

Tunnel Try-on: Excavating Spatial-temporal Tunnels for High-quality Virtual Try-on in Videos

Anonymous Authors

The structure of this supplementary material can be summarized as follows. Section A provides more implementation details. Section B discusses the comparison with Animation methods. Section C presents more explanation and results on the user studies. Section D discusses the limitations and future work of our Tunnel Try-on. Moreover, more comprehensive visualization results are displayed in Section E. Note that all figures in the main text and appendices have corresponding videos in the supplementary materials.

A IMPLEMENTATION DETAILS

A.1 Focus Tunnel Extraction

We designed a tunnel extraction rule based on the pose map to ensure stable and accurate extraction of regions centered around the person while covering the areas requiring try-on. Specifically, we first used the pose map extracted by DW-Pose [6]. Then, depending on the reference clothing type (upper or lower), we calculated the minimum bounding box and expanded it outwardly. For the upper clothing try-on, we computed the minimum bounding box of the upper body in the pose map as the initial box. Then, we extended the bottom boundary of the box to the knee position and the top boundary by the height distance from the shoulders to the head. For the lower clothing try-on, we calculated the minimum bounding box of the lower body in the pose map as the initial box. Then, we extended the top boundary of the box to the upper third point from the shoulder to the hip and expanded the bottom boundary by 0.25 times the height. If the relevant points do not exist in the pose map, the top and bottom boundaries of the box are uniformly expanded by 0.25 times the height. Then, we adjusted the width to be equal to the new height. For areas beyond the image, we performed padding operations.

A.2 Focus Tunnel Smoothing

The tunnel obtained through focus tunnel extraction may introduce unexpected jitter due to errors in the bounding box prediction and errors introduced during the outward expansion process. To address this issue, we propose the focus tunnel smoothing strategy to eliminate these jitters and achieve a smoother, more stable tunnel.

Specifically, we first smooth the tunnel using a Kalman filter, as shown in Algorithm 1. Then, we add a low-pass filter in Algorithm 2 to filter outliers. The visualized curves of the center coordinates and the size of the focus tunnel before and after filtering can be seen in figure 1. Evidently, after filtering, the jitter in the focus tunnel disappears, and the transition in the tunnel becomes smoother.

B COMPARISON WITH ANIMATION

Image Animation enables images to move according to a specified pose sequence, simulating effects similar to real videos. Integrating image visual try-on methods and image animation methods may lead to a solution method for video visual try-on.

Algorithm 1: Kalman Filter.

Input: Raw box coordinate \mathbf{x} , number of the tunnel boxes N .
Result: Smoothed box coordinate $\hat{\mathbf{x}}$.

```
1 Initialize  $P_0 = \mathbf{x}_1, \hat{\mathbf{x}}_0 = \mathbf{x}_1, Q = 0.001, R = 0.0015, t = 1$ .
2 repeat
3   Project the state ahead  $\hat{\mathbf{x}}_t^- = \hat{\mathbf{x}}_{t-1}$ .
4   Project the error covariance ahead  $P_t^- = P_{t-1} + Q$ .
5   Compute the Kalman Gain  $K_t = P_t^- (P_t^- + R)^{-1}$ .
6   Update the estimate  $\hat{\mathbf{x}}_t = \hat{\mathbf{x}}_t^- + K_t(\mathbf{x}_t - \hat{\mathbf{x}}_t^-)$ .
7   Update the error covariance  $P_t = P_t^- (1 - K_t)^{-1}$ .
8    $t \leftarrow t + 1$ .
9 until  $t > N$ ;
Output:  $\hat{\mathbf{x}}$ 
```

Algorithm 2: Low-pass Filter.

Input: The smoothed box coordinate of Kalman Filter $\hat{\mathbf{x}}$, number of the tunnel boxes N , filter window size L .

Result: Final box coordinate \mathbf{y} .

```
1 Initialize filter window  $W = \hat{\mathbf{x}}_{1:L}, i = 1, r = 0.1$ .
2 repeat
3    $j = 1, W_1 = \hat{\mathbf{x}}_i$ .
4   repeat
5      $W_{j+1} = W_{j+1} * r + W_j * (1 - r)$ .
6      $j \leftarrow j + 1$ .
7   until  $j > L$ ;
8    $\mathbf{y}_i = W_L$ .
9    $i \leftarrow i + 1$ .
10 until  $i > N$ ;
Output:  $\mathbf{y}$ 
```

To implement this integration, we utilize the initial frames generated by our model as input and feed them into the popular image animation framework Magic Animate [5]. The resulting generated video is depicted in figure 2.

It can be observed that directly integrating animation methods with image try-on methods to generate video try-on results presents at least two notable shortcomings. Firstly, animation can only generate static backgrounds due to the lack of background information from the original video. However, in practical applications, the movement of individuals and camera angles, or non-static backgrounds, can lead to continuous changes in the background. In such cases, the fusion of the animated video try-on results with the background appears rigid. Secondly, since the temporal control conditions are solely based on the pose sequence, animation methods struggle to produce high-fidelity try-on results when there are

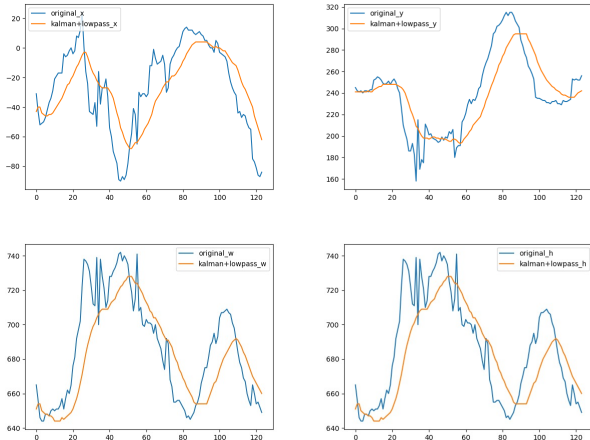


Figure 1: The effect of focus tunnel smoothing. The horizontal axis represents the frame index, and the vertical axis represents the corresponding values. It can be observed that the orange curve (after smoothing) exhibits less jitter compared to the blue curve (before smoothing), resulting in smoother transitions.



Figure 2: Results of Magic Animate. The animation method cannot handle changing backgrounds (third row), and the fidelity of the generated try-on videos is very low (all rows).

significant movements in the individual or variations in perspective. These limitations severely restrict the direct application of animation methods to real-world video try-on scenarios.

C USER STUDY

This section provides a detailed introduction to the criteria used in the user study conducted in the main text. Specifically, "Quality" denotes the image quality, encompassing aspects like artifacts, noise levels, and distortion. "Fidelity" measures the ability to preserve details compared to the reference clothing image. "Smoothness" evaluates the temporal consistency of the generated videos.

In figure 3, we present a typical example from the user study. While showing some continuity, FW-GAN [2] lags significantly behind other methods in generation quality and fidelity. PBAFN [3] accurately warps reference clothing only in close-up shots of the person, achieving satisfactory try-on effects, but it exhibits notable warping errors in other scenarios. AnyDoor [1] struggles to generate detailed clothing items such as letter logos. Stable VITON [4] generally demonstrates better generation quality and fidelity than the aforementioned methods. However, AnyDoor and Stable VITON still exhibit significant variations in generated results between frames, even under fixed random seed conditions, resulting in noticeable temporal inconsistencies. In contrast, our method maintains consistent, high-quality try-on results during person movements, showcasing significantly superior generation quality, fidelity, and smoothness compared to other methods.

D LIMITATIONS AND FUTURE WORK

Our Tunnel Try-on relies on accurate parsing to refine fine-grained inpainting masks required by the generation model. When parsing model segmentation results are erroneous, it may lead to leaks, resulting in generation failure. Therefore, we believe that accurate parsing results or mask generation methods with less reliance on parsing can further enhance the performance of our method.

E MORE QUALITATIVE RESULTS

We provide additional qualitative results demonstrating the high spatio-temporal consistency of our Tunnel Try-on. Figures 4, 5, and 6 showcase more try-on results involving various types of video motions. Additionally, Figures 7 and 8 present extra try-on results featuring various types of reference clothing on our collected dataset.

REFERENCES

- [1] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. 2023. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481* (2023).
- [2] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. 2019. Fw-gan: Flow-navigated warping gan for video virtual try-on. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1161–1170.
- [3] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. 2021. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8485–8493.
- [4] Jeongho Kim, Gyojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. 2023. StableVITON: Learning Semantic Correspondence with Latent Diffusion Model for Virtual Try-On. *arXiv preprint arXiv:2312.01725* (2023).
- [5] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. 2023. MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model. *ArXiv abs/2311.16498* (2023). <https://api.semanticscholar.org/CorpusID:265466012>
- [6] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. 2023. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4210–4220.



Figure 3: More examples of person movements varying in distance from the camera.

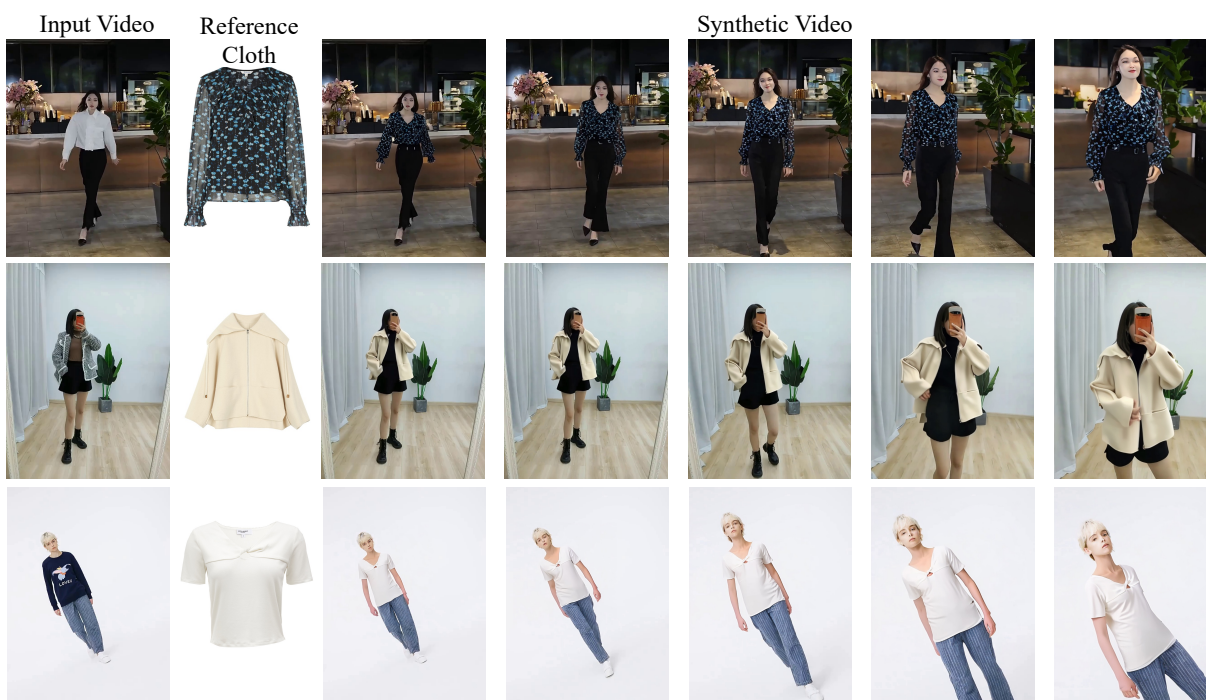


Figure 4: More examples of person movements varying in distance from the camera.

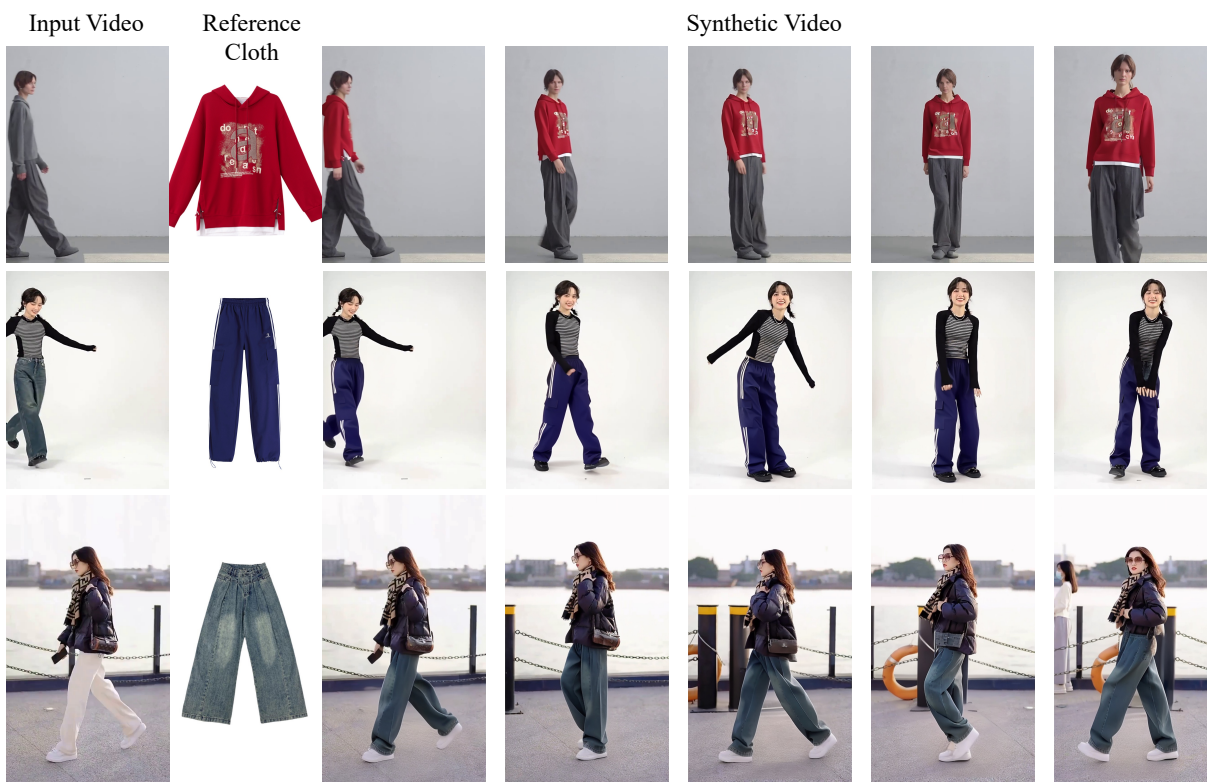


Figure 5: More examples of person movements parallel to the camera.



Figure 6: More examples of camera movements.



Figure 7: Extra results of bottoms try-on.

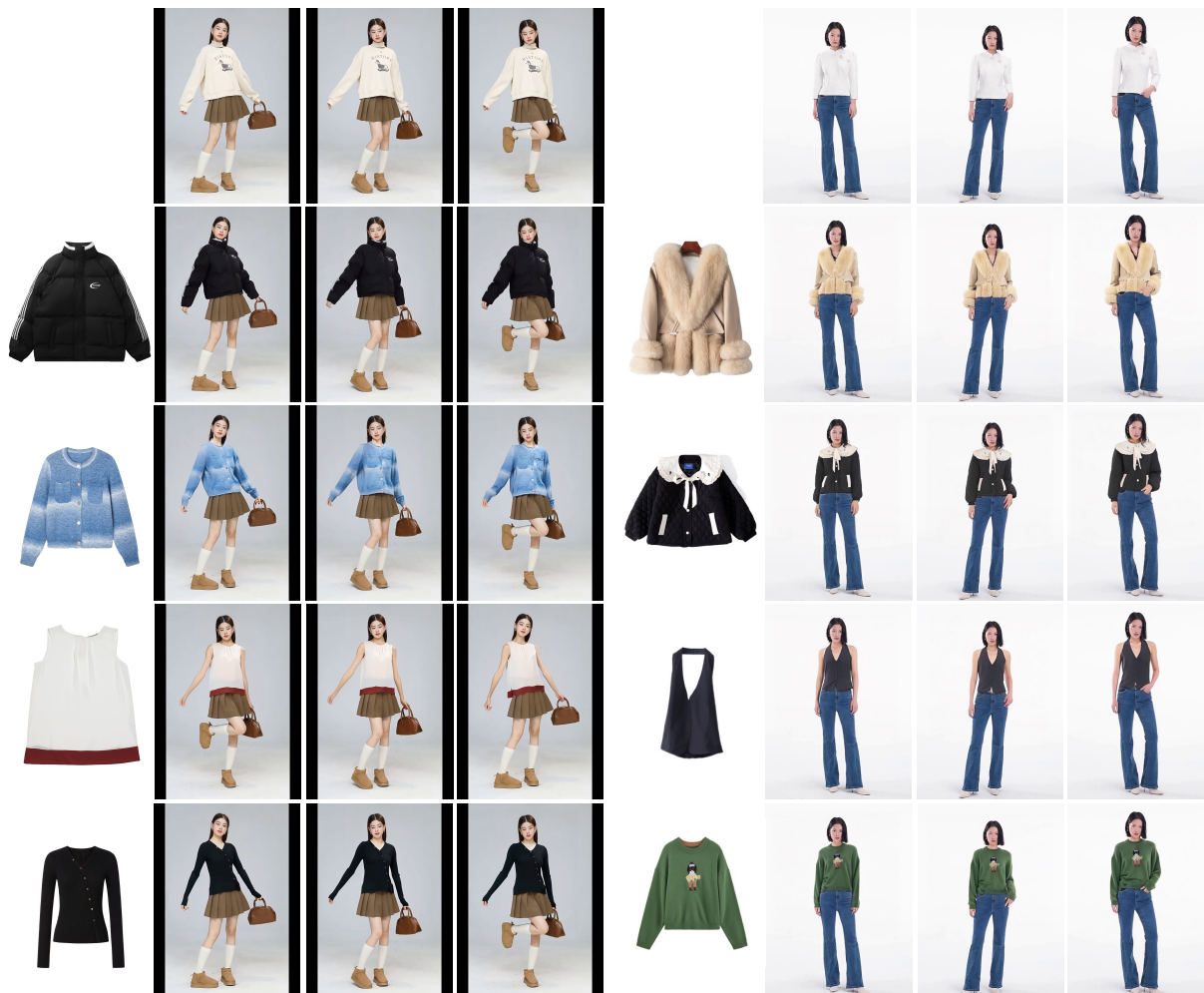


Figure 8: Extra results of tops try-on.