garment artifacts while maintaining strong identity preservation.

Decoupling Classifier-Free Guidance. Classifier-Free Guidance (CFG) [4] is crucial for improving the fidelity and adherence to conditioning signals in diffusion models. However, applying a single guidance scale to all conditions can be suboptimal when different aspects of generation require varying levels of control. We therefore employ a decoupled CFG strategy that applies different guidance scales to identity preservation and garment transfer. Concretely, we use the following equation to estimate the noise predicted at each timestep t:

$$\begin{aligned}\epsilon_t = \epsilon_\theta(\emptyset) &+ w_u \cdot (\epsilon_\theta(\mathcal{P}, \mathcal{R}) - \epsilon_\theta(\emptyset)) \\ &+ w_a \cdot (\epsilon_\theta(\mathcal{P}, \mathcal{R}, \mathcal{G}) - \epsilon_\theta(\mathcal{P}, \mathcal{R})),\end{aligned} \quad (3)$$

where w_u controls the strength of user identity and pose adherence, while w_a independently controls garment transfer fidelity, and $\{\mathcal{P}, \mathcal{S}, \mathcal{G}\}$ are VAE-encoded latents of corresponding condition images. This decoupled approach proves essential for balancing the competing objectives of maintaining user identity while accurately transferring garment appearance.

C.2. Experimental Details

Training Datasets. Our training data comprises pose transfer data and multi-pose try-on data. For Pose Transfer Data, we collected 81K paired image samples along with 80K video sequences from publicly available fashion websites. The image pairs and video frames provide supervision for learning pose-transfer task while maintaining garment consistency across a wide range of user identities. For this data, garment image(s) of either top or bottom garments are available for each sample, but not both. For multi-pose virtual try-on data, we utilize a licensed dataset consisting of 24K videos captured from 268 unique users. Each video sequence captures a single user performing a standardized set of movements across multiple garments, against a plain background. We pair random videos of the same users and then sample frames from each one to train our model for joint clothing and posing task. Garment images for both top and bottom garments are available for each sample in this dataset.

Implementation Details. We initialize our model using SDXL-Base-1.0 weights. For the multi-stream blocks, we initially clone the weights from the original stream to initialize parallel streams. Training images are resized to 1024×768 resolution while preserving aspect ratios through body mask-guided resizing and white-pixel padding. We employ Fully Sharded Data Parallel (FSDP) training across 8 H100 GPUs with an effective batch size of 768 to train our model end-to-end. To optimize memory usage, we combine bf16 mixed-precision training with an 8-bit AdamW optimizer — Stage 1 runs for 60K steps and Stage 2 for 55K steps, both using a linear learning rate warmup to 5e-5. For generation, we use the DDIM Sampler [10] sampler for 50 denoising steps.

D. Additional Results

In Table 1, we report results for Clothe and Pose using more pipeline-based baselines. The additional baselines have been underlined in the Method column. Specifically, we introduce a proprietary Kolors VTON model by KlingAI [5] and pair it with the Leffa [12], Kontext [6]

Table 1. Additional comparisons for quantitative evaluation of Clothe and Pose task. 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
Kolors+Leffa	0.259	16.916	80.368	0.259	16.779	80.599	0.250	17.104	81.518	0.246	17.024	81.962
IDM-VTON+Kontext	0.313	16.222	79.627	0.284	16.820	80.651	0.302	16.544	80.875	0.280	16.884	81.686
OOTDiffusion+Kontext	0.316	15.953	78.915	0.291	16.391	79.902	0.304	16.312	80.302	0.286	16.577	81.050
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
Kolors+Kontext	0.314	16.238	79.500	0.285	16.823	80.559	0.300	16.577	80.827	0.281	16.936	81.619
Kolors+`gpt-image`	0.313	15.828	77.930	0.292	16.281	79.103	0.309	16.020	79.316	0.288	16.328	80.251
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**

and the proprietary `gpt-image-1` [8] text editing model by OpenAI. Additionally, we also report results for IDM-VTON+Kontext and OOTDiffusion+Kontext baselines for completeness. We observe that the Kolors VTON-based pipelines perform better than their open-source/academic counterparts due to the better quality try-ons produced by Kolors VTON. Overall, pipeline-based methods struggle at Clothe and Pose, Qwen-Image-Edit struggles in synthesizing back views but performs decently in other pose configurations, and our method excels and outperforms all the baselines in all pose configurations, across all the metrics.

E. Failure Cases and Limitations

Failure Cases. We visualize our common failure cases in Figure 2. These include: i) difficulty in synthesizing partially visible faces in back views, ii) bad anatomy hands, and iii) different side-face view in generated image and the ground truth. Difficulties in synthesizing partially visible faces in back-views, and back-views in general, arises due to the skewness of training data towards frontal and side poses. We attribute the bad anatomy of hands to a combination of the underlying base model SDXL [9], which has been known to struggle with hands, and the error in accurate hand pose estimation by the the ViTPose [11] model which is used to skeleton generation. Finally, for any given frontal view of the face, there are multiple possible side-views and this results in different side-view face synthesis by our model, in comparison to the available ground truth.

Limitations of proposed model. Our method processes frontal and back views of the top and garments as inputs. Even if the actual garment conditions are missing, we feed it a gray image to act as "null" condition and denote its absence. This design limits the efficiency of our approach because both front and back view of the garments are not necessary to synthesize every target pose, i.e., both views are only required when both front and back garments are visible in the target pose. We believe making Clothe and Pose models more efficient by choosing the garment views adaptively is an important direction for future work. Additionally, our current approach is limited to single-person scenarios and extending our unified framework to support multi-person editing simultaneously remains an exciting and open challenge.

F. Ethics Statement

We are committed to conducting research that adheres to the highest ethical standards. The human subject data collection for this work was carried out with careful consideration of participant welfare.

Data Collection Conditions. The data collection was conducted in a fully equipped professional studio environment. To ensure the safety and comfort of the participants, each subject was assisted by a dedicated support team of five professionals, including photographers and stylists. All sessions were conducted during regular working hours with appropriate breaks.

Informed Consent. We confirm that informed consent was obtained from all participants. Each human subject signed a formal service agreement prior to the session, explicitly consenting to the data capture and its usage for research purposes. Participants were informed about the nature of the research, how their data would be processed, and their rights regarding data usage. All participants were over 18 years of age.

Fair Compensation. All human subjects were fairly and generously compensated for their time and effort for data collection. The compensation rates provided were highly competitive (comparable to the hourly rates of senior engineering roles in the country of origin), ensuring the ethical treatment of all participants involved.

Figure 2. **Failure Cases**. Our method suffers from two common failure modes: difficulty in synthesizing partial side face in back-view try-ons (top row), and bad anatomy of hands(top and bottom row). Additionally, the user identity can get altered slightly compared to the ground truth in side-views because for a given front-view of the face in the user image, many possible side-views exist in the solution space.

References

[1] 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] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 1

[3] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996. 1

[4] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[5] KlingAI. Klingai virtual try on api, 2025. 3

[6] 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

[7] 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

[8] OpenAI. gpt-image-1 api, 2025. 4

[9] 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. 4

[10] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *CoRR*, abs/2010.02502, 2020. 3

[11] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *NeurIPS*, 2022. 4

[12] 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. 3

[13] 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. 1

[14] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *CVPR*, 2019. 1