FUSEDVTON: EARLY IMAGE FUSION FOR EFFICIENT VIRTUAL TRY-ON

Anonymous authors
Paper under double-blind review



Figure 1: **FusedVTON combines strong quality with low inference cost.** Left: four garment-conditioning paradigms and their legend. Upper right: DressCode unpaired FID versus inference TFLOPs for SD1.5 methods at 512×384. Lower right: FLUX.1 compute and latency scaling for one to four garments on Garments2Look at 1024×768 (logarithmic vertical axes). Circle size encodes single-garment latency; stars mark FusedVTON (ours). Efficiency excludes external preprocessing.

ABSTRACT

Existing image-based virtual try-on methods typically use explicit garment warping or model garment–person correspondence through reference networks, cross-attention, or self-attention over concatenated tokens. Although effective, these designs introduce architectural complexity and computational overhead, especially in multi-garment try-on, where additional garments require additional warping operations or garment representations. We view garment transfer as coarse spatial placement followed by non-rigid generative refinement and propose **FusedVTON** to implement this formulation. Reference garments are resized and placed within the clothing-agnostic person image before VAE encoding. The resulting composite is processed through a single, fixed-length image-token stream without a separate garment branch or appended garment tokens. Experiments on VITON-HD, DressCode, DressCode-MR, and Garments2Look show competitive quality. Separate profiling at 512×384 without classifier-free guidance measures 53.9% and 51.7% fewer model-inference FLOPs than the corresponding SD1.5 and FLUX.1 CatVTON variants. At 1024×768, counted model-inference computation remains constant across one to four garment inputs in the evaluated scaling setup.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Image-based virtual try-on aims to synthesize a person wearing reference garments while preserving garment appearance and person identity. Warping-based methods (Han et al., 2018; Wang et al., 2018; Ge et al., 2021; Gou et al., 2023) explicitly align garments before image synthesis. Attention-based approaches instead integrate garment features during generation, using dedicated conditioning branches in StableVITON (Kim et al., 2024) and IDM-VTON (Choi et al., 2024), or spatial and token concatenation in CatVTON (Chong et al., 2024), Any2Any Tryon (Guo et al., 2025), and Voost (Lee & Kwak, 2025). These attention-based designs allocate each reference garment a separate set of features or tokens in addition to those representing the person.

These separate garment representations introduce computation for feature extraction and attention-based interaction with the person. The overhead increases in multi-garment virtual try-on, which synthesizes a person wearing multiple reference garments simultaneously. WG-VITON (Park & Park, 2022) separately warps upper- and lower-body garments, while Any-Fit (Li et al., 2024), BootComp (Choi et al., 2025), and MMTryon (Zhang et al., 2024) integrate per-garment features through attention. FastFit (Chong et al., 2025) avoids repeated reference encoding through caching, but attention still processes a separate cache for each garment. Thus, the processed conditioning grows with garment count. This motivates letting multiple garments share the target person's existing spatial canvas.

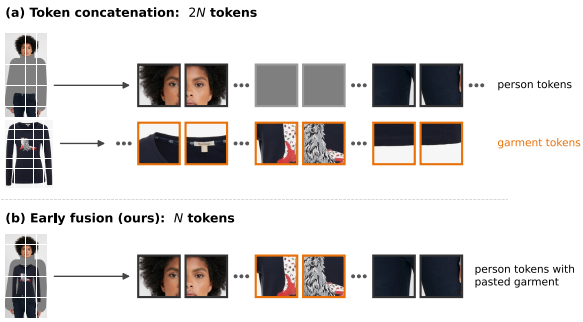

Figure 2: **Garment conditioning within a fixed token grid.** Token concatenation processes separate person and garment grids ($2N$ tokens), while early fusion places the garment within the person's grid (N tokens).

While recent work reduces this computational overhead through architectural simplification and feature caching, a separate question concerns the role of the additional garment representation itself. The approaches above vary how garment features are extracted and integrated, while retaining features or tokens allocated separately from the target person. These designs do not by themselves distinguish the contribution of separate garment representations from that of the additional spatial resolution they provide. This motivates the following question:

Can virtual try-on be achieved by incorporating reference garments within the target person's existing spatial token grid, without allocating additional garment tokens?

In concatenation-based methods such as CatVTON, garment and person representations are processed by a shared backbone and interact through self-attention. This shared processing does not establish whether an additional garment token grid is necessary for virtual try-on. *We hypothesize that reference garments can instead be represented within the person's existing spatial grid* (Fig. 2), using the clothing-agnostic region whose original clothing appearance is masked out but whose spatial positions remain part of the model input. *Additional garment tokens may improve detail preservation by allocating more spatial positions to garment appearance.*

We propose FusedVTON, which fuses reference garments with the clothing-agnostic person directly in image space before VAE encoding. Aspect-ratio-preserving placement supplies approximate location and scale; the backbone performs subsequent body-conditioned refinement. The fused appearance condition and DensePose are supplied through channel-wise conditioning, preserving one spatial grid without appended garment tokens. A fine-tuned backbone then adapts garment shape, boundaries, occlusions, and appearance to the target person. FusedVTON supports Stable Diffusion (Rombach et al., 2022) and FLUX.1 (Black Forest Labs, 2024) without a separate garment encoder, ReferenceNet, appended garment tokens, or category-specific network modules. At fixed resolution, additional garments change the fused image's content without enlarging the denoiser's spatial grid. Figure 6 measures the resulting inference-cost scaling. We evaluate FusedVTON on single- and multi-garment benchmarks in terms of generation quality and computational efficiency. Our contributions are summarized as follows:

- We investigate early garment–person fusion within the clothing-agnostic person image, allowing garment appearance and person context to share a spatial representation before VAE encoding.
- We propose **FusedVTON**, which combines coarse garment placement with body-conditioned generation using SD1.5 and FLUX.1, without a separate garment branch or appended garment tokens. At fixed image resolution, denoiser computation is independent of garment count.
- Experiments show competitive single- and multi-garment quality. Separate profiling measures 53.9% and 51.7% lower model-inference FLOPs than same-backbone CatVTON at 512×384, with constant counted computation over the evaluated garment counts at fixed resolution.

2 Related Work

Explicit Garment Warping IVTON. A fundamental challenge in image-based virtual try-on is bridging the large spatial discrepancy between the reference garment and the target person. Early approaches address this problem by explicitly aligning geometry before image synthesis. Representative methods, including CP-VTON (Wang et al., 2018), ACGPN (Yang et al., 2020), and PF-AFN (Ge et al., 2021), first estimate a spatial transformation to warp the garment to the target body configuration before generating the try-on image. This design has also been adopted by diffusion-based frameworks such as DCI-VTON (Gou et al., 2023) and FIP-VTON (Dam et al., 2024), where warped garments are incorporated as additional conditioning signals. Warped-condition quality can affect subsequent synthesis, and garment alignment errors may appear in the generated result.

Two-Stream Interaction in IVTON. Recent diffusion-based methods replace explicit geometric alignment with learned feature interaction during denoising. TryOnDiffusion (Zhu et al., 2023) first introduced cross-attention to learn garment-person correspondence during diffusion. Building on this, StableVITON (Kim et al., 2024), IDM-VTON (Choi et al., 2024), CatVTON (Chong et al., 2024), CAT-DM (Zeng et al., 2024), PromptDresser (Kim et al., 2025), Leffa (Zhou et al., 2025), Any2Any Tryon (Guo et al., 2025), and Voost (Lee & Kwak, 2025) process the garment and target person as two separate visual streams, relying on cross-attention or token-concatenation-based self-attention to model their interaction throughout denoising. Consequently, maintaining separate representations of garments and persons has become the dominant paradigm in modern IVTON. Reference branches introduce additional feature computation, while spatial concatenation increases the sequence length processed by the denoiser. Although these approaches differ in feature extraction and interaction, they retain garment features or tokens allocated separately from those representing the target person. Dedicated encoders introduce additional feature computation, while spatial or token concatenation increases the sequence length processed by attention.

FusedVTON incorporates garment appearance within the clothing-agnostic person image before VAE encoding. Coarse placement specifies approximate garment location and scale without estimating a learned non-rigid deformation. The resulting condition retains the person's existing spatial grid, without a separate garment branch or additional garment tokens. A generative backbone then refines the composite into the final try-on image, conditioned on DensePose.

3 Methodology

We investigate whether virtual try-on can be achieved by representing reference garments within the target person's existing spatial grid. FusedVTON implements this hypothesis through early image fusion: garments are incorporated into the clothing-agnostic person image before VAE encoding. The resulting condition supplies garment appearance without a separate garment branch or additional spatial tokens. Figure 3 provides an overview.

3.1 Garment Conditioning within a Fixed Spatial Grid

Given a person image $I^{\mathrm{p}} \in \mathbb{R}^{3\times H\times W}$ and reference garments $\{C_i\}_{i=1}^{G}$, our goal is to synthesize the person wearing these garments while preserving identity and non-garment regions. Spatial concatenation, as used in CatVTON (Chong et al., 2024), allocates additional positions to the reference garment. We instead construct a conditioning image $I^{\mathrm{f}} \in \mathbb{R}^{3\times H\times W}$ within the existing image dimensions. The clothing-agnostic region provides space for garment appearance: its original clothing content is removed, while the surrounding person and background remain available as context.

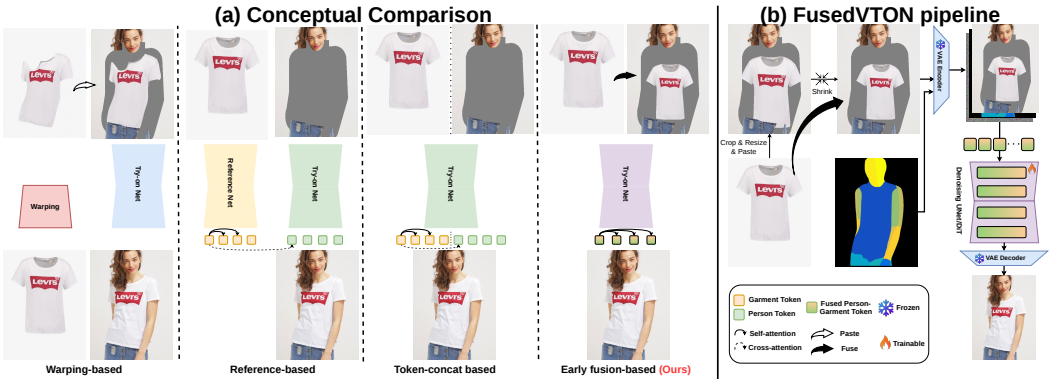

Figure 3: (a) Comparison of garment conditioning strategies. FusedVTON incorporates reference garments within the person image before VAE encoding, retaining the existing spatial grid. (b) The fused image and DensePose provide channel-wise latent conditioning to the generation backbone.

Within this fixed grid, placement aims to retain as much garment resolution as the available region permits. Let $\{(C_i, M_i^{\mathrm{g}}, q_i)\}_{i=1}^{G}$ denote the reference items, their foreground masks, and their categories (tops, bottoms, dresses, shoes, or bags; bags are used only for Garments2Look). Category q_i determines the dilated clothing-agnostic placement region A_i, detailed in Appendix A.

We crop each garment to its tight foreground bounding box and fit the largest axis-aligned box of the same aspect ratio inside A_i. Cropping removes surrounding background, while the largest feasible placement allocates more image area to garment appearance. The aspect-ratio constraint preserves garment proportions, and placement within A_i supplies approximate location without estimating a body-aligned non-rigid deformation. Among equally sized feasible placements, we select the one closest to the region centroid. For multiple garments, overlapping boxes are separated and refitted.

Denoting the resulting box by r_i, we resize and place the RGB crop using Lanczos interpolation and its mask using nearest-neighbor interpolation:

$$\bar{C}_i = \mathcal{T}_{r_i}^{\mathrm{Lanczos}}(C_i^{\mathrm{crop}}), \qquad M_i^{\mathrm{f}} = \mathcal{T}_{r_i}^{\mathrm{NN}}(M_i^{\mathrm{crop}}), \tag{1}$$

where $\mathcal{T}_{r_i}$ maps the crop to its allocated box on the person image.

Let $\mathcal{S}$ index the selected garments and $M^{\mathrm{a}} = \bigvee_{i\in\mathcal{S}} A_i$ denote their combined clothing-agnostic mask. We replace the original appearance within this mask with a constant gray value g and composite the placed garments:

$$\begin{aligned} I^{\mathrm{a}} &= (1 - M^{\mathrm{a}}) \odot I^{\mathrm{p}} + g M^{\mathrm{a}}, \\ I^{\mathrm{f}} &= \mathrm{Composite}\left(I^{\mathrm{a}}, \{(\bar{C}_i, M_i^{\mathrm{f}})\}_{i\in\mathcal{S}}\right). \end{aligned} \tag{2}$$

Here, $\odot$ denotes element-wise multiplication with the mask broadcast over RGB channels, and $\mathrm{Composite}$ pastes the transformed crops using their foreground masks. The resulting image supplies garment appearance at a coarse location and scale, with subsequent body-conditioned refinement by the trained backbone. Adding garments changes the condition's content without expanding its spatial grid, although the available area limits the resolution retained per garment.

3.2 Latent Conditioning and Training

We supply the fused-image latent and target pose condition through the denoiser input, without a separate garment encoder or appended garment tokens. In paired training, let $I^{\star}$ denote the ground-truth target image. A pretrained VAE encoder $\mathcal{E}$ (Rombach et al., 2022) encodes it and I^{f}:

$$z_0 = \mathcal{E}(I^{\star}), \qquad z_{\mathrm{f}} = \mathcal{E}(I^{\mathrm{f}}). \tag{3}$$

The pose condition z_{p} is derived from the target DensePose representation P (Güler et al., 2018). FLUX uses $z_{\mathrm{p}} = \mathcal{E}(P)$, while SD1.5 directly resizes the DensePose RGB image to the latent spatial resolution using bilinear interpolation. For the FLUX flow path, we sample $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and form

$$z_t = (1 - \sigma_t) z_0 + \sigma_t \epsilon, \qquad y_t = \epsilon - z_0. \tag{4}$$

Table 1: FLUX denoiser compute at 512×384, grouped by module family. "Compute share" is each family's percentage of FusedVTON FLOPs; "saving share" is its share of the FLOPs avoided relative to CatVTON-FLUX.1. Only shared families are shown; $<1\%$ of the saving comes from baseline-specific operations. The FLOP counter excludes normalization and element-wise operations.

	Module family	Scaling in L	Compute share	Saving share
FLUX.1	MLP projections	$\Theta(L)$	69.3%	63.9%
	Attention projections (Q/K/V/O)	$\Theta(L)$	26.7%	24.6%
	Attention score/value products	$\Theta(L^2)$	4.0%	11.1%
	Input/output projections	$\Theta(L)$	$< 0.1\%$	$< 0.1\%$
	Conditioning and AdaLN	$\Theta(1)$	$< 0.1\%$	0.0%

Here σ_t is the noise level at timestep t. The SD1.5 instantiation uses its native noise-prediction target $y_t = \epsilon$. We condition the denoiser by concatenating these maps along the channel dimension:

$$X_t = [\, z_t; z_{\mathrm{f}}; z_{\mathrm{p}} \,], \qquad \hat{y}_t = f_\theta(X_t, t), \tag{5}$$

where $[\,\cdot;\cdot\,]$ denotes channel-wise concatenation. To accommodate the additional conditioning channels, we adapt the denoiser input projection accordingly, reuse pretrained noisy- and masked-image weights, and zero-initialize the DensePose weights. Appendix B details the backbone-specific channel layouts. This introduces neither a separate garment branch nor additional garment tokens.

For image classifier-free guidance, training zeros z_{f} with probability 0.1 while always retaining z_{p}. For multi-garment inputs, z_{f} represents the complete selected outfit, so guidance is applied jointly rather than separately to each garment. At inference, two denoiser evaluations give

$$\hat{y}_t^{\mathrm{cfg}} = \hat{y}_t^0 + s_{\mathrm{img}}\big(\hat{y}_t^{\mathrm{f}} - \hat{y}_t^0\big), \tag{6}$$

where $\hat{y}_t^0$ uses a zero fused-image condition and $\hat{y}_t^{\mathrm{f}}$ uses z_{f}; both retain DensePose. We use $s_{\mathrm{img}} = 2.5$ for the reported FLUX models.

Computational advantage of fused conditioning. Early fusion reduces the image-token sequence rather than the transformer width. Let L be the image-token length in each FLUX block. At 512×384, FusedVTON processes 768 image tokens, whereas spatial concatenation in CatVTON-FLUX.1 (Chong et al., 2024) processes 1,536. Table 1 separates this cost by module family.

Only the attention score and value products scale quadratically in L; the Q/K/V/output and MLP projections are token-linear. In the single-garment setting, spatial concatenation doubles L. It therefore quadruples the attention products but only doubles the projections, which constitute 96.0% of FusedVTON's counted FLOPs at this resolution. This composition explains why CatVTON-FLUX.1 uses $2.09\times$, rather than $4\times$, as many counted FLOPs per denoising forward pass (Table 2). Of the avoided computation, 88.4% comes from token-linear transformer projections and 11.1% from quadratic attention products. However, token-linear scaling is not garment-independent: spatial concatenation appends one image-sized token grid per garment, so every garment adds another full grid of the dominant projections, while the attention-product cost grows quadratically as the grids accumulate. FusedVTON composites all garments before tokenization and retains one fixed-length grid, so neither cost grows with garment count at fixed resolution.

4 Experiments

Datasets. We evaluate single-garment try-on on VITON-HD (Choi et al., 2021) and DressCode (Morelli et al., 2022), and multi-garment try-on on DressCode-MR (Chong et al., 2025) and Garments2Look (Hu et al., 2026). Their train/test splits have $11{,}647/2{,}032$, $48{,}392/5{,}400$, $25{,}779/2{,}400$, and $28{,}711/1{,}511$ samples, respectively. Paired and unpaired protocols use 512×384. Target parses, keypoints, and clothing-agnostic masks build the fused images; DensePose (Güler et al., 2018) provides pose conditioning. Appendix A details datasets and preprocessing.

Implementation Details. We instantiate FusedVTON with Stable Diffusion v1.5 inpainting (Rombach et al., 2022) and FLUX.1-Fill (Black Forest Labs, 2024). Both reuse pretrained

Table 2: Generation quality and computational efficiency on VITON-HD and DressCode, grouped by Stable Diffusion-based and FLUX.1-based methods. Subscripts p and u denote paired and unpaired settings, respectively. All metrics, including efficiency, use 512×384 resolution. Within each backbone group, the best and second-best reported results are shown in **bold** and underlined, respectively. "–" indicates that no pretrained model is available for the corresponding dataset.

Method	Conditioning Style	VITON-HD						DressCode						Efficiency	
		$FID_p \downarrow$	$KID_p \downarrow$	$SSIM_p \uparrow$	$LPIPS_p \downarrow$	$FID_u \downarrow$	$KID_u \downarrow$	$FID_p \downarrow$	$KID_p \downarrow$	$SSIM_p \uparrow$	$LPIPS_p \downarrow$	$FID_u \downarrow$	$KID_u \downarrow$	TFLOPs↓	Latency (s)↓
Stable Diffusion-based methods															
LaDI-VTON	Pre-warp	7.052	1.478	0.8696	0.0620	9.604	1.809	5.198	1.794	0.8982	0.0525	7.049	2.429	33.3	0.998
DCI-VTON	Pre-warp	6.011	0.494	0.8892	0.0525	9.438	0.853	–	–	–	–	–	–	31.7	1.180
StableVITON	ReferenceNet	6.471	1.484	**0.8912**	**0.0491**	9.491	1.644	–	–	–	–	–	–	102.0	3.898
IDM-VTON	ReferenceNet	5.911	0.785	0.8663	0.0618	9.143	1.086	–	–	–	–	–	–	135.2	6.852
OOTDiffusion	ReferenceNet	6.014	0.461	0.8473	0.0606	9.166	**0.755**	3.747	0.732	0.8971	0.0430	7.362	2.557	46.1	1.414
PromptDresser	ReferenceNet	11.484	6.071	0.8100	0.1102	13.376	6.258	7.361	3.152	0.8651	0.0788	8.271	3.107	146.2	5.394†
CatVTON	Token concat	**5.670**	**0.431**	0.8780	0.0561	9.359	1.182	3.210	0.578	0.9081	**0.0395**	5.298	1.232	66.1	0.995
FusedVTON	Early image fusion	5.871	0.455	0.8857	0.0583	**9.071**	1.178	**3.149**	**0.557**	**0.9180**	0.0404	**4.807**	**0.875**	**30.5**	**0.839**
FLUX.1-based methods															
Any2AnyTryon	Token concat	7.221	0.957	0.8608	0.0867	9.202	1.166	2.620	0.702	**0.9380**	0.0346	5.244	1.237	1,201.4	2.720
CatVTON	Token concat	6.605	0.651	0.8473	0.0855	9.778	1.215	4.283	1.109	0.8862	0.0535	6.356	1.788	606.9	1.664
RefTon	Token concat	6.110	1.103	0.8635	0.0629	9.263	1.290	3.392	0.935	0.9033	0.0420	5.328	1.100	1,188.3	3.024
FusedVTON	Early image fusion	**5.089**	**0.207**	**0.8889**	**0.0471**	**8.153**	**0.239**	**2.543**	**0.205**	0.9159	**0.0332**	**4.695**	**0.623**	**292.9**	**1.255**

Table 3: Multi-garment comparison on DressCode-MR and Garments2Look. Subscripts p and u denote paired and unpaired settings. All metrics use 512×384 resolution. Efficiency is measured with four garments on Garments2Look over model inference, excluding external preprocessing. Within each dataset, the best and second-best results are shown in **bold** and underlined, respectively.

Method	DressCode-MR						Garments2Look						Efficiency	
	$FID_p \downarrow$	$KID_p \downarrow$	$SSIM_p \uparrow$	$LPIPS_p \downarrow$	$FID_u \downarrow$	$KID_u \downarrow$	$FID_p \downarrow$	$KID_p \downarrow$	$SSIM_p \uparrow$	$LPIPS_p \downarrow$	$FID_u \downarrow$	$KID_u \downarrow$	TFLOPs↓	Latency (s)↓
CatVTON (FLUX.1)	8.545	0.902	0.8555	0.0675	11.318	1.491	12.98	0.515	0.8919	0.0394	16.85	2.042	1688.1	3.725
FastFit (FLUX.1)	7.797	**0.540**	0.8702	0.0557	**10.951**	**1.241**	13.17	0.500	0.9035	0.0416	**15.48**	**0.968**	383.0	1.921
RefTon (FLUX.1)	9.833	1.990	0.8656	0.0651	12.486	2.402	13.76	1.089	0.9064	0.0389	16.56	1.204	3302.0	7.645
FusedVTON (FLUX.1)	**7.287**	0.587	**0.8807**	**0.0468**	10.960	1.549	**12.86**	**0.443**	**0.9077**	**0.0366**	16.29	1.410	**292.9**	**1.189**

masked-image input weights for the fused-image latent and zero-initialize the added DensePose channels. SD expands its input convolution from 9 to 11 channels and removes cross-attention; FLUX reduces its 384-channel input projection to 192 channels and removes the text pathway. We optimize the remaining denoisers. Both variants are trained for 80,000 steps on the combined VITON-HD and DressCode training sets at 512×384, with learning rate 10^{-5} and image-condition dropout 0.1. SD uses AdamW with batch size 128; FLUX uses Adafactor with batch size 16. Appendix B gives the remaining hyperparameters.

Evaluation Metrics. We report SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), FID (Heusel et al., 2017), and KID for paired and FID and KID for unpaired evaluation. We separately profile model-inference FLOPs and latency over all required modules, excluding external preprocessing. Profiling uses 50 steps for SD and 28 for FLUX, without classifier-free guidance, at batch size one in bfloat16 on an NVIDIA B200. Appendix C gives further evaluation details.

4.1 Quantitative Comparison

Cloth-to-person (C2P). Table 2 reports generation quality and efficiency on VITON-HD and DressCode under paired and unpaired settings. FusedVTON maintains comparable quality across both backbone groups. Under SD1.5, its VITON-HD paired FID and KID are 5.871 and 0.455, compared with CatVTON's 5.670 and 0.431. These results suggest that garment appearance can be represented within the person's existing spatial grid while retaining comparable quality.

Inference efficiency. Table 2 also shows reduced model-inference cost. Relative to CatVTON with the same backbone, FusedVTON reduces FLOPs by 53.9% under SD1.5 and 51.7% under FLUX.1, with latency decreasing from 0.995 to 0.839 s and from 1.664 to 1.255 s, respectively. Early fusion removes the additional garment-token grid processed during denoising.

Multi-garment try-on. Table 3 shows comparable multi-garment quality on DressCode-MR and Garments2Look. FusedVTON and FastFit obtain similar DressCode-MR unpaired FID values of 10.960 and 10.951. For four garments on Garments2Look at 512×384, FusedVTON requires 292.9 TFLOPs and 1.189 s, compared with FastFit's 383.0 TFLOPs and 1.921 s. This extends the quality–efficiency trade-off to conditioning on multiple items.

Figure 6 shows efficiency from one to four garments at 1024×768 on Garments2Look, the only one of the two datasets with four-garment outfits. FusedVTON's model-inference computation is

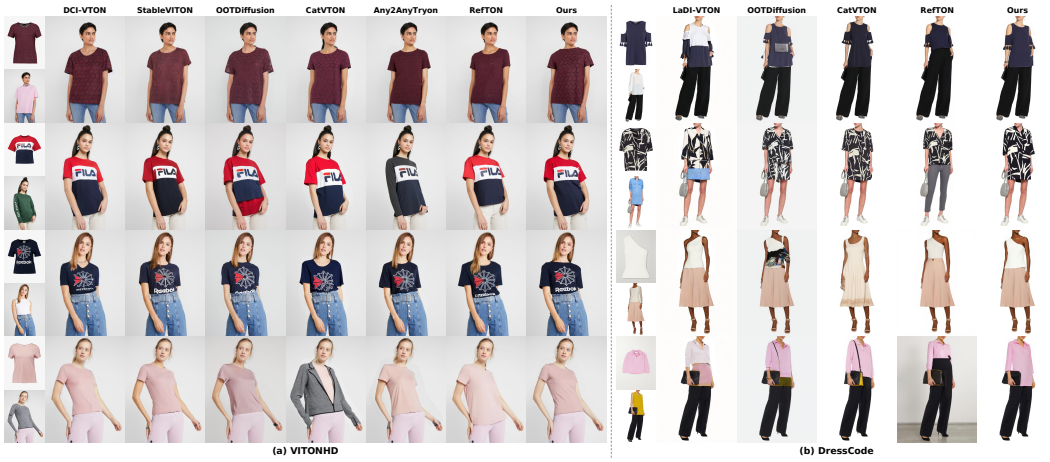

Figure 4: Qualitative comparison on VITON-HD and DressCode under the unpaired cloth-to-person (C2P) setting. Each row shows the target garment above the source person in the narrow reference column, followed by method outputs. VITON-HD compares DCI-VTON, StableVITON, OOTDiffusion, CatVTON, Any2AnyTryon, RefTON, and ours; DressCode compares LaDI-VTON, OOTDiffusion, CatVTON, RefTON, and ours. Each dataset includes four examples. Zoom in for details.

constant as garments are added, and its measured latency stays nearly constant (2.736–2.739 s), whereas both costs grow for all three baselines. CatVTON and RefTon concatenate a separate token stream per garment, while FusedVTON fuses all garments into a single image before VAE encoding, so the backbone always processes one image-token stream regardless of garment count.

4.2 Qualitative Comparison

Cloth-to-person (C2P). Figure 4(a) shows VITON-HD examples with repeated patterns and printed graphics. FusedVTON retains recognizable FILA lettering, the Reebok graphic, and the burgundy shirt's pattern. Figure 4(b) shows DressCode examples with different garment silhouettes, including a cut-out shoulder top and an asymmetric white neckline. These examples show garment appearance preserved as it adapts to the target person. Appendix E provides more examples.

Multi-garment try-on. Figure 5 compares selected two- to four-item outfits from DressCode-MR and Garments2Look. On DressCode-MR, FusedVTON retains recognizable sportswear lettering and sweater stripes in the three-item examples, together with the requested bottoms and shoes. The two-item examples show patterned dresses combined with heels. The Garments2Look examples extend the comparison to a four-item outfit with a bag. These results illustrate joint transfer of garment appearance within one conditioning canvas.

4.3 Controlled Analysis

Qualitative marker-transport probe. The fused condition and noisy latent are concatenated along channels and share one spatial token grid, so attention maps cannot separate source garment tokens from output tokens. We therefore probe correspondence by intervention rather than by attention visualization. We insert a small colored marker at a chosen location on the garment in the fused condition I^{f}, generate the try-on result, and detect where the marker appears in the output. Each marker is inserted and generated in a separate run, so markers do not interact. Whereas attention maps indicate where a model attends, this probe shows where the model transports each garment location. Figure 7 connects each inserted marker to its detected output location. In the displayed examples, markers near the neckline, chest graphics, and lower torso map to corresponding garment regions. These mappings suggest spatially organized garment transfer under fused conditioning.

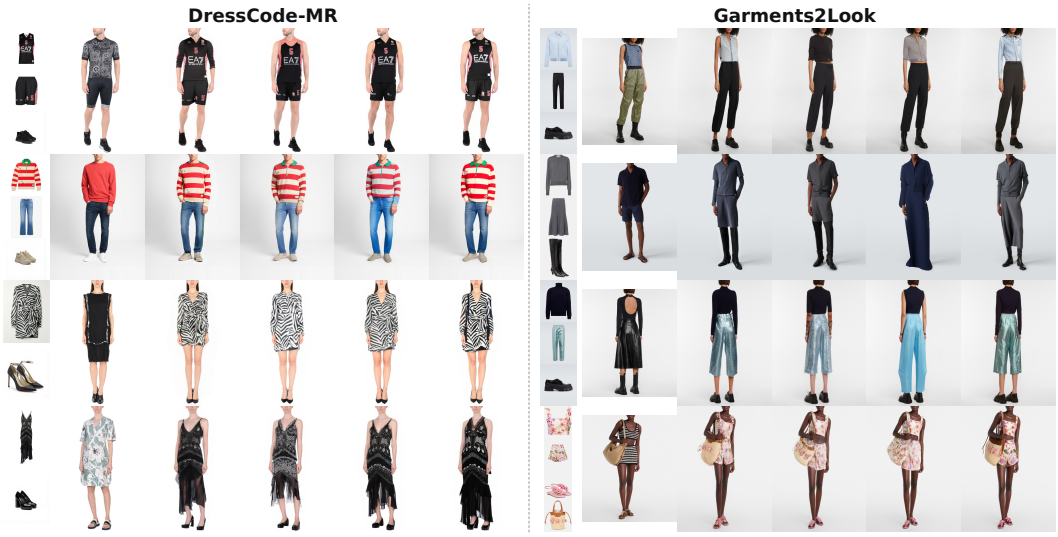

Figure 5: Multi-garment comparison on DressCode-MR (left) and Garments2Look (right). Both panels share one column order: garment references, source person, CatVTON, FastFit, RefTON, and FusedVTON (Ours). DressCode-MR has two three-item and two two-item outfits; Garments2Look has three three-item and one four-item outfit. Shoes and bags each count as one item.

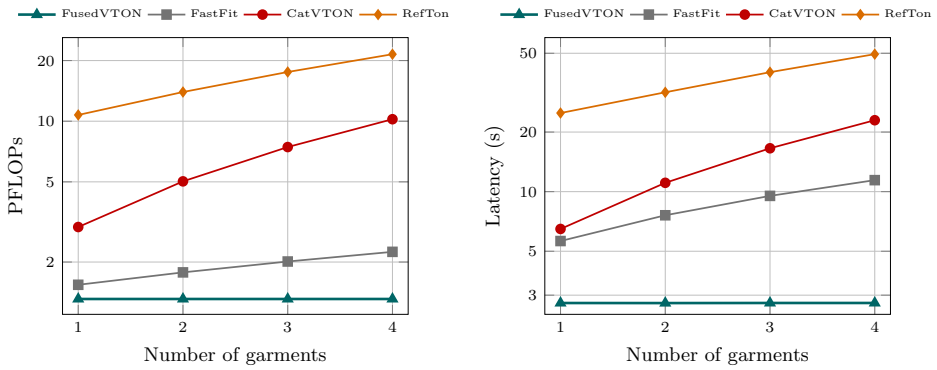

Figure 6: FusedVTON maintains constant model-inference compute and nearly constant latency as garment count increases. We profile one to four nested garment inputs from one Garments2Look outfit at 1024×768, using FLUX.1, 28 steps, no classifier-free guidance, and batch size one on an NVIDIA B200. FLOPs (left, PFLOPs) and latency (right) cover VAE encoding, method-specific encoders and cache construction, denoising, and VAE decoding, excluding external preprocessing. FastFit uses our FLUX.1 port with full-resolution references. Log-scale vertical axes.

4.4 Ablation Study

Cross-backbone evaluation. We apply the same fusion and training objective to Stable Diffusion v1.5 and FLUX.1 without changing the formulation. Table 2 reports both FusedVTON instantiations in the main comparison. Under SD1.5, FusedVTON is compared directly with CatVTON using the same backbone, while the FLUX.1 row shows that the same fusion formulation can also be instantiated with a different generative backbone.

Fusion strategy. Table 5 compares fusion space and garment placement. With region-fit placement, image-space fusion obtains paired LPIPS of 0.0583, compared with 0.071 for latent-space fusion. Random placement within image-space fusion gives 0.064, and the remaining metrics follow the same direction, except unpaired KID, where random placement is lower (1.08 vs. 1.178). These results support forming the joint conditioning image before VAE encoding and using structured placement.



Figure 7: **Qualitative marker-transport probe.** Each colored marker is inserted on the garment in the fused condition and generated in a separate run. Markers on the source garment show where each was inserted, markers on the generated result show where it was detected, and lines connect each inserted marker to its detected location.

Table 4: Component analysis on VITON-HD (FLUX.1). "+ Mask" adds an explicit mask channel, "w/o DP" removes DensePose, and "Ours" is the complete configuration. Best in bold.

Metric	+ Mask	w/o DP	Ours
Paired FID ↓	5.1184	5.2104	**5.0889**
Paired KID ↓	0.3361	0.3640	**0.1992**
Paired PSNR ↑	**21.5375**	20.9649	21.3663
Paired SSIM ↑	**0.8906**	0.8817	0.8889
Paired LPIPS ↓	**0.0462**	0.0491	0.0471
Paired DISTS ↓	**0.0750**	0.0760	0.0756
Unpaired FID ↓	8.4897	8.2281	**8.1532**
Unpaired KID ↓	0.7055	0.4067	**0.2205**

Table 5: Fusion-strategy ablation on VITON-HD (SD1.5). Ours uses image-space fusion with region-fit placement. Best results are in bold; second-best are underlined.

Metric	Latent space Region fit	Image space Random	Image space Region fit (Ours)
Paired FID ↓	<u>5.97</u>	6.08	**5.871**
Paired KID ↓	0.79	<u>0.63</u>	**0.455**
Paired LPIPS ↓	0.071	<u>0.064</u>	**0.0583**
Paired SSIM ↑	0.861	<u>0.875</u>	**0.8857**
Unpaired FID ↓	9.32	<u>9.21</u>	**9.071**
Unpaired KID ↓	1.34	**1.08**	<u>1.178</u>

Component analysis. Table 4 evaluates two component variants of FusedVTON with the FLUX.1 backbone on VITON-HD. Removing DensePose degrades all reported paired and unpaired metrics relative to the complete configuration. Adding an explicit mask channel improves paired PSNR, SSIM, LPIPS, and DISTS, but the complete configuration retains lower paired and unpaired FID and KID. Thus, DensePose provides a consistent benefit in this comparison, whereas the explicit mask channel produces a trade-off between reconstruction-oriented and distributional metrics.

5 Conclusion

We presented **FusedVTON**, a simple, efficient virtual try-on framework that fuses the garment and clothing-agnostic person into a single conditioning image before tokenization. This formulation treats virtual try-on as image editing and avoids explicit garment warping, separate garment branches, dedicated interaction modules, additional garment tokens, and an inpainting mask channel. Experiments show competitive quality with 53.9% and 51.7% lower profiled model-inference FLOPs than the SD1.5 and FLUX.1 CatVTON variants at 512×384, respectively. Counted inference FLOPs remain constant across the evaluated garment counts at fixed resolution. Our results suggest early image-space fusion as a practical direction for simpler, more efficient virtual try-on.

AI Use Statement

In this work, we used generative AI tools only for language assistance and manuscript polishing. We reviewed all AI-assisted revisions and take responsibility for the final content of this work, including its text, claims, and artifacts.

Ethics Statement

All datasets used in this work are publicly available, and we use them in accordance with their licensing terms. We collected no new data and used no personally identifiable or sensitive information beyond the images released with these datasets. Virtual try-on has many beneficial applications. We are committed to the responsible advancement of this field and encourage continued research toward clear ethical guidelines for its deployment.

REFERENCES

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul 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.

Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *European Conference on Computer Vision*, pp. 206–235. Springer, 2024.

Yisol Choi, Sangkyung Kwak, Sihyun Yu, Hyungwon Choi, and Jinwoo Shin. Controllable human image generation with personalized multi-garments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Zheng Chong, Xiao Dong, Haoxiang Li, Shiyue Zhang, Wenqing Zhang, Xujie Zhang, Hanqing Zhao, Dongmei Jiang, and Xiaodan Liang. Catvton: Concatenation is all you need for virtual try-on with diffusion models. *arXiv preprint arXiv:2407.15886*, 2024.

Zheng Chong, Yanwei Lei, Shiyue Zhang, Zhuandi He, Zhen Wang, Xujie Zhang, Xiao Dong, Yiling Wu, Dongmei Jiang, and Xiaodan Liang. Fastfit: Accelerating multi-reference virtual try-on via cacheable diffusion models. *arXiv preprint arXiv:2508.20586*, 2025.

Phuong Dam, Jihoon Jeong, Anh Tran, and Daeyoung Kim. Time-efficient and identity-consistent virtual try-on using a variant of altered diffusion models. In *European Conference on Computer Vision*, pp. 35–51. Springer, 2024.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022.

Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8485–8493, 2021.

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 *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 7599–7607, 2023.

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7297–7306, 2018.

Hailong Guo, Bohan Zeng, Yiren Song, Wentao Zhang, Jiaming Liu, and Chuang Zhang. Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19085–19096, 2025.

Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7543–7552, 2018.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Junyao Hu, Zhongwei Cheng, Waikeung Wong, and Xingxing Zou. Garments2look: A multi-reference dataset for high-fidelity outfit-level virtual try-on with clothing and accessories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pp. 206–228. Springer, 2024.

Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8176–8185, 2024.

Jeongho Kim, Hoiyeong Jin, Sunghyun Park, and Jaegul Choo. Promptdresser: Improving the quality and controllability of virtual try-on via generative textual prompt and prompt-aware mask. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16026–16036, 2025.

Seungyong Lee and Jeong-gi Kwak. Voost: A unified and scalable diffusion transformer for bidirectional virtual try-on and try-off. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

Yuhan Li, Hao Zhou, Wenxiang Shang, Ran Lin, Xuanhong Chen, and Bingbing Ni. AnyFit: Controllable virtual try-on for any combination of attire across any scenario. *arXiv preprint arXiv:2405.18172*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress Code: High-Resolution Multi-Category Virtual Try-On. In *Proceedings of the European Conference on Computer Vision*, 2022.

Soonchan Park and Jinah Park. WG-VITON: Wearing-guide virtual try-on for top and bottom clothes. *arXiv preprint arXiv:2205.04759*, 2022.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11410–11420, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pp. 4596–4604, 2018.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 589–604, 2018.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Jianhao Zeng, Dan Song, Weizhi Nie, Hongshuo Tian, Tongtong Wang, and An-An Liu. Cat-dm: Controllable accelerated virtual try-on with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8372–8382, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Xujie Zhang, Ente Lin, Xiu Li, Yuxuan Luo, Michael Kampffmeyer, Xin Dong, and Xiaodan Liang. Mmtryon: Multi-modal multi-reference control for high-quality fashion generation. *arXiv preprint arXiv:2405.00448*, 2024.

Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI Artificial Intelligence Research*, 3:9150038, 2024.

Zijian Zhou, Shikun Liu, Xiao Han, Haozhe Liu, Kam Woh Ng, Tian Xie, Yuren Cong, Hang Li, Mengmeng Xu, Juan-Manuel Pérez-Rúa, et al. Learning flow fields in attention for controllable person image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2491–2501, 2025.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4606–4615, 2023.

A Data and Preprocessing

Datasets and splits. VITON-HD (Choi et al., 2021) contains 13,679 frontal-view person–garment pairs of upper-body garments, split into 11,647 training and 2,032 test pairs. DressCode (Morelli et al., 2022) contains 53,792 full-body pairs across upper-body garments, lower-body garments, and dresses, split into 48,392 training and 5,400 test pairs. Both have native resolution 1024×768. DressCode-MR (Chong et al., 2025) contains 28,179 outfit-level samples, with 25,779 for training and 2,400 for testing. For Garments2Look (Hu et al., 2026), we use 30,222 processed outfit-level samples, with 28,711 for training and 1,511 for testing. Both multi-garment datasets pair each person with multiple reference items. Our DressCode-MR evaluation excludes bags; Garments2Look uses tops, bottoms, dresses, shoes, and bags. We follow the standard paired and unpaired protocols for VITON-HD and DressCode and evaluate joint transfer of multiple references for DressCode-MR and Garments2Look. All images are resized to 512×384 in our experiments.

Human parsing and DensePose. All anatomy-dependent operations use Sapiens (Khirodkar et al., 2024) human parsing with its 29-class label set. We use its face-and-neck label as the head, its two hand labels as the hands, and its foot, shoe, and sock labels as the feet. The pose condition P is the DensePose (Güler et al., 2018) rendering of the target person, resized to 512×384 together with the person image. Target parses, pose keypoints, and available clothing-agnostic masks are used to construct the fused images; DensePose is supplied separately as latent pose conditioning.

Clothing-agnostic masks. The combined mask M^{a} in Eq. (2) is the clothing-agnostic mask of the target person. For VITON-HD, we take the union of the dataset's clothing-agnostic mask and the Sapiens upper-clothing region, so that the mask covers sleeves, hems, and collars that the original mask can leave exposed. For DressCode, we use clothing-agnostic masks derived from the Sapiens parse for each of the three categories. Pixels inside M^{a} are set to the constant gray value $g = 127/255$ before the garments are composited.

Construction of A_i. The placement region A_i of garment i is the clothing-agnostic mask with the head and hands removed; for lower-body and full-body garments, the feet are removed as well. Removing these parts prevents a garment from being placed over anatomy that must be preserved. Hands and other small holes enclosed by the region, each at most 2% of the region's area, can split an otherwise wide region and force a much smaller box. We therefore admit them when fitting the box and remove them from the placed garment afterward. The placed garment consequently never covers a hand and never extends outside A_i.

Box fitting. Before placement, each in-shop garment image and its foreground mask are resized to the 512×384 canvas, and the garment is cropped to the tight bounding box of its mask. Given a crop of size $h \times w$, we binary-search the scale s over 14 iterations. For each candidate s, a box of size $\mathrm{round}(sh) \times \mathrm{round}(sw)$ fits if A_i, eroded by a structuring element of that size, is non-empty. Among the feasible top-left corners at the largest feasible scale, we select the one whose box center is closest to the centroid of A_i. Background pixels of the crop are zeroed before resizing, so no background color leaks into the garment boundary.

Multiple garments. When the boxes of two garments overlap, the vertical extent they share is split at a row between the two garments' own parse regions on the target person, or at the midpoint of the shared extent when either region is unavailable. The upper garment keeps the part of its region above the split row and the lower garment keeps the part below it; each garment always retains its own parse region. Both boxes are then refitted. We repeat this procedure for up to four passes.

Multi-garment training data. During multi-garment fine-tuning, each training sample contains at least two reference items. Each item is dropped independently with probability 0.2, and at least one item is always retained. For Garments2Look, the reference items come from five categories: tops, bottoms, dresses, shoes, and bags (the upper, lower, overall, shoes, and bags slots in the data). Garment foreground masks are estimated from the product images with BiRefNet (Zheng et al., 2024). We exclude items whose annotations are flagged for manual review. After removing looks with fewer than two usable items, 24,941 of the 28,711 training looks remain.

Garments2Look test protocol. Garments2Look does not provide an official split. We therefore construct a product-disjoint split from the connected components of the bipartite graph between looks and product identifiers. Components enter the test set only while a 5% budget permits, so no product appears in both training and test data; this yields the $1{,}511$ test looks. We evaluate the $1{,}319$ of these looks that contain two to four usable reference items, counting bags. 247 of these looks include a bag. In the paired setting, each wearer is conditioned on the product images of their own look, and the corresponding photograph serves as ground truth. In the unpaired setting, the $1{,}319$ looks are shuffled with seed 42, and each wearer receives the next look in the resulting cycle as a donor. If the donor has no product that differs from the wearer's in a shared slot, we advance through the cycle to the first donor that does. Only shared slots with a different product are donated. The wearer keeps their own image, clothing-agnostic mask, DensePose, and placement regions. This procedure uses $1{,}242$ distinct donor looks. The wearer–donor assignment is fixed and shared by all methods.

B Training Configuration

Input projections. For FLUX.1-Fill, the input projection of FLUX.1-Fill-dev maps 384 input channels to the model width: 64 channels for the packed noisy latent, 64 for the packed masked-image latent, and 256 for the packed mask. We retain the first 128 columns, which serve the noisy latent and the fused-image latent z_{f}, and discard the mask columns. We then append 64 zero-initialized columns for the packed DensePose latent, resulting in 192 input channels. At 512×384, the 16-channel VAE latent of size 64×48 is packed into 2×2 patches, giving 768 image tokens. The context embedder and the text-dependent layers in every block are removed, so no text is encoded, and all remaining transformer parameters are trained. For Stable Diffusion, the inpainting UNet's 9-channel input convolution is rebuilt with 11 channels: 4 for the noisy latent, 4 for the fused-image latent (initialized from the masked-image weights), and 3 zero-initialized channels for the DensePose image. The mask channel is dropped. Cross-attention is replaced by an identity operation, the VAE is frozen, and the full UNet is trained.

Optimization. Both single-garment variants are trained for $80{,}000$ steps on the combined VITON-HD and DressCode training sets. Table 6 lists the training hyperparameters. Adafactor is used without relative step sizes or parameter scaling, with an update clipping threshold of 1.0; its update is applied to each parameter as soon as that parameter's gradient is available during the backward pass. Target and condition latents are taken from the mode of the VAE posterior. For FLUX, training timesteps are drawn uniformly from the scheduler's $1{,}000$ training timesteps, and the guidance embedding is fixed to 1.0. Both multi-garment models are fine-tuned for $30{,}000$ steps from an $80{,}000$-step single-garment FusedVTON (FLUX.1) checkpoint.

Table 6: Training hyperparameters. "MR" denotes multi-garment fine-tuning.

	FusedVTON (FLUX.1)	FusedVTON (SD1.5)	MR: DressCode-MR	MR: Garments2Look
Initialization	FLUX.1-Fill-dev	SD1.5 inpainting	FusedVTON (FLUX.1), 80k steps	
Training data	VITON-HD + DressCode	VITON-HD + DressCode	DressCode-MR	Garments2Look
Resolution	512×384	512×384	512×384	512×384
Objective	flow matching	ϵ-prediction	flow matching	flow matching
Batch size	16	128	16	16
Gradient accumulation	1	1	1	1
Optimizer	Adafactor (Shazeer & Stern, 2018)	AdamW (Loshchilov & Hutter, 2019)	Adafactor	Adafactor
Learning rate	10^{-5}	10^{-5}	10^{-5}	10^{-5}
Schedule	constant, 500 warmup	constant, no warmup	constant, 500 warmup	constant, 500 warmup
Weight decay	0.01	0.01	0.01	0.01
AdamW $(\beta_1, \beta_2, \epsilon)$	–	$(0.9,\ 0.999,\ 10^{-8})$	–	–
Gradient clipping	1.0	1.0	1.0	1.0
Precision	bf16	bf16	bf16	bf16
Image-condition dropout	0.1	0.1	0.1	0.1
Per-garment dropout	–	–	0.2	0.2
Fine-tuning steps	–	–	30,000	30,000
Seed	42	42	42	42

C Inference and Evaluation Protocol

Sampling. On VITON-HD and DressCode, the FLUX.1 model uses 28 Euler flow-matching steps with a static timestep shift of 3.0, image guidance $s_{\text{img}} = 2.5$ (Eq. (6)), a guidance embedding of 1.0, and seed 42. The SD1.5 model uses 50 DDIM (Song et al., 2021) steps with image guidance 1.5 and seed 42. All outputs are generated at 512×384.

Test sets. VITON-HD is evaluated on its 2,032 test pairs and DressCode on its 5,400 test pairs, 1,800 per category, which are pooled into a single score. The unpaired setting uses each dataset's standard unpaired test list.

Ground truth and resizing. Ground-truth images are resized from their original resolution to 512×384 with Lanczos interpolation and stored losslessly as PNG. Outputs from methods that generate at a different resolution are resized to 512×384 with the same filter and also stored as PNG, so every method is scored against identical references at identical resolution.

Metrics. FID and KID are computed with clean-fid (Parmar et al., 2022). KID uses clean-fid's default estimator, which averages the unbiased kernel MMD over 100 random subsets of at most 1,000 images; all KID values in this paper are multiplied by 10^3. PSNR and SSIM are computed on images scaled to $[0, 1]$. LPIPS (Zhang et al., 2018) uses the SqueezeNet backbone, and DISTS (Ding et al., 2022) uses its reference implementation. Paired metrics are computed between each output and its own ground truth. For unpaired outputs, where no aligned reference exists, we report only FID and KID against the real test images.

Efficiency. FLOPs are counted with PyTorch's `FlopCounterMode` on the GPU, over the full model-inference span: VAE encoding, any required reference encoders or learned warpers, cache construction, the denoising loop, and VAE decoding. Counting on the GPU ensures that the attention kernels are included. All methods run without classifier-free guidance, with 50 steps for Stable Diffusion-based methods and 28 steps for FLUX.1-based methods. For latency, the GPU is synchronized before and after the same span, which is timed with a wall-clock timer. Measurements use an idle NVIDIA B200 in bfloat16 at batch size one. Image loading, saving, model loading, and external preprocessing, including parsing, pose and mask estimation, and garment crop, resize, and paste, are excluded from both FLOPs and latency.

D Limitations and Discussion

Because FusedVTON places all garments within one fixed canvas, the resolution available to each garment is bounded by the clothing-agnostic region. This keeps computation constant as garments are added, but it can limit the detail retained for small items or densely packed outfits. The fused image is also built from human parsing, DensePose, and clothing-agnostic masks, so its placement depends on the quality of these inputs. Our evaluation follows the standard benchmark protocols at 512×384 with in-shop garment images, and extending it to higher resolutions and in-the-wild photographs is a natural next step. The fixed canvas also suggests simple extensions, such as adaptive allocation of canvas area among garments or a few additional tokens for detail-heavy items, that keep the single-grid design. Finally, early fusion reduces the number of tokens per denoiser evaluation, so it can be combined with methods that reduce sampling steps or numerical precision. Its relative savings are expected to grow at higher resolutions, where quadratic attention accounts for a larger share of computation (Table 1).

E Additional Qualitative Results

Figures 8–9 show eight additional unpaired VITON-HD examples, and Figures 10–11 show eight additional DressCode examples spanning upper-body garments, lower-body garments, and dresses. Each row places the target garment above the source person in the narrow reference column, followed by the method outputs. These examples complement the main comparisons with additional lettering, prints, garment shapes, and poses.

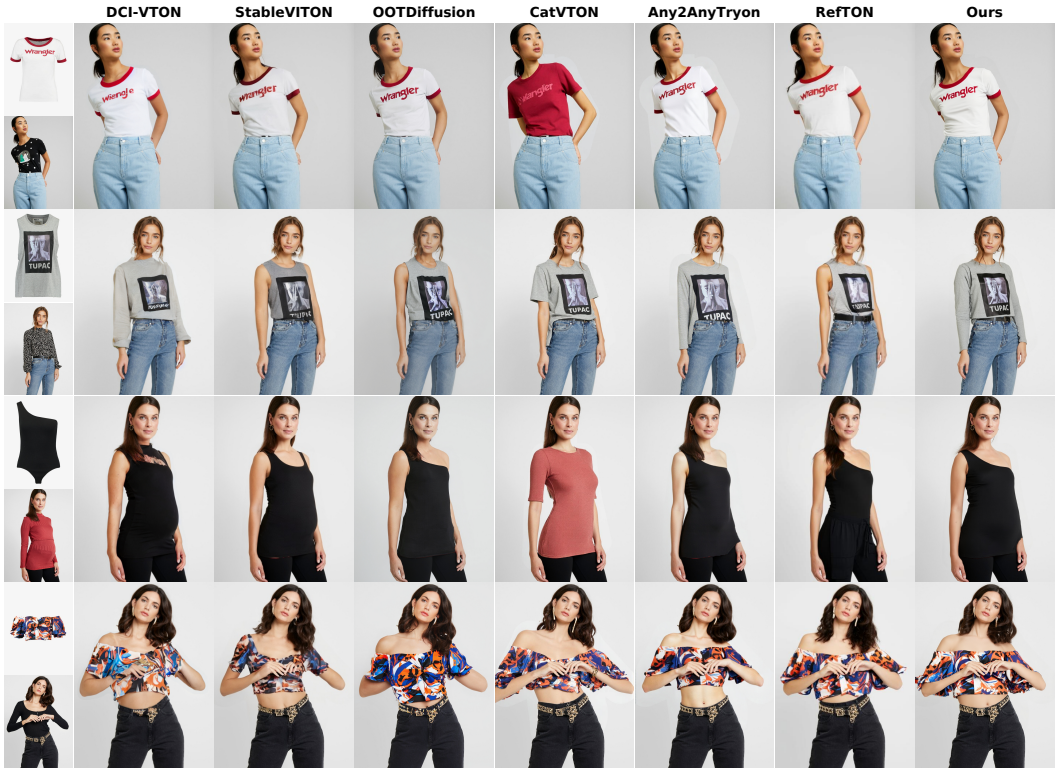

Figure 8: Additional unpaired try-on examples on VITON-HD (1/2). Columns after the references show DCI-VTON, StableVITON, OOTDiffusion, CatVTON, Any2AnyTryon, RefTON, and Fused-VTON (Ours). Our results use the 1024×768 outputs.

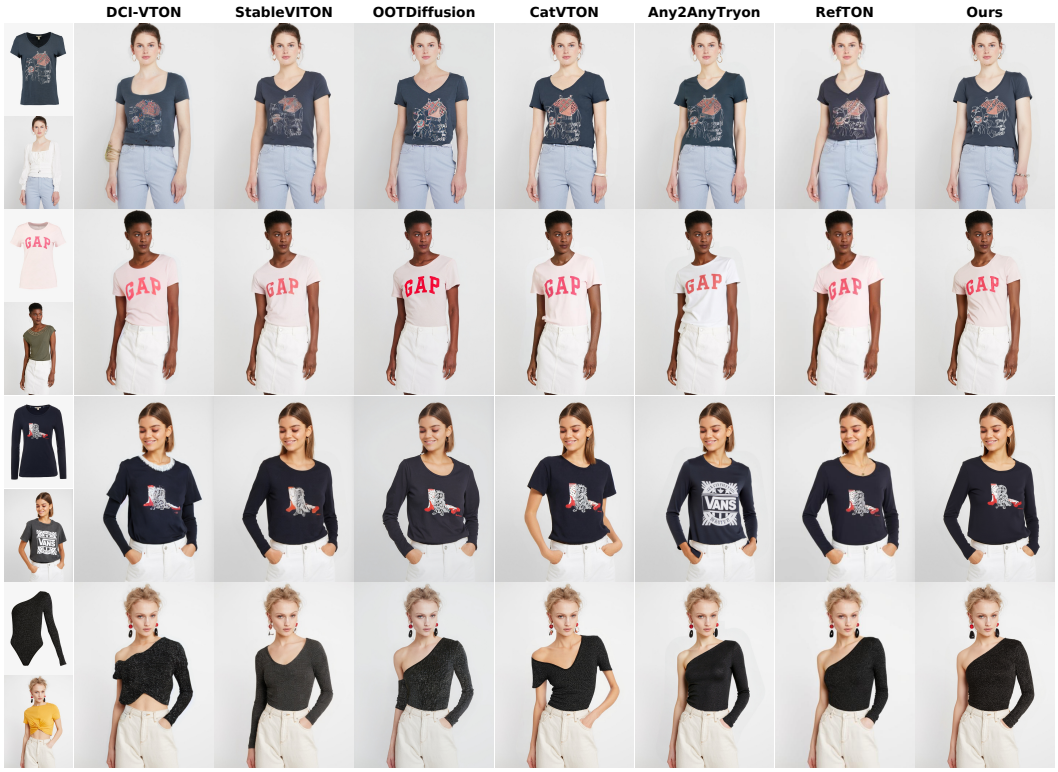

Figure 9: Additional unpaired try-on examples on VITON-HD (2/2). The reference and method order matches Figure 8. Our results use the 1024×768 outputs.



Figure 10: Additional unpaired try-on examples on DressCode (1/2). Columns after the references show LaDI-VTON, OOTDiffusion, CatVTON, RefTON, and FusedVTON (Ours). Our results use the 1024×768 outputs.



Figure 11: Additional unpaired try-on examples on DressCode (2/2). The reference and method order matches Figure 10. Our results use the 1024×768 outputs, except the third row, which uses the selected 512×384 output.

Figure 12 shows six additional DressCode-MR examples that do not appear in the main comparison. Each combines upper-body, lower-body, and shoe references.

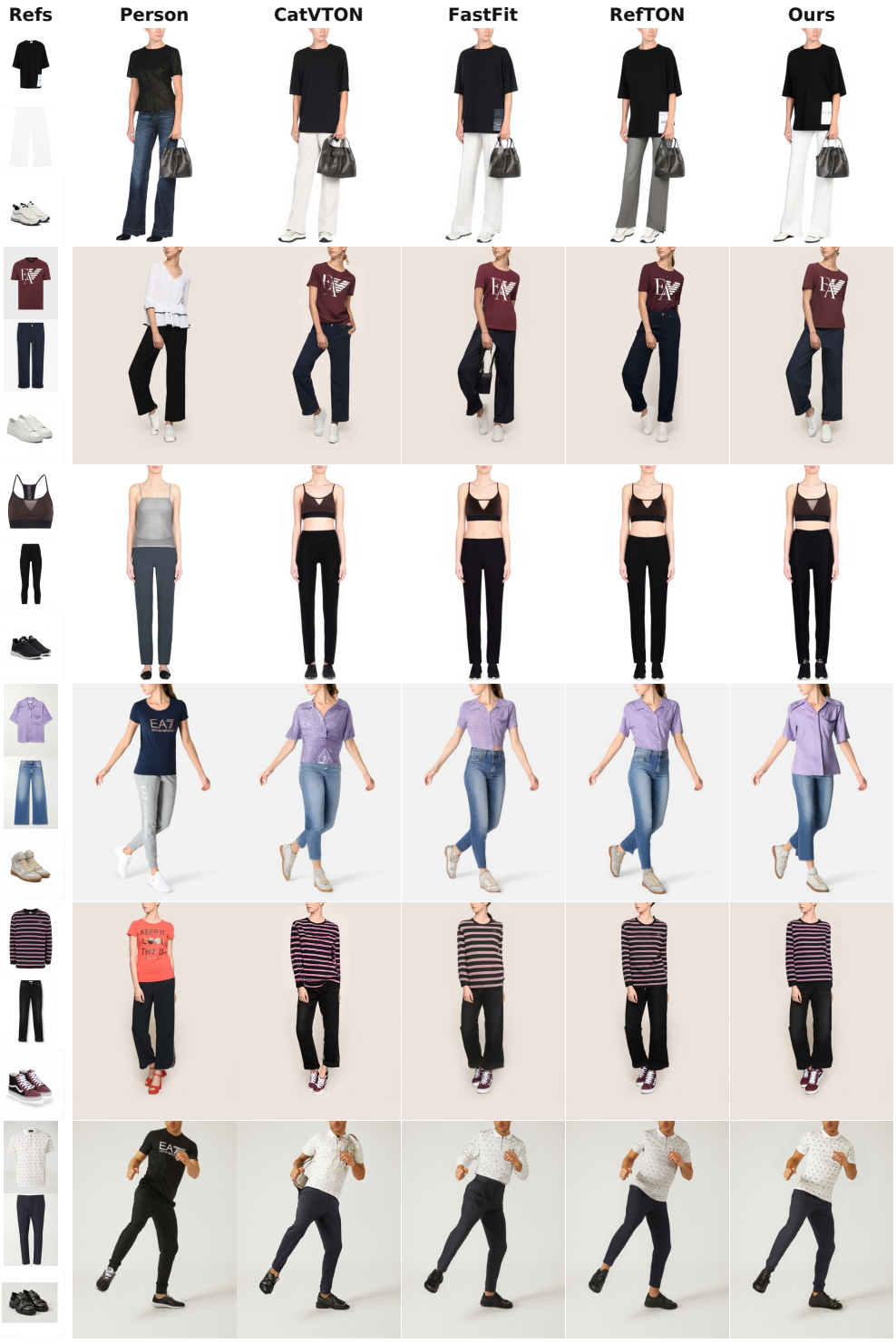

Figure 12: Additional multi-garment try-on examples on DressCode-MR. Each row shows the reference items, source person, and outputs from CatVTON, FastFit, RefTON, and FusedVTON (Ours).