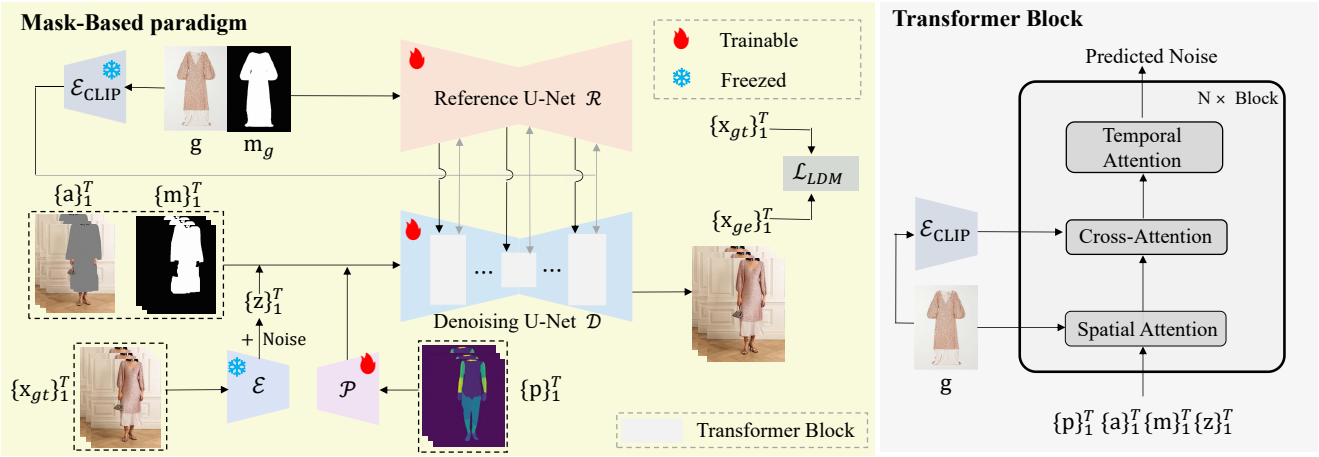

Figure 1. The pipeline of our mask-based video virtual try-on paradigm.

Appendix

1. Baseline Architecture

The pipeline of our mask-based baseline model is shown in Fig. 1. Specifically, in the training stage, given the source person video $\{\mathbf{x}_{gt}\}_1^T = \{\mathbf{x}_1, \cdots, \mathbf{x}_T\} \in \mathbb{R}^{T\times 3\times H\times W}$ and garment image $\mathbf{g} \in \mathbb{R}^{3\times H\times W}$, where H and W denote the height and width of each frame and T denotes the length of the source video, we first extract the corresponding inputs respectively for the denoising U-Net $\mathcal{D}$ and reference U-Net $\mathcal{R}$ with the same architecture and initial weight. The input latent $\{\mathbf{d}\}_1^T$ of $\mathcal{D}$ is the concatenation of the noisy latent $\{\mathbf{z}\}_1^T \in \mathbb{R}^{T\times 4\times h\times w}$, where $\{\mathbf{z}\}_1^T$ is formed by adding Gaussian noise to the latent $\mathcal{E}(\{\mathbf{x}_{gt}\}_1^T)$ and $h = H/8$ and $w = W/8$, the cloth-agnostic video latent $\mathcal{E}(\{\mathbf{a}\}_1^T)$ and a binary agnostic mask sequence $\{\mathbf{m}\}_1^T$. The pose information is introduced to the $\mathcal{D}$ by adding the human pose latent $\mathcal{P}(\{\mathbf{p}\}_1^T)$ to $\{\mathbf{d}\}_1^T$, where $\mathcal{P}$ represents the pose encoder in [4]. The input latent $\mathbf{r}$ of $\mathcal{R}$ is the concatenation of the garment latent $\mathcal{E}(\mathbf{g})$ and the binary garment mask $\mathbf{m}_g$. The transfer of fine-grained garment information is conducted by the spatial attention between the corresponding features of the layers in $\mathcal{D}$ and $\mathcal{R}$. Besides, the high-level semantic CLIP feature $\mathcal{E}_{\text{CLIP}}(\mathbf{g})$ of the garment is inserted to $\mathcal{D}$ through a cross attention to enhance the ability of the garment transfer. The coherence of generated video is ensured by the temporal attention [5, 6] which is devised after each cross attention layer in $\mathcal{D}$.

2. Inference Scheme

Compared to existing video virtual try-on methods, our PEMF-VVTO provides a more intelligent and user-friendly virtual try-on experience. Specifically, according to most human videos with simple action postures, it is enough to achieve accurate and desirable try-on results even without the guidance of point alignments. However, when dealing with more complex human videos (*e.g.* street dance videos) or aiming at more controllable try-on results, users can manually click on the matching point pairs between a randomly selected single video frame and reference garment image, thereby acquiring a more coherent and natural generation. Besides, to enable the smooth and consistent try-on results for long videos, we follow [4] to employ a sliding window strategy in the inference stage. Concretely, the long video will be first split to multiple overlapping segments. Then, our model will conduct the inference on each segment. The final result of overlapping frames is acquired by taking the average over each inference.

3. Datasets

Following our baseline [4], we train our model on two image datasets, VITON-HD [1] and DressCode [8], and one video dataset, ViViD [4]. The commonly used VVT [3] dataset is selected as the evaluation dataset to obtain a fair performance comparison. Besides, we also choose more challenging StreetVTON image dataset [2] and TikTok video dataset [7] to show our visualization results.

- VITON-HD includes 13,679 pairs of upper-body model and garment images, with 2,032 pairs choosing as testing data.
- DressCode contains 15,363, 8,951 and 2,947 pairs of full-body model respectively corresponding upper-body, lower-body and dress garment images.
- StreetVTON is a subset collected from DeepFashion to conduct wild image try-on task, containing 2,089 person images with complex backgrounds and body postures.
- VVT consists of 791 video clips with a resolution of 256×192, which is split into a training set of 661 clips and a test set of 130 clips.
- ViViD contains 9,700 model videos and garment images with a resolution of 832×624. Following [4], we divide it into 7,759 videos as the training set and 1,941 videos as

Figure 2. The visualization results for the generation of pseudo-person data and try-on results of mask-free model. The results of top row and bottom row are respectively from VITON-HD and ViViD datasets. *Best viewed with Acrobat Reader. Click the images to play the video clips.*

the test set.

- TikTok comprises 340 dance videos that captures a single person dancing with complex body movements. The virtual try-on for this dataset is more challenging than simple human model videos in VVT and ViViD.

4. More Visualization Results

In this section, we will show the visualizations of pseudo-person data, image virtual try-on results and video virtual try-on results.

4.1. Image Virtual Try-on

In Fig. 3, we present more visualization results of VITON-HD and StreetVTON datasets. Our generated results achieve both the accurate garment transfer and background consistency, demonstrating the generalizability of our method on image try-on data. Besides, we also choose multiple garments of the same type but different styles to conduct the virtual try-on with the same person. As Fig. 4 shown, our method can also produce superior try-on results in this more challenging setting.

4.2. Video Virtual Try-on

To better exhibit video virtual try-on results, we make a video demo to intuitively reflect the performance of our method. It contains all video results in the article and provides more visualization results of ViViD and TikTok datasets. For more detailed information, please watch the video demo provided in the supplementary materials.

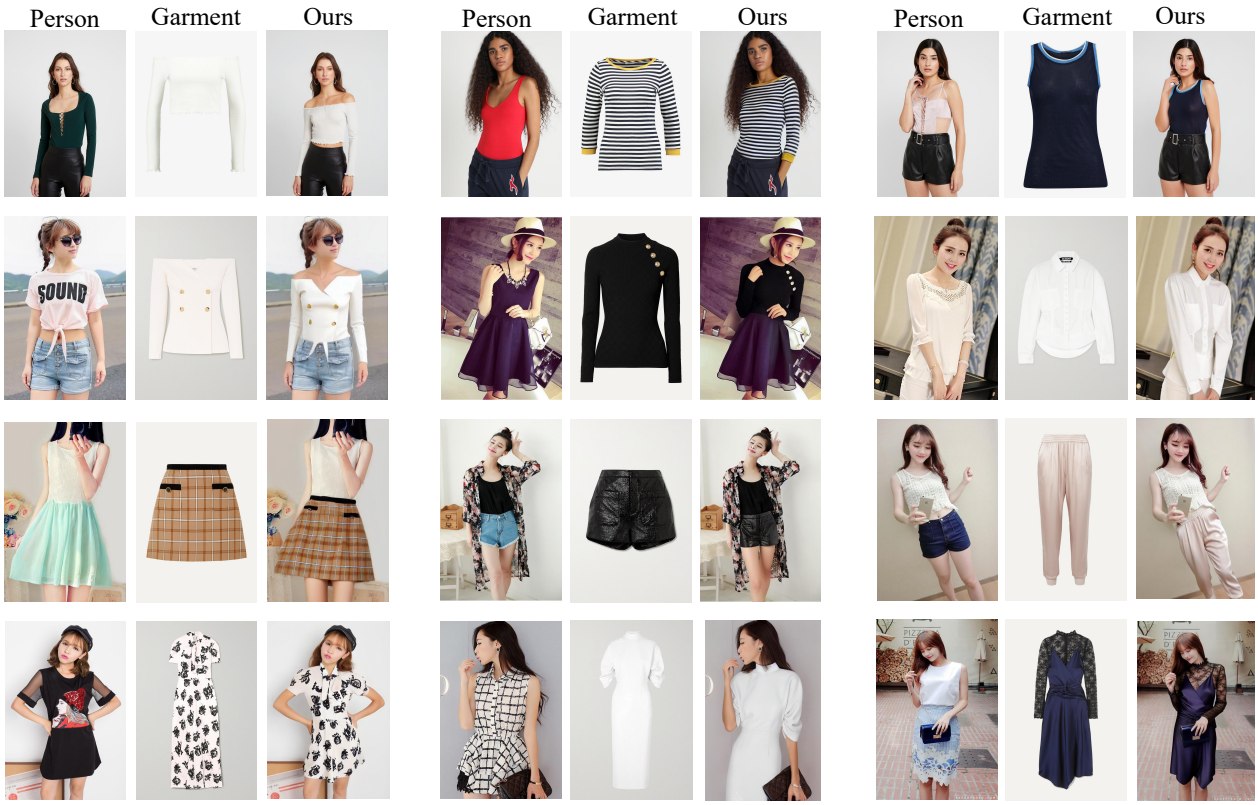

Figure 3. The visualization results for VITON-HD and StreetVTON datasets. Rows from top to bottom respectively denote the VITON-HD try-on results with upper garments, StreetVTON try-on results with upper garments, StreetVTON try-on results with lower garments and StreetVTON try-on results with dress garments.

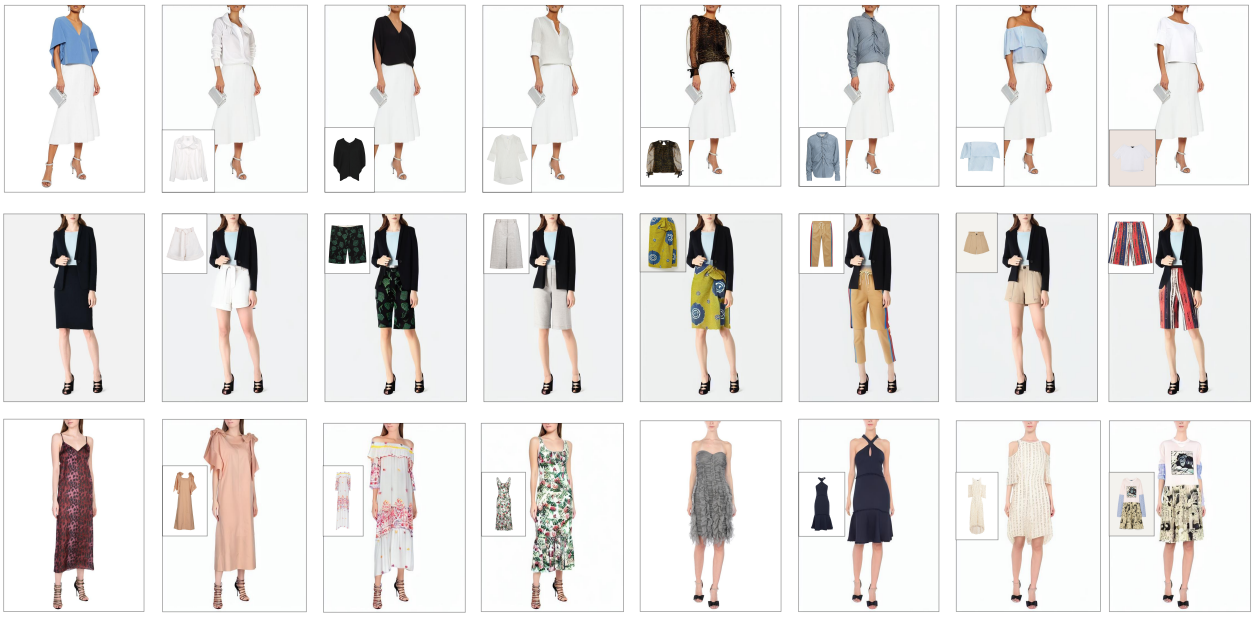

Figure 4. The visualization results for DressCode dataset. Rows from top to bottom denote the DressCode try-on results with upper, lower and dress garments.