One Model for All: Unified Try-On and Try-Off in Any Pose via LLM-Inspired Bidirectional Tweedie Diffusion

Supplementary Material

7. Implementation Details

7.1. Training and Inference Details

We first train the model on the VITON-HD dataset with a resolution of 384×512 for 20K iterations. Then, keeping the same learning rate and batch size, we fine-tune the model on both the VITON-HD and DeepFashion-MultiModal datasets using a higher resolution of 768×1024. For data augmentation, we enhance the background color of the generated garment. Specifically, we use a tensor of the same size as the input garment, with all values set to 255, and concatenate it with the garment latent $\mathcal{E}(I_g)$ along the channel dimension. To align the latent with the UNet input along the channel dimension, we apply separate convolutional layers to each component of the joint input, projecting their channels to 320. Each convolutional layer is initialized with the first several channels of the corresponding layer in the pretrained UNet. During inference, if the try-on task is required, the inputs needed are the person image I_p, the garment I_g, and the person's portrait I_h, whereas for the try-off task, only I_p is needed, with the other inputs set to 0. Our implementation is based on the PyTorch deep learning framework (version 2.1.2), with the diffusion model adapted from HuggingFace's Diffusers library.

7.2. Evaluation Metrics

In our experiments, we adopt a variety of evaluation metrics commonly used in generative tasks. Among them, SSIM [41], LPIPS [51], KID [1], and FID [15] are widely used general metrics in related work. This section focuses on the detailed computation of several additional quantitative metrics used in our method, including DINO similarity [29], CLIP-I [32], and LLM-based Image Scoring.

CLIP-I. CLIP focuses on semantic alignment similarity between images. Specifically, we utilize the CLIP-ViT-B/32 model as the feature extractor. Given a pair of images, the model encodes them into two 512-dimensional feature vectors. We then compute the cosine similarity between these vectors to measure their semantic similarity—a higher similarity indicates more semantically consistent content.

DINO Similarity. DINO similarity focuses on structural and fine-grained detail similarity between images. We utilize the DINOv2-Base model to extract features for a pair of images. For each image, we apply mean pooling across the patch embeddings from the final layer of the model's output, resulting in a 768-dimensional feature vector. We then calculate the cosine similarity to measure.

LLM-based Image Scoring. We provide the model image, garment image, and the try-on result as input to GPT-4o-mini. The prompt for the multimodal large language model is presented in Fig. 9.

8. Additional Ablation Studies

Experimental Setup for SMPL-X Ablation. Specifically, we trained two low-resolution models on VITON-HD, using DensePose (provided by the VITON-HD preprocessing) and SMPL-X respectively as structural conditioning, which were concatenated along the input channels. The configuration uses a resolution of 384×512, a per-GPU batch size of 8, and two 80-GB A100 GPUs for 45,000 training steps. At inference, the guidance scale is set to 2.0.

Ablation Studies for Joint Training Strategy. To evaluate the impact of unified training on performance, we conduct an ablation using the same low-resolution configuration as described above. In the joint-training setting, each batch is duplicated and divided into two halves, one for try-off training and the other for try-on training. Tab. 6 compares two training setups with no performance degradation. We attribute this to our bidirectional Tweedie diffusion mechanism, which explicitly defines the context and generation task by adjusting the combination of noising targets and conditions, thereby avoiding cross-task context confusion. Moreover, unified training encourages the model to learn bidirectional, shared feature correlations between garments and the human body, yielding more essential representations and improving the performance and robustness of both tasks.

Table 6. **Quantity ablation of unified training strategy.** where "(u)" indicates that the metric is computed in the unpaired setting. Training the network to learn both try-on and try-off tasks does not degrade performance.

Setting	SSIM↑	LPIPS↓	CLIP-I (u)↑	FID (u)↓	KID (u)↓
Try-on training	**0.840**	**0.150**	0.866	11.919	3.603
Joint training	0.839	0.152	**0.872**	**11.781**	**3.602**

9. More Qualitative Results

9.1. Virtual Try-on

Person-to-person Virtual Try-on. Fig. 10 presents additional try-on comparative results in the person-to-person scenario on the VITON-HD dataset. Specifically, when the input clothing is not an exhibition garment, the input warp cloth often exhibits incomplete contours and distorted textures, which further exacerbates the artifacts in the try-on results, ultimately leading to suboptimal performance. Fig. 11 and Fig. 12 present more generated results of our proposed OMFA model in this task, further demonstrating that our model maintains excellent detail preservation in both the try-off and try-on steps, leading to robust and high-fidelity results.

Multi-pose Virtual Try-on. As shown in Fig. 13, we present pose transfer try-on results on the VITON-HD dataset. Specifically, we select three different target poses and replace the original pose parameters with the corresponding SMPL-X parameters to enable pose variation in try-on. We also provide try-on results under the original pose as a reference. Since we only replaced the pose parameters of SMPL-X while keeping the shape parameters unchanged, the generated body meshes exhibit different poses but consistent body shape, which helps achieve natural and identity-consistent try-on results.

9.2. Virtual Try-off

As shown in Fig. 14, we present additional try-off comparison results on the VITON-HD dataset. Additionally, we performed garment reconstruction on two open-source datasets DressCode and DeepFashion-MultiModal and visualized the quantitative results, as illustrated in Fig. 15 and Fig. 16. As demonstrated above, our method effectively handles complex poses and occlusions, accurately restoring the garment's canonical shape while highly preserving its texture and structural details.

9.3. User Study

We conducted a user study with 50 participants using the model trained on the VITON-HD dataset. Each participant was randomly assigned 10 samples from a pool of 50 for evaluation, with each sample containing six different virtual try-on results generated in the person-to-person scenario. Participants were asked to choose the best result using three criteria: image fidelity, human identity, and garment consistency. We totaled the number of times each method was chosen as the best across all test samples and calculated the average voting proportion for each method. As shown in Tab. 7, our method had the highest average voting proportion among all examples, indicating visually superior results and a significant advantage in human evaluation.

Table 7. **User study results.** We report the best-choice rate for our method and seven other methods, including StableVITON, OOT-Diffusion, and CatVTON.

	OMFA	IDM-VTON	MV-VTON	others
Best Choice Rate	46%	8.6%	18.8%	26.6%

10. Limitations

Due to a lack of paired data for multi-layer garments, our proposed method does not provide multi-layer try-on/try-off. Furthermore, our architecture is only intended for a single garment input, whereas multiple garment inputs may dramatically extend the input sequence. In the future, we will incorporate more in-the-wild data to develop computationally efficient virtual try-on solutions that are more in line with real-world application scenarios.

LLM-based Try-on Image Evaluation Prompt

Task Definition:
You will be given three images: a person image, a cloth image, and a generated try-on image (synthesized from the first two to simulate realistic try-on). As an evaluator, assess the generated image based on garment-person integration quality, garment preservation accuracy, and person identity consistency. Focus on whether the generated image faithfully reflects the original garment's structure and texture while preserving the person's appearance and natural pose.

Scoring Criteria:
The reasonableness of the try-on result is typically judged based on the following visual features:
1. *General Image Quality*: Evaluates overall realism and aesthetics, including clothing integration with, color harmony, and stylistic consistency. Avoid anatomical errors, artifacts, unnatural wrinkles, or inconsistent lighting.
2. *Garment Preservation*: Assesses preservation of original clothing structure, texture, and fine details (patterns, text, stripes, logos) in the generated image.
3. *Size Fitness*: Evaluates whether clothing fits naturally, reflecting its original silhouette (loose, fitted, etc.) while adapting to body shape. Avoid floating, stretching, scaling issues, or misalignment.
4. *Pose Reasonableness*: Checks for natural, anatomically correct poses with coherent skeleton structure, realistic limb placement, distinct fingers, and consistent facial orientation. Avoid misalignments, merged fingers, or distortions.

Scoring Range:
You need to give a score based on the above scoring criteria ranging from 0 to 10. Your score should be rounded to two decimal places, and should aim to reflect clear distinctions.

Input format:
You will receive three images: the first is the person image, the second is the cloth image, and the third is the try-on result.

Output Format:
Your Score

Note: You must adhere to the specified output format, which means that only the scores need to be output, excluding your analysis process.

Figure 9. **The prompt for GPT-4o-mini to evaluate try-on results quality.**

Figure 10. **More qualitative comparison of try-on results on VITON-HD.** *Please zoom in to better observe the details.*

Figure 11. **More qualitative results on VITON-HD.** *Please zoom in to better observe the details.*

Figure 12. **More qualitative results on VITON-HD.** *Please zoom in to better observe the details.*

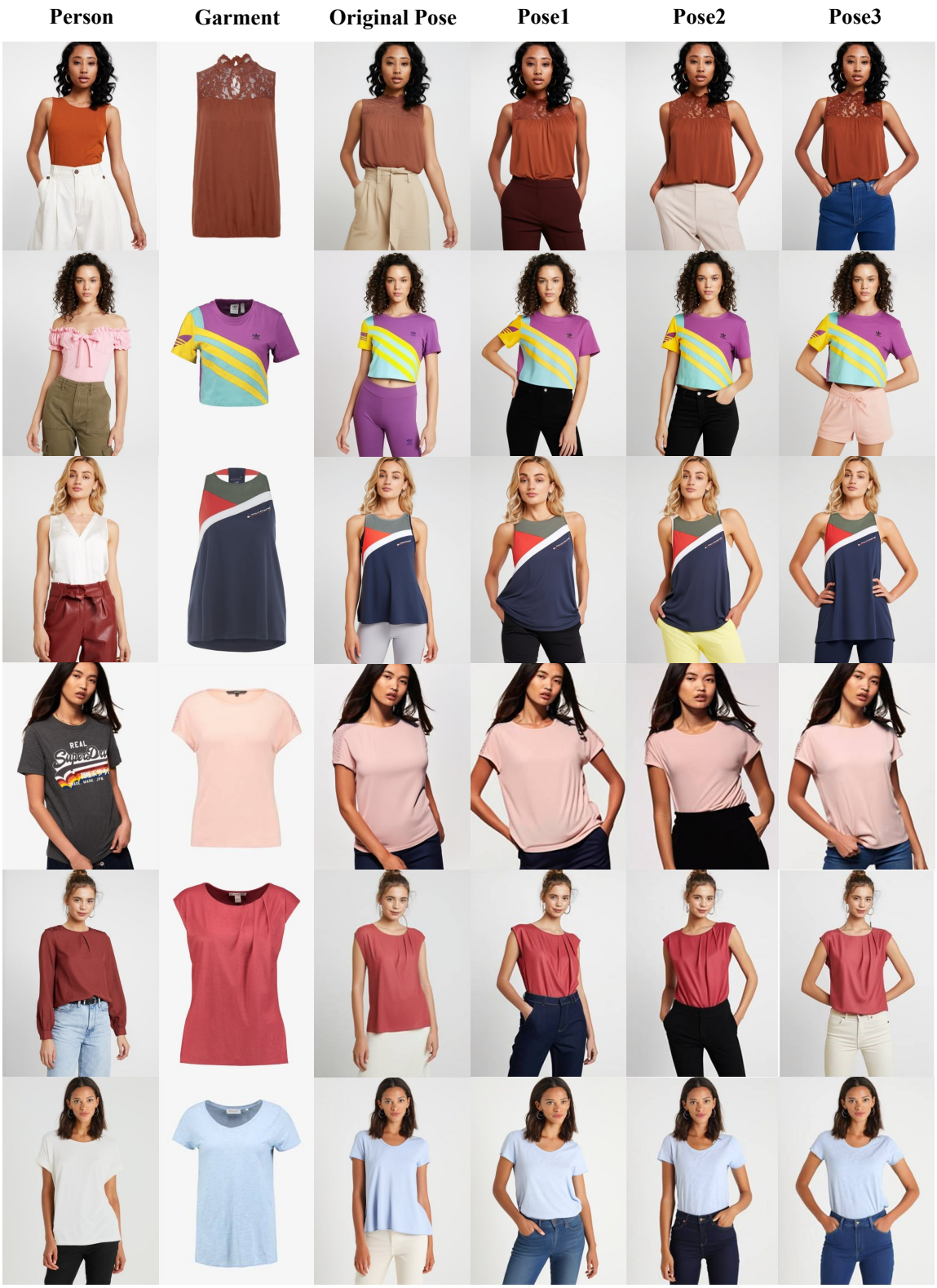

Figure 13. **Qualitative result of multi-pose try-on on VITON-HD.** *Please zoom in to better observe the details.*

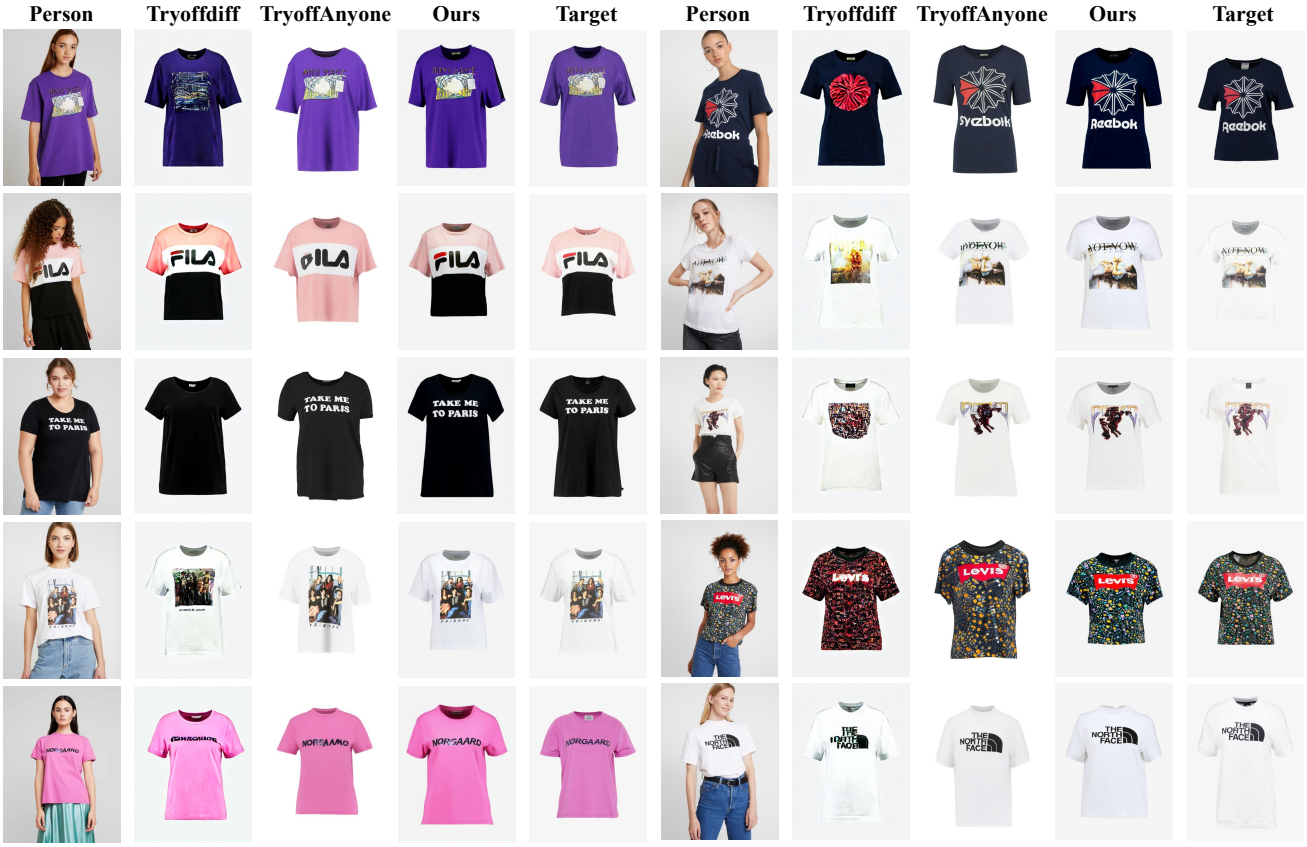

Figure 14. **More qualitative comparison of try-off results on VITON-HD.** *Please zoom in to better observe the details.*

Figure 15. **Qualitative try-off results on DressCode upperbody.** *Please zoom in to better observe the details.*

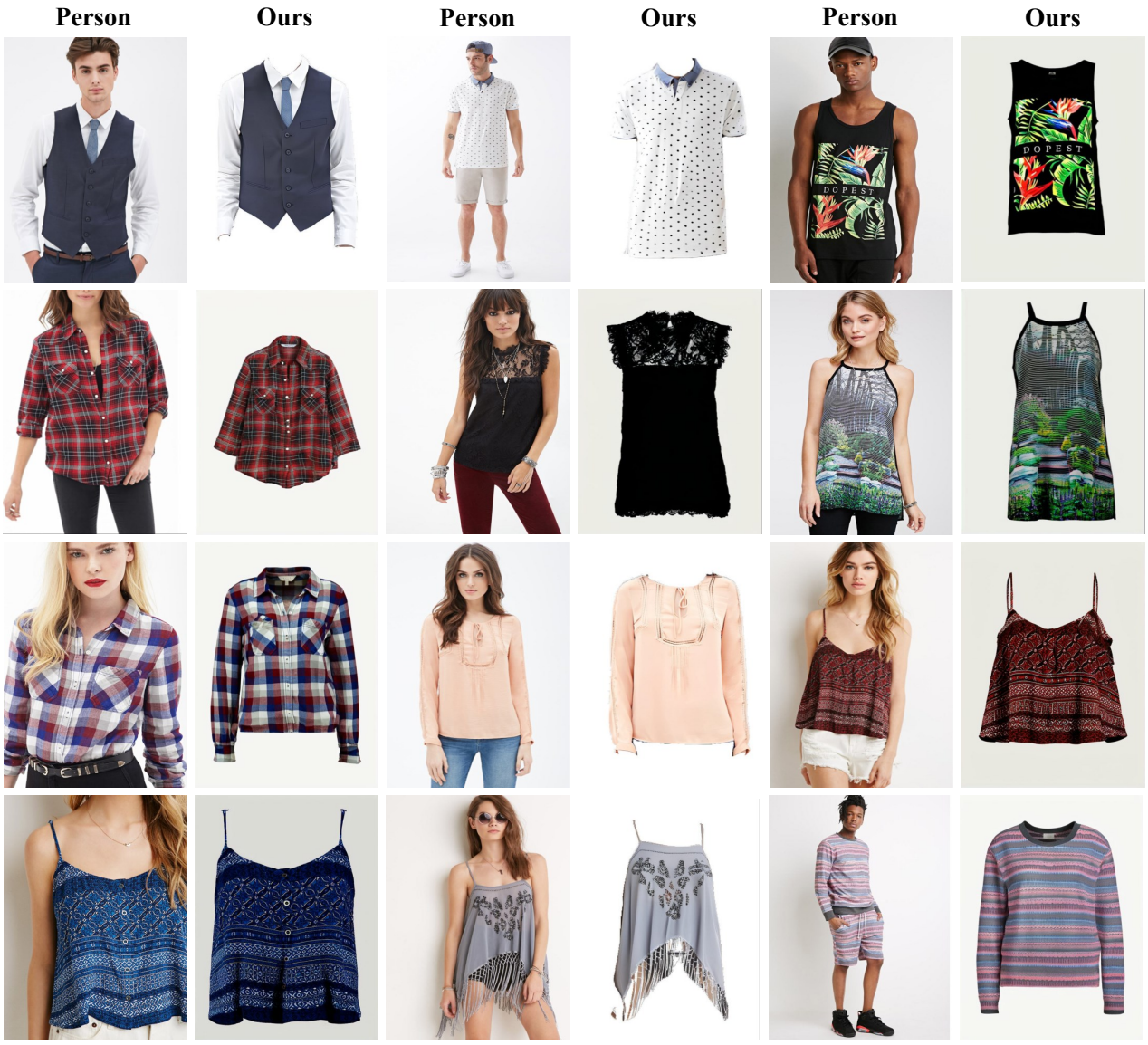

Figure 16. **Qualitative try-off results on DeepFashion-MultiModal.** *Please zoom in to better observe the details.*