ITA-MDT: Image-Timestep-Adaptive Masked Diffusion Transformer Framework for Image-Based Virtual Try-On

Supplementary Material

6. Masked Diffusion Transformer for Image-Based Virtual Try-On (MDT-IVTON)

This section gives further explanation of Sec. 3.1.

6.1. Patchified Latent Formulation

The Garment Agnostic Map, DensePose, and Garment Agnostic Mask images have initial shapes of $A, P, M_X \in \mathbb{R}^{3,h,w}$, where h and w are the height and width, and 3 corresponds to the RGB channels. This image is encoded into the latent space by a VAE encoder, transforming them into $\mathcal{E}(A), \mathcal{E}(P), \mathcal{E}(M_X) \in \mathbb{R}^{4,H,W}$, where $H = h/8$ and $W = w/8$. These are concatenated with the noised latent representation $z_t \in \mathbb{R}^{4,H,W}$, resulting in a combined tensor $\mathbb{R}^{16,H,W}$.

This combined tensor is patchified into a representation $L \in \mathbb{R}^{p,D}$, with patch $p = \frac{H \cdot W}{\text{patch size}^2}$ and D is the hidden layer embedding dimension. Positional embeddings $\in \mathbb{R}^{p,D}$ are added to L, forming the final patchified latent representation. This representation undergoes denoising in the encoder and decoder blocks of MDT-IVTON. Within these blocks, L serves as the query in the cross-attention mechanism, interacting with the condition c to incorporate garment-specific information.

6.2. Condition Formulation

The garment image is processed through the Salient Region Extractor (SRE) and Image-Timestep Adaptive Feature Aggregator (ITAFA) to obtain the garment feature F_g and the salient region feature F_s. These features have shapes $F_g, F_s \in \mathbb{R}^{s,d}$, where s is the sequence length of the patch tokens from the image encoder and d is the embedding dimension of the image encoder.

Both F_g and F_s are projected to align with MDT-IVTON's embedding dimension D, resulting in $\mathbb{R}^{s,D}$. These are then concatenated along the sequence dimension to form $\mathbb{R}^{2s,D}$. Time embedding $T \in \mathbb{R}^D$ is added to all sequences, formulating the final condition c, which acts as the key and value in the cross-attention mechanism to guide denoising in MDT-IVTON.

6.3. Denoising Objective

The primary objective function minimizes the mean squared error (MSE) between the predicted noise and the actual noise in the noised latent z_t at each timestep t, following the standard diffusion objective:

$$L_{\text{denoise}} = \mathcal{E}_{z_t, c, \epsilon \sim \mathcal{N}(0,I), t} \left[\|\epsilon - \epsilon_\theta(z_t, c, t)\|^2 \right], \quad (13)$$

where $z_t \in \mathbb{R}^{4,H,W}$ is the noised latent representation at timestep t, $c \in \mathbb{R}^{2s,D}$ is the condition, $\epsilon \in \mathbb{R}^{4,H,W}$ is the Gaussian noise added during the forward diffusion process, and $\epsilon_\theta(z_t, c, t) \in \mathbb{R}^{4,H,W}$ is the model's predicted noise. This objective guides the model to learn to reverse the forward diffusion process.

6.4. Mask Reconstruction Objective

The Mask Reconstruction Objective operates on the masked latent representation L_m, derived by applying a binary mask $M_L \in \mathbb{R}^p$ to the patchified latent L. The mask M_L indicates masked tokens with 0 and unmasked tokens with 1.

The Side-Interpolator reconstructs the masked tokens in L_m by leveraging the semantic information from the unmasked tokens as follows:

1. The unmasked tokens $L_u \in \mathbb{R}^{p_u,D}$, where p_u is the number of unmasked tokens, provide semantic context.
2. The masked tokens $L_m \in \mathbb{R}^{p_m,D}$, where $p_m = p - p_u$, are reconstructed by interacting with L_u in the Side-Interpolator.
3. The reconstruction leverages an attention mechanism:
   - L_m serves as the query, representing the masked tokens requiring reconstruction.
   - L_u serves as the key and value, encoding semantic and spatial context from the unmasked tokens.
   - Attention weights computed between L_m (query) and L_u (key) determine how information from L_u (value) is used to reconstruct L_m.
4. The output $L'_m \in \mathbb{R}^{p_m,D}$ replaces the masked tokens in L_m, forming a refined latent representation.

The reconstruction loss is computed as:

$$L_{\text{mask}} = \mathcal{E}_{y,c,t} \left[\|L'_m - L_m\|^2 \right],$$

ensuring spatial coherence and semantic consistency in masked regions.

6.5. Inpainting Objective

The inpainting loss focuses on regions defined by the Garment Agnostic Mask M_X. In the latent space, M_X is encoded as $\mathcal{E}(M_X) \in \mathbb{R}^{4,H,W}$. Among its four channels, the first channel, $\mathcal{E}(M_X)_0 \in \mathbb{R}^{H,W}$, closely resembles the binary mask and is used for loss computation, which is visualized in Fig. 7

This channel is normalized to ensure numerical stability, and the inpainting loss is calculated as:

$$L_{\text{inpaint}} = \mathcal{E}_{y,c,t} \left[\|\mathcal{E}(M_X)_0 \cdot (y_t - y_{t-1})\|^2 \right],$$

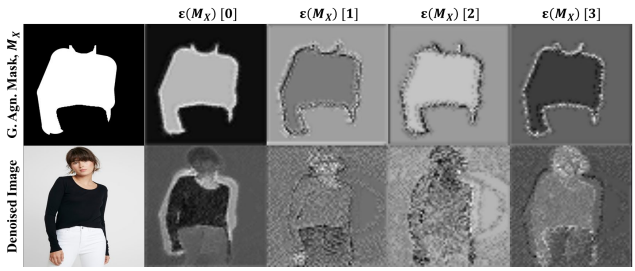

Figure 7. The top row visualizes the four normalized channels of the Garment Agnostic Mask in the latent space. The bottom row visualizes the corresponding four channels of the latent image, each with the mask from the matching channel applied.

where $y_t, y_{t-1} \in \mathbb{R}^{4,H,W}$ are latent representations at consecutive timesteps. This loss emphasizes garment-specific details in masked regions.

6.6. Overall Loss Function

The overall loss integrates denoising, mask reconstruction, and inpainting objectives:

$$L_{\text{total}} = L_{\text{denoise}} + L_{\text{mask}} + L_{\text{inpaint}}.$$

This formulation balances global structure, spatial coherence, and fidelity of garment-specific details, ensuring high-quality virtual try-on results.

7. Image-Time Adaptive Feature Aggregator (ITAFA)

This section elaborates on the ITAFA of Sec. 3.3 in detail.

Feature Complexity Components. Complexity components for the input feature tensor f are calculated as below.

7.1. Feature Sparsity

S quantifies the proportion of near-zero activations in the feature embeddings, providing insight into the sparsity of structural activations within f. Given a threshold δ, sparsity is defined as:

$$S = \frac{1}{H \times s \times d} \sum_{i=1}^{H} \sum_{j=1}^{s} \sum_{k=1}^{d} \mathbb{I}(|f_{ijk}| < \delta), \qquad (14)$$

where H is the number of hidden layers of the image encoder, s is the sequence length of the patch tokens, d is the embedding dimension, and $\mathbb{I}$ is the indicator function that equals 1 when $|f_{ijk}| < \delta$ and 0 otherwise. This function allows for counting the proportion of elements in f that are below the threshold (i.e., near zero) providing a measure of sparsity.

7.2. Feature Variance

V reflects the variability across activations, capturing structural complexity and richness of detail in f:

$$V = \frac{1}{H \times s \times d} \sum_{i=1}^{H} \sum_{j=1}^{s} \sum_{k=1}^{d} (f_{ijk} - \overline{f})^2 \qquad (15)$$

where $\overline{f}$ is the mean activation across all embeddings.

7.3. Gradient Magnitude

G measures local variations in feature embeddings by computing spatial gradients along the sequence and embedding dimensions. This component captures texture and fine details, calculated as:

$$\Delta f_{i,j,k} = \sqrt{(f_{i,j+1,k} - f_{i,j,k})^2 + (f_{i,j,k+1} - f_{i,j,k})^2} \qquad (16)$$

$$G = \frac{1}{H \times (s-1) \times (d-1)} \sum_{i=1}^{H} \sum_{j=1}^{s-1} \sum_{k=1}^{d-1} \Delta f_{i,j,k} \qquad (17)$$

$\Delta f_{i,j,k}$ represents the gradient magnitude at each index (i, j, k) in the feature embedding tensor. G then averages the Δf values over the entire feature embedding tensor to capture the overall texture complexity.

The combined complexity score vector, $[S, V, G] \in \mathbb{R}^3$, captures the structural and textural complexity of f.

8. Salient Region Extractor (SRE)

This section describes the SRE algorithm in detail, elaborating on Sec. 3.3.

8.1. Entropy Map Computation

The entropy map X_e provides a measure of the information content for each pixel in the grayscaled version of the input garment image, $\mathcal{X}_{\text{gray}}$. The Shannon Entropy [33] is utilized to capture local texture complexity and information density.

Local Neighborhood Definition. For each pixel $\mathcal{X}_{\text{gray}}(i, j)$, we consider a local neighborhood of size 5×5, denoted as $N_{i,j}$, centered at pixel (i, j):

$$N_{i,j} = \{\mathcal{X}_{\text{gray}}(m, n) \mid i{-}2 \leq m \leq i{+}2,\ j{-}2 \leq n \leq j{+}2\}, \qquad (18)$$

where the neighborhood is truncated near the image borders to fit within the image dimensions.

Probability Distribution of Intensities. In each neighborhood $N_{i,j}$, we count the frequency of each pixel intensity

value from 0 to 255, and calculate the probability p_k of each intensity value k as:

$$p_k = \frac{1}{|N_{i,j}|} \sum_{(m,n)\in N_{i,j}} \mathbb{I}(\mathcal{X}_{\text{gray}}(m,n) = k), \quad (19)$$

where $|N_{i,j}| = 25$ is the number of pixels in the neighborhood, and $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if the condition is true and 0 otherwise.

Shannon Entropy Calculation. The Shannon entropy $H_{i,j}$ for the neighborhood $N_{i,j}$ is calculated as:

$$H_{i,j} = -\sum_{k=0}^{255} p_k \log_2(p_k), \quad (20)$$

where p_k represents the probability of intensity k within the neighborhood. If $p_k = 0$, the corresponding term is considered zero, as $p_k \log_2(p_k) = 0$ for $p_k = 0$.

Constructing the Entropy Map. The entropy map X_e is constructed by assigning the computed entropy value $H_{i,j}$ to each pixel (i, j) in the image:

$$X_e(i,j) = H_{i,j}, \quad (21)$$

resulting in an entropy map $X_e \in \mathbb{R}^{H\times W}$ that provides a grayscale representation of the information content for each pixel.

Interpretation of the Entropy Map. The resulting entropy map X_e reflects the complexity of each region in the image:

- **High Entropy**: Regions with higher entropy indicate greater variability in pixel intensities, suggesting areas with rich textures, edges, or details.
- **Low Entropy**: Regions with lower entropy represent uniform areas with little variation.

This process effectively highlights the most informative areas of the input image.

8.2. High-Entropy Region Identification

To isolate regions of interest, a binary mask X_m is generated by thresholding the entropy map with a pre-defined entropy threshold E:

$$X_{\text{m}}(i,j) = \begin{cases} 1 & \text{if } X_e(i,j) > E, \\ 0 & \text{otherwise.} \end{cases} \quad (22)$$

The entropy threshold E is empirically set to 0.8. If no high-entropy regions are detected, adaptive thresholding is applied, gradually lowering the threshold until a region is found or reaches a minimum value. If this adjustment fails.

8.3. Entropy Centroid Localization

The centroid (x_c, y_c) of the high-entropy region is computed as the center of mass:

$$(x_c, y_c) = \frac{\sum_{i,j} X_m(i,j) \cdot (i,j)}{\sum_{i,j} X_m(i,j)}. \quad (23)$$

If no high-entropy regions are found, the fallback behavior sets the centroid to the image center, ensuring robustness in cases of low entropy.

8.4. Circular Region Expansion

The region around the centroid is initially bounded by a square of width and height l_{min}, set to 224, centered at (x_c, y_c). l_{min} defines the minimum height and width of the Salient Region to prevent overly small regions and extreme aspect ratios. This bounding box expands outward in a circular pattern (i.e., up, right, down, left), repeating this sequence until no further expansion is needed. For each direction, the algorithm checks whether the newly added edge pixels contain more high-entropy pixels than a given threshold to determine if expansion should continue.

8.5. Region Extraction

After the Circular Region Expansion, the bounding box is adjusted to match the aspect ratio of the original image to minimize distortion. The adjustment involves expanding either the height or width, depending on the current bounding box's aspect ratio compared to the original $\mathcal{X}_{\text{gray}}$. The final Salient Region X_s extracted maintains the aspect ratio, preserving visual consistency. By preserving the original aspect ratio, the model can effectively perceive the salient region in the context of the full garment, minimizing potential spatial confusion. The extracted region is then resized to 224×224 to be processed as a diffusion condition.

9. Denoising with Classifier-Free Guidance

The denoising process employs Classifier-Free Guidance (CFG) [12] to dynamically balance unconditional and conditional noise predictions during each timestep t of the diffusion process.

9.1. DDIM Sampling

The iterative denoising process is implemented using a DDIM sampling loop, which refines the noisy latent representation $z_t \in \mathbb{R}^{4,H,W}$ over a predefined number of diffusion steps γ_{steps}. At each timestep, the model predicts the noise ϵ_t to compute the latent representation for the next timestep z_{t-1}:

$$z_{t-1} = \text{DDIM}(\epsilon_t, z_t, t, \gamma_{\text{steps}}), \quad (24)$$

where the process continues until z_0, clean latent representation. The sampling loop iteratively combines conditional and

unconditional noise predictions using the guidance mechanism described below.

9.2. Unconditional and Conditional Predictions

At each timestep t, the noisy latent representation z_t is processed by the denoising model ϵ_θ to produce two noise predictions:

$$\epsilon_{\text{uncond}} = \epsilon_\theta(z_t, t, \emptyset), \tag{25}$$

$$\epsilon_{\text{cond}} = \epsilon_\theta(z_t, t, c), \tag{26}$$

where $\epsilon_{\text{uncond}} \in \mathbb{R}^{4,H,W}$ is the unconditional noise prediction generated without the condition c, and $\epsilon_{\text{cond}} \in \mathbb{R}^{4,H,W}$ is the noise prediction with condition c informed.

9.3. Dynamic CFG Scaling

The classifier-free guidance scale α_{cfg} determines the strength of the conditional guidance, which is dynamically adjusted at each timestep using a cosine-based power scaling function:

$$\delta_{\text{scale}} = \frac{1 - \cos\left(\left(1 - \frac{t}{\gamma_{\text{steps}}}\right)^{\beta_{\text{scale}}} \cdot \pi\right)}{2}, \tag{27}$$

where γ_{steps} is the total number of diffusion steps, and β_{scale} is the power scaling factor. The intermediate scale factor $\delta_{\text{scale}} \in [0, 1]$ ensures a smooth adjustment of guidance strength, starting with weaker conditional emphasis in earlier steps and gradually increasing its influence.

The effective guidance scale α_{eff} for timestep t is computed as:

$$\alpha_{\text{eff}} = 1 + (\alpha_{\text{cfg}} - 1) \cdot \delta_{\text{scale}}. \tag{28}$$

9.4. Guided Noise Prediction

The final noise prediction ϵ_t is a weighted combination of the unconditional and conditional predictions:

$$\epsilon_t = \epsilon_{\text{uncond}} + \alpha_{\text{eff}} \cdot (\epsilon_{\text{cond}} - \epsilon_{\text{uncond}}), \tag{29}$$

where $\epsilon_t \in \mathbb{R}^{4,H,W}$.

10. Implementation Details

This section provides further details on the implementation of the proposed model, as outlined in Sec. 4. The foundational architecture for MDT-IVTON is derived from the MDTv2 XL model, configured with a depth of 28 layers, including 4 decoding layers, a hidden size of 1152, and 16 attention heads per layer. We utilize RGB images of size $512 \times 512 \times 3$ as both the input reference images and the generated output results. To ensure a fair comparison with prior works, the Variational Autoencoder (VAE) from Stable Diffusion XL [26] is employed, encoding images into a latent representation z with dimensions $64 \times 64 \times 4$.

During training, we use 1000 diffusion steps, whereas 30 steps are used for the generation results reported in Table 1. The mask ratio in the MDT training scheme is set to 0.3. For optimization, we use an initial learning rate of 1×10^{-4} with a batch size of 6. Training stability is enhanced through the use of an Exponential Moving Average (EMA) with a rate of 0.9999. All other hyperparameters, including the optimizer and learning rate scheduler, follow the configurations used in DiT [24], ensuring consistency with existing diffusion-based transformer approaches. For inference, the model is evaluated using 30 sampling steps γ_{steps}. The Classifier-Free Guidance scale α_{cfg} is set to 2.0, and the power scaling factor β_{scale} is set to 1.0. We report the performance of the model trained for 2 million steps.

11. Dataset and Evaluation Metrics

This section gives details on datasets used and the evaluation metrics, mentioned in Sec. 4. The VITON-HD dataset comprises 13,679 images of human models paired with upper garment images, with person images provided at a resolution of 1024×768. This dataset is widely used for virtual try-on tasks and features relatively simple poses, where subjects stand in straightforward, static positions with solid-colored backgrounds. The simplicity of backgrounds and poses makes VITON-HD an ideal testbed for evaluating the model's performance in basic virtual try-on scenarios.

The DressCode dataset offers a more diverse collection, containing over 50,000 high-resolution images (1024×768) across three categories: upper-body garments, lower-body garments, and dresses. Specifically, it includes 17,650 images of upper-body garments, 17,650 images of lower-body garments, and 17,650 images of dresses. Similar to VITON-HD, DressCode features consistent, simpler poses against plain backgrounds. However, it includes a more diverse set of garment styles, offering additional challenges in fitting and transferring intricate patterns, logos, and textures accurately during virtual try-on tasks. Both datasets serve as benchmarks, with VITON-HD focusing on basic pose and background handling, and DressCode testing the model's ability to preserve detailed garment features across various clothing types.

To evaluate performance, we employ several widely-used metrics in virtual try-on research. LPIPS (Learned Perceptual Image Patch Similarity) measures perceptual similarity by comparing deep features from neural networks, with lower LPIPS scores indicating greater perceptual closeness to ground truth. SSIM (Structural Similarity Index) evaluates the structural integrity of generated images by quantifying similarity in luminance, contrast, and structure; higher SSIM values indicate better preservation of the original structure. FID (Fréchet Inception Distance) assesses quality and diversity by comparing the feature distributions of generated and real images, with lower FID values denoting closer align-

ment to real image distributions. We report both paired and unpaired FID results. While FID is commonly used to assess unpaired results in VTON task, paired FID is also informative as it directly compares generated images with their corresponding ground-truth images, which does not exist for unpaired generation.

12. Ablation Study

12.1. Analysis on ITAFA

For the learnable parameter α of Eq. (10) which controls the balance between timestep information and image complexity in the aggregation process, the final value of our final model is 0.655. This indicates that the model emphasizes timestep information, while still incorporating image complexity.

The image complexity distribution of the garments in VITON-HD dataset and DressCode dataset are organized in Table 2.

Data	Sparsity		Variance		Gradient	
	Avg.	Std.	Avg.	Std.	Avg.	Std.
V-HD	0.125	0.011	1.139	0.077	0.462	0.013
DC-U	0.127	0.010	1.156	0.081	0.459	0.015
DC-L	0.128	0.008	1.161	0.065	0.450	0.014
DC-D	0.129	0.008	1.157	0.057	0.456	0.015

Table 2. The V-HD, DC-U, DC-L, and DC-D denote the VITON-HD, DressCode Upper-body subset, Lower-body subset, and Dresses subset, respectively. **Avg.** and **Std.** denote the average and standard deviation of the values, respectively.

The results highlight subtle differences between the VITON-HD and DressCode datasets, particularly for upper-body garments. DressCode garments exhibit slightly higher sparsity and variance compared to those in VITON-HD, suggesting that individual garments in DressCode may contain patterns with relatively greater complexity, such as logos or prints, which contribute to increased pixel intensity variations. In contrast, the lower average gradient magnitude in DressCode samples indicates that these patterns often have smoother transitions or softer boundaries, likely due to similar colors between garments and their prints or repetitive designs with subtle changes. Meanwhile, the relatively higher gradient values in VITON-HD garments suggest that they may include simpler, more distinct patterns, such as large logos with sharp edges and contrasting colors.

Although the differences exist, they are subtle. The similar garment complexity across the datasets explains the small gap in FID scores between the VITON-HD and DressCode Upper-body datasets in Table 1. The performance of our model on DressCode dataset, including Lower-body and Dresses, are shown in Table 3.

Subset	LPIPS↓	SSIM↑	FID(p./unp.)↓
Upper-body	0.034	0.951	5.412/10.069
Lower-body	0.052	0.931	6.109/12.335
Dresses	0.080	0.883	6.957/10.662

Table 3. Performance of our ITA-MDT on three subsets of DressCode dataset. **p.** and **unp.** denotes paired and unpaired generation evaluation, respectively.

12.2. Analysis on SRE

Figure 8 illustrates examples of the Entropy Map X_e and the corresponding extracted Salient Region X_s from a given garment image X. When no dominant high-entropy cluster is detected, such as solid-colored and uniformly patterned garments, SRE tends to extract a broader region or even the entire garment, as shown in Figure 10 (right). While this may seem less selective, it remains beneficial by reducing background region, capturing the full garment with higher-resolution information that enriches local detail representation. Examples of such cases are shown in Figure 9.

A potential concern is whether high-entropy elements such as wrinkles (possibly irrelevant to the target garment) might be mistakenly emphasized. However, in the virtual try-on setting, garments are photographed under controlled conditions, where such elements are likely part of the intended design. As a result, our entropy-based method remains robust for this task. Figure 10 (left) shows that SRE performs well even on garments with wrinkles, using both dataset and real-world examples.

To accelerate training and evaluation, Salient Regions were preprocessed in advance. The SRE process averaged about 1.563 seconds per image in our experimental environment.

12.3. Analysis on Mask Reconstruction Objective

The mask reconstruction objective of the Masked Diffusion Transformer (MDT) with its Side-Interpolator module is the key component that transforms a Diffusion Transformer (DiT) into an MDT. The impact of the mask reconstruction objective on training efficiency and performance is illustrated in Figure 11. The figure compares the early training progression of our ITA-MDT with and without the mask reconstruction objective. The mask reconstruction objective produces a steeper learning curve, indicating faster convergence.

12.4. Analysis on Sampling Steps

The number of sampling steps γ_{steps} used during inference directly influences the trade-off between inference speed and the quality of generated results. Table 4 highlights this trade-off. Higher γ_{steps} leads to improvement in FID but slight degradation in SSIM and LPIPS. This discrepancy may have

Figure 8. Garment image X with its Entropy Map X_e and its Salient Region X_s extracted with our Salient Region Extractor. Images are from the upper-body, lower-body, and dresses subset of the DressCode dataset.

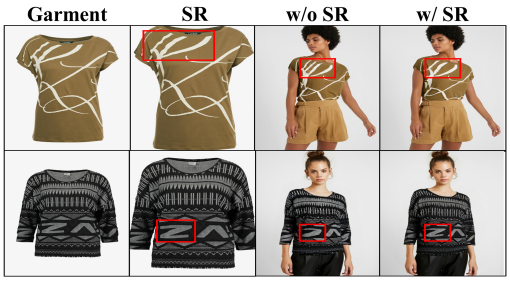

Figure 9. Results with and without Salient Region (SR).

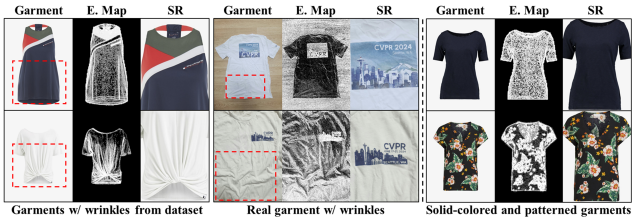

Figure 10. Salient Region Extractor (SRE) outputs: E. Map and SR refer to Entropy Map and Salient Region, respectively. Red box indicates wrinkled regions.

arisen because higher sampling steps refine fine-grained textures, which improve perceptual quality captured by FID, but may slightly alter pixel-level structural consistency, affecting SSIM and LPIPS. Additionally, the longer sampling trajectory may have introduced small deviations in structure as the latent representation evolves.

While we balance these factors by using $\gamma_{\text{steps}} = 30$, a reduction of γ_{steps} can be considered for accelerated use cases where inference speed is prioritized over marginal improvements in perceptual quality.

Figure 11. Comparison of training efficiency between ITA-MDT with and without the mask reconstruction objective with side-interpolator of Masked Diffusion Transformer(MDT). Evaluated with VITON-HD paired.

γ_{steps}	Inf. Time (s)	LPIPS↓	SSIM↑	FID↓
20	**3.207**	0.084	**0.888**	5.799
25	3.912	0.084	**0.888**	5.594
30	4.606	**0.083**	0.885	5.462
35	5.284	**0.083**	0.885	5.355
40	5.951	**0.083**	0.885	5.322
45	6.938	**0.083**	0.884	5.347
50	7.427	0.084	0.884	**5.293**

Table 4. Trade-off between sampling steps (γ_{steps}), Inference Time (Inf. Time), and image quality metrics on VITON-HD paired evaluation. Note that the optimal classifier-free guidance scale α_{cfg} and power scaling factor β_{scale} of our model were determined using γ_{steps} of 30.

13. Qualitative Results and Comparison

We provide a detailed overview of the qualitative results of our ITA-MDT framework and its comparison to previous methods, highlighting its superior fidelity in capturing the texture and color of the garments.

- Figure 12: Qualitative comparison of the effect of each component of the ITA-MDT framework on VITON-HD.
- Figure 13: Qualitative comparison between our ITA-MDT and previous methods on the VITON-HD.
- Figure 14: More qualitative comparison between our ITA-MDT and previous methods on the VITON-HD.
- Figure 15: Qualitative comparison between our ITA-MDT and previous methods on the DressCode Upper-body.
- Figure 16: More qualitative comparison between our ITA-MDT and previous methods on the DressCode Upper-body.
- Figure 17: Qualitative results of our ITA-MDT on the DressCode Upper-body.
- Figure 18: Qualitative results of our ITA-MDT on the DressCode Lower-body.
- Figure 19: Qualitative results of our ITA-MDT on the DressCode Dresses.

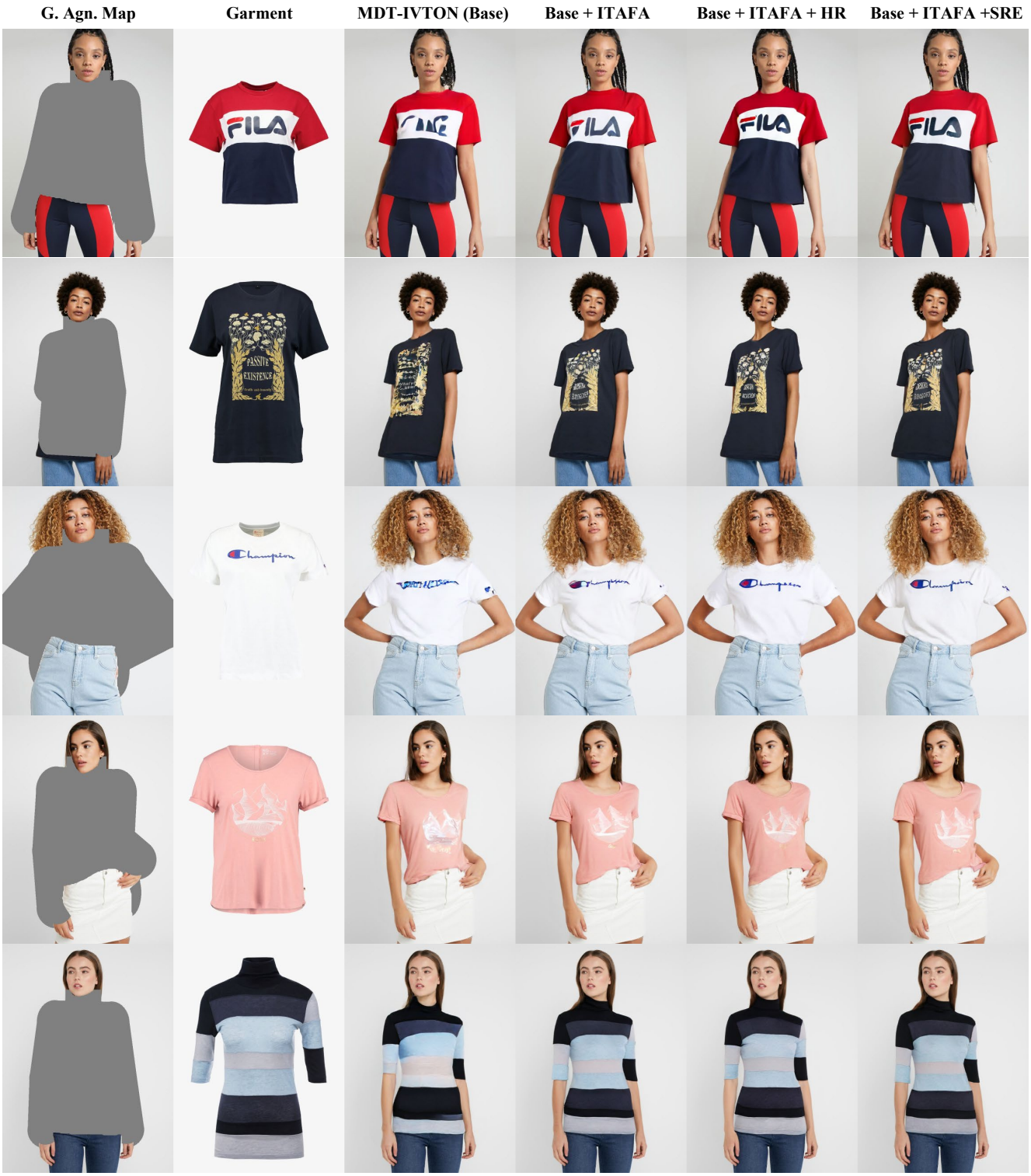

Figure 12. Qualitative comparison of the effect of each component of the ITA-MDT framework on VITON-HD. HR refers to the use of single High-resolution $(448 \times 448 \times 3)$ garment image to formulate condition vector c.

Figure 13. Qualitative comparison between our ITA-MDT and previous methods on the VITON-HD.

Figure 14. More qualitative comparison between our ITA-MDT and previous methods on the VITON-HD.

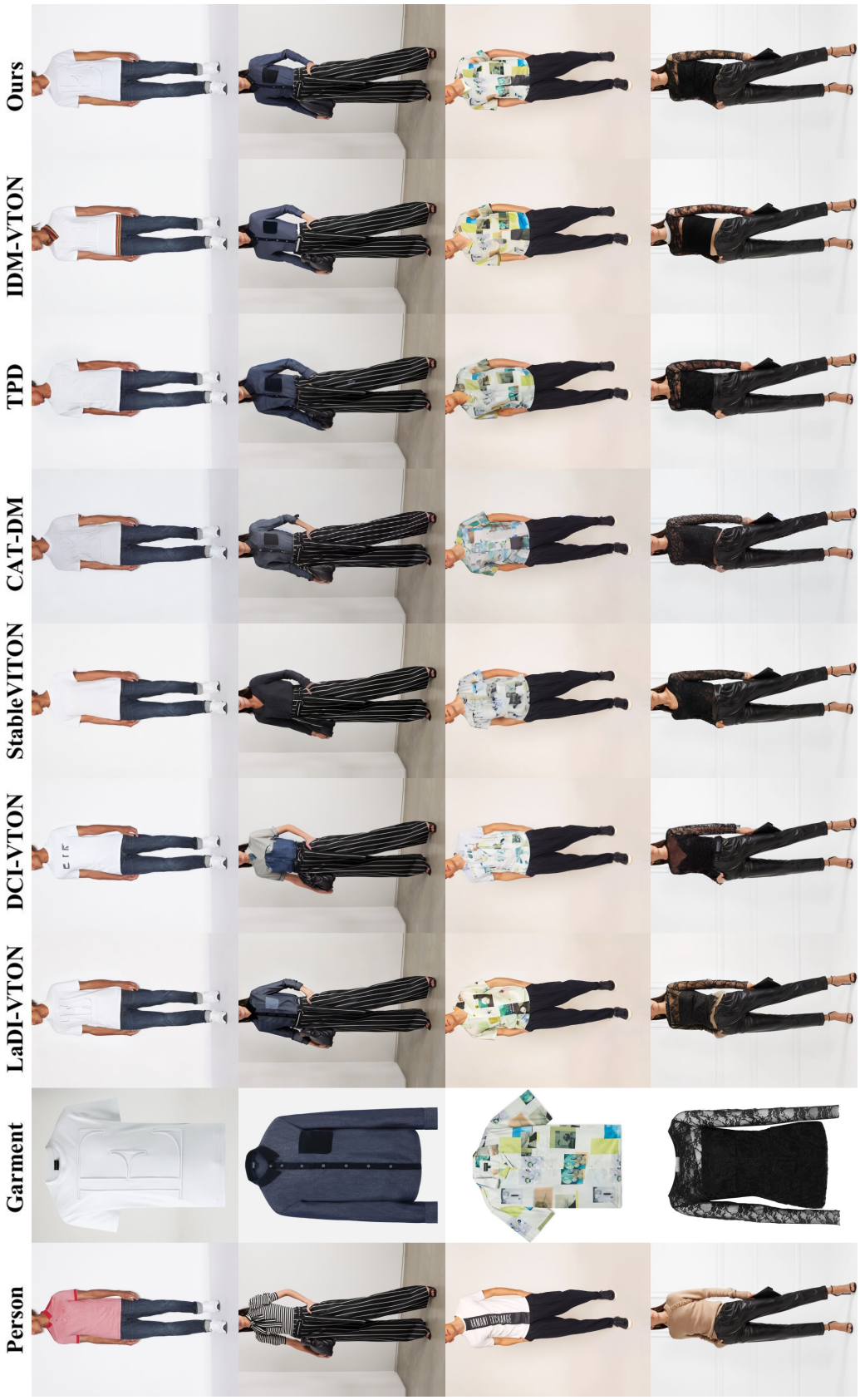

Figure 15. Qualitative comparison between our ITA-MDT and previous methods on the DressCode Upper-body.

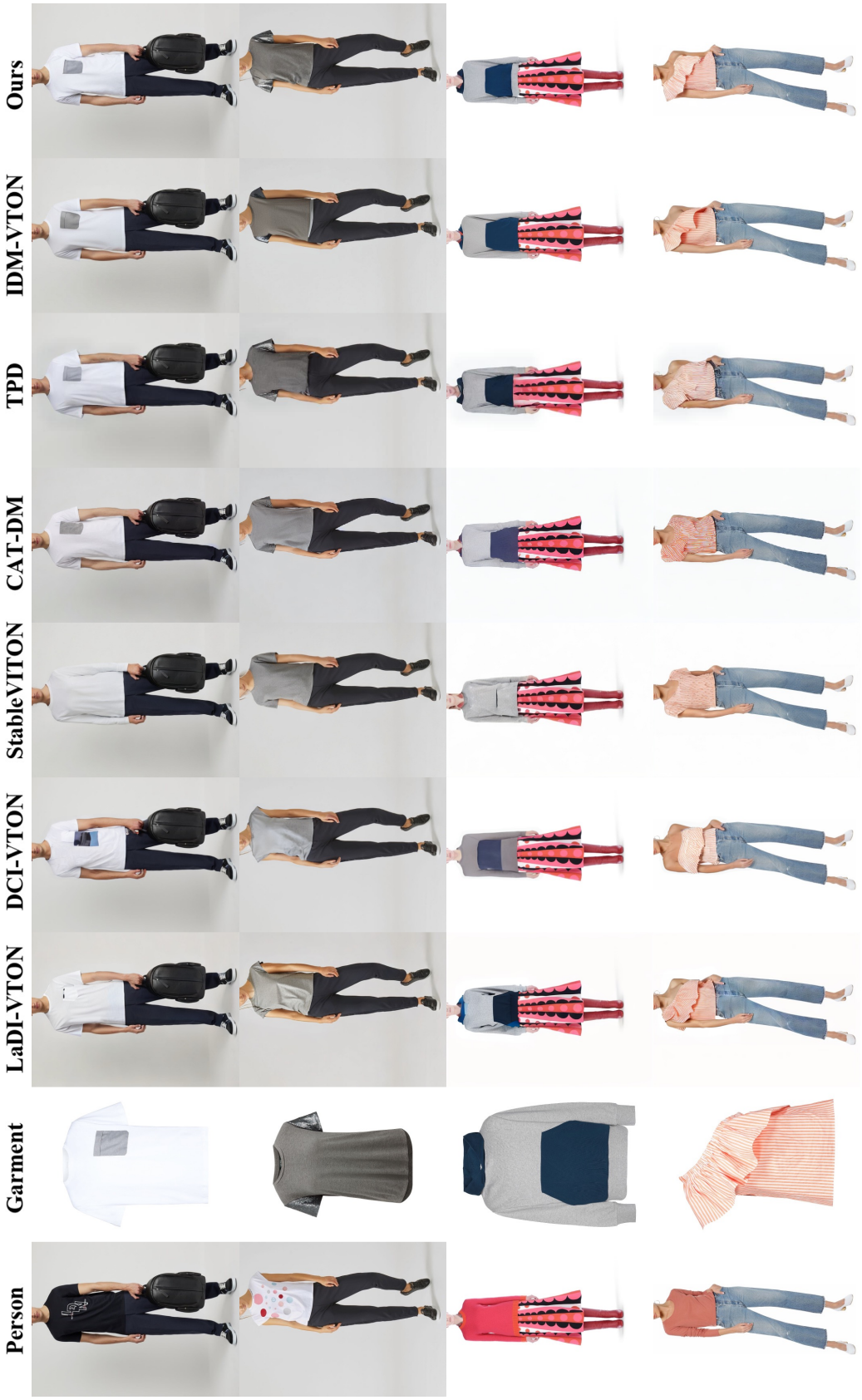

Figure 16. More qualitative comparison between our ITA-MDT and previous methods on the DressCode Upper-body.

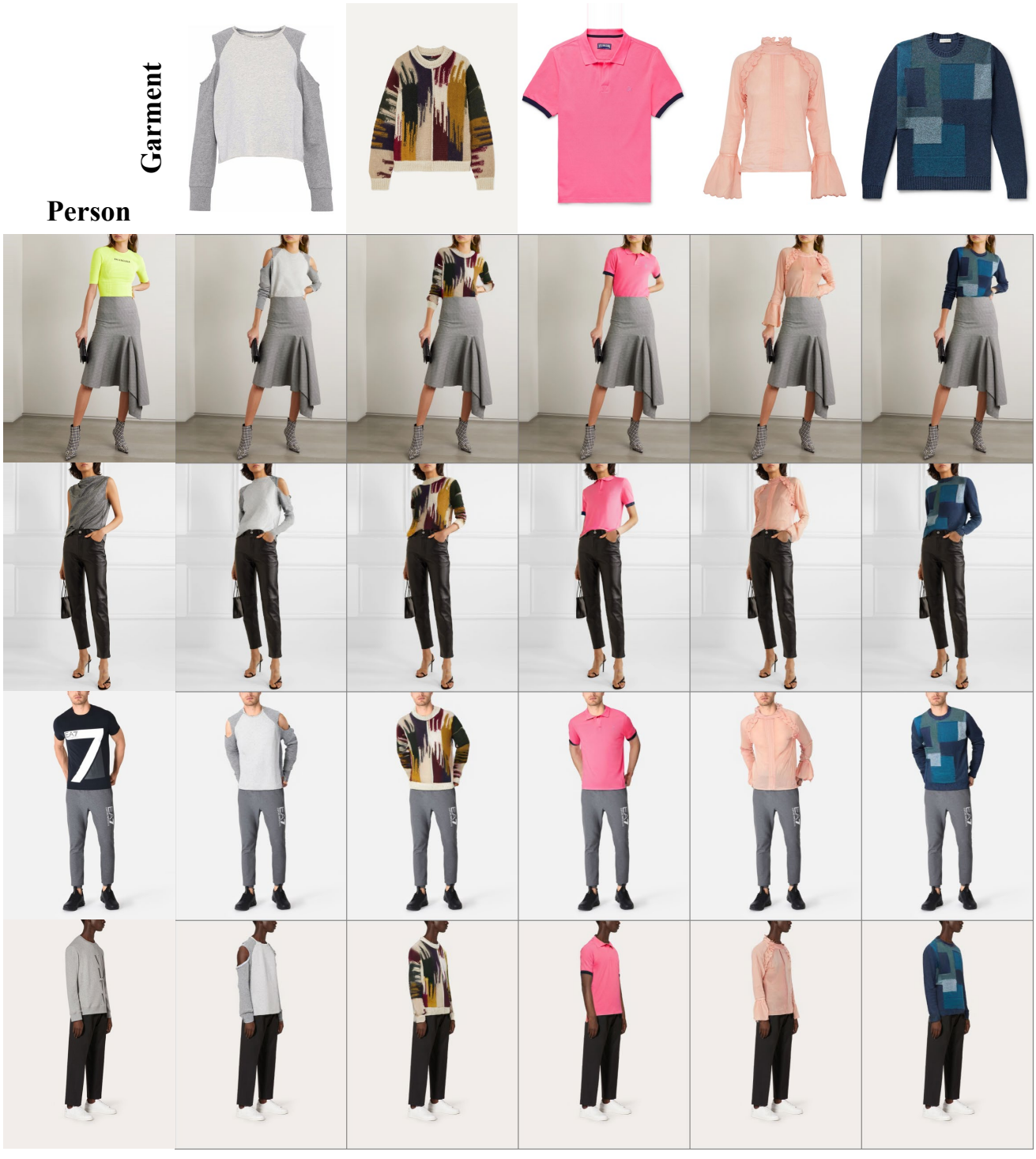

Figure 17. Qualitative results of our ITA-MDT on DressCode Upper-body.

Figure 18. Qualitative results of our ITA-MDT on DressCode Lower-body.

Figure 19. Qualitative results of our ITA-MDT on DressCode Dresses.