MOBILE-VTON: High-Fidelity On-Device Virtual Try-On

Supplementary Material

A. On-Device Deployment Details

We evaluate the practical feasibility of MOBILE-VTON under a fully on-device setting. The model is deployed on a Xiaomi 17 Pro Max equipped with a Snapdragon 8 Gen 5 chipset and Qualcomm Hexagon NPU. All experiments are conducted in Airplane Mode to ensure that no external computation or network communication is involved.

The full inference pipeline runs entirely on the mobile device. Model weights are preloaded onto the NPU, incurring negligible runtime loading overhead. GarmentNet and TryOnNet execute in parallel on the NPU, while the VAE operates on the CPU during encoding and is transferred to the NPU only for the final decoding stage. The end-to-end inference time for generating one full-resolution (1024×768) try-on result is approximately 80 seconds on the mobile NPU. This latency corresponds to a complete diffusion pipeline executed without step reduction, pruning, or system-level acceleration techniques. We adopt INT8 quantization during inference, as most Android NPUs (including Qualcomm Hexagon) require INT8 execution. In contrast, Apple A-series NPUs support BF16 inference (as reported in SnapGen), indicating that the proposed architecture is compatible with both major mobile hardware ecosystems. These results demonstrate the feasibility of deploying diffusion-based virtual try-on models fully on-device without reliance on cloud-based computation.

B. Additional Ablation Study

To better understand the contribution of each component in MOBILE-VTON, we conduct controlled ablation experiments on the DressCode upper-body setting. Specifically, we evaluate the impact of (a) the Latent Alignment (LA) module, (b) Trajectory-Consistent Garment supervision (TCG), and (c) the distillation training strategy.

As shown in Table 1, removing the LA module significantly degrades perceptual metrics (LPIPS increases from 0.088 to 0.168, SSIM drops from 0.893 to 0.827), indicating that latent-level garment alignment is crucial for preserving fine-grained appearance details. Removing TCG supervision also leads to consistent degradation across all metrics, suggesting that enforcing trajectory-consistent feature transfer across diffusion timesteps helps stabilize garment representation and reduces structural drift. Most notably, training without distillation results in severe performance collapse (FID increases from 10.211 to 113.59), demonstrating that direct training from scratch without teacher guidance fails to converge to a meaningful solution under the lightweight model capacity.

Figure 1. Qualitative ablation results on the DressCode upper-body test set. Removing key components degrades garment detail, structural consistency, or overall synthesis quality.

Table 1. Quantitative ablation results on the DressCode upper-body test set (UN setting). Each component contributes to improved perceptual quality and distribution alignment.

Method	LPIPS↓	SSIM↑	CLIP-I↑	FID↓ (UN)	KID↓ (UN)
Ours	**0.088**	**0.893**	**0.833**	**10.211**	**2.023**
w/o TCG	0.108	0.881	0.804	10.914	2.939
w/o LA	0.168	0.827	0.704	13.652	5.091
w/o Distill	0.354	0.788	0.438	113.59	106.66

Qualitative comparisons in Fig. 1 further corroborate these observations. LA improves garment texture fidelity, TCG enhances structural coherence, and distillation provides essential supervision for stable training. Together, these components form a complementary design in which each contributes to model stability and perceptual quality.

C. Discussion on GarmentNet

GarmentNet is trained under Trajectory-Consistent Garment (TCG) supervision, where real garment images are directly used as reconstruction targets across diffusion timesteps. This design provides explicit supervision from the true data distribution and promotes semantic consistency throughout the denoising trajectory. Under this training setting, introducing an additional adversarial loss for GarmentNet did not yield measurable improvements in either quantitative or perceptual metrics, while increasing training complexity and optimization instability. We therefore do not incorporate a GAN loss into the final design. This observation is in line with prior studies emphasizing the importance of training stability and supervision quality in preventing shortcut dependency and over-memorization in constrained models [5–7]. Related work also highlights that carefully conditioned learning signals are essential for robustness and cross-modal generalization in gener-

ative systems [1–4, 8–10].

D. Impact of Different Dataset Quality

While prior studies consistently report better results from fine-tuning on the VITON-HD dataset, we observe the opposite trend with our lightweight MOBILE-VTON: it achieves superior performance when fine-tuned on the DressCode dataset. To investigate how dataset quality influences the performance of lightweight virtual try-on models, we conduct fine-tuning experiments on both VITON-HD and DressCode. All models are initially fine-tuned on a combined dataset consisting of both VITON-HD and DressCode, and are subsequently fine-tuned individually on each dataset. We perform evaluations on both test sets to assess generalization ability and cross-domain robustness.

As shown in Table 2, our model achieves consistent improvements across all metrics when fine-tuned on DressCode, even when evaluated on the VITON-HD test set. For example, on VITON-HD, SSIM improves from 0.932 to 0.935, LPIPS drops by 9.3%, and CLIP-I increases by 1.56%, while realism-oriented metrics such as FID and KID also show notable gains. The performance gap is even larger when evaluating on the DressCode test set: SSIM improves by 1.82%, LPIPS by 13.9%, and KID by 9.5%.

Visual comparisons in Fig. 2 illustrate that models trained on DressCode yield sharper garment textures and more coherent alignment, particularly in fine-grained regions such as logos and sleeves. We attribute these consistent gains to the higher sensitivity of lightweight models to data quality. The DressCode dataset, with its uniform resolution, consistent garment framing, and clearer visual features, offers more stable learning signals, which are essential for on-device models with limited parameter budgets. In contrast, VITON-HD contains a mix of low- and high-quality images, with varying degrees of compression and pose ambiguity, which can degrade the training dynamics for small models. These findings underscore the importance of dataset curation and quality when designing practical, efficient mobile-based VTON systems.

E. Limitations

A key limitation of MOBILE-VTON lies in its difficulty in accurately reproducing garments with textual elements, such as logos, printed slogans, or brand names. These failures are often manifested as blurred, distorted, or partially missing characters in the generated try-on results. This issue primarily stems from two factors: first, the model is trained entirely from scratch without leveraging any large-scale pretraining on text-aware image corpora; second, garments containing prominent textual content are relatively rare in current virtual try-on datasets, which limits the model's exposure to such patterns during training. As a result, the model lacks the capacity to establish robust visual-text associations and generalize to unseen text styles or layouts. This limitation is particularly evident in cases involving stylized fonts, curved text, or small printed labels.

Figure 2. Qualitative comparison of try-on results fine-tuned on DressCode (a) and VITON-HD (b). DressCode-trained models produce sharper textures and more accurate alignment, especially in fine-grained regions (highlighted in red). Zoom in for details.

Table 2. Comparison of MOBILE-VTON fine-tuned on VITON-HD and DressCode upper-body. "UN" indicates unpaired setting; FID is multiplied by 100.

Dataset	LPIPS↓	SSIM↑	CLIP-I↑	FID↓ UN	KID↓ UN
VITON-HD					
VITON-HD	0.102	0.877	0.818	10.558	2.685
DressCode	**0.088**	**0.893**	**0.833**	**10.211**	**2.023**
DressCode Upper-body					
VITON-HD	0.058	0.932	0.832	13.527	2.473
DressCode	**0.053**	**0.935**	**0.845**	**12.775**	**1.917**

References

[1] Ziming Hong, Tianyu Huang, Runnan Chen, Shanshan Ye, Mingming Gong, Bo Han, and Tongliang Liu. Adlift: Lifting adversarial perturbations to safeguard 3d gaussian splatting assets against instruction-driven editing. *arXiv preprint*

arXiv:2512.07247, 2025. 2

[2] Zhuo Huang, Chang Liu, Yinpeng Dong, Hang Su, Shibao Zheng, and Tongliang Liu. Machine vision therapy: Multimodal large language models can enhance visual robustness via denoising in-context learning. In *Forty-first International Conference on Machine Learning*, 2024.

[3] Zhuo Huang, Gang Niu, Bo Han, Masashi Sugiyama, and Tongliang Liu. Towards out-of-modal generalization without instance-level modal correspondence. In *The Thirteenth International Conference on Learning Representations*, 2025.

[4] Runqi Lin, Bo Han, Fengwang Li, and Tongliang Liu. Understanding and enhancing the transferability of jailbreaking attacks. In *The Thirteenth International Conference on Learning Representations*, . 2

[5] Runqi Lin, Chaojian Yu, Bo Han, and Tongliang Liu. On the over-memorization during natural, robust and catastrophic overfitting. In *The Twelfth International Conference on Learning Representations*, . 1

[6] Runqi Lin, Chaojian Yu, Bo Han, Hang Su, and Tongliang Liu. Layer-aware analysis of catastrophic overfitting: Revealing the pseudo-robust shortcut dependency. In *Forty-first International Conference on Machine Learning*, .

[7] Runqi Lin, Chaojian Yu, and Tongliang Liu. Eliminating catastrophic overfitting via abnormal adversarial examples regularization. *Advances in Neural Information Processing Systems*, 36:67866–67885, 2023. 1

[8] Runqi Lin, Alasdair Paren, Suqin Yuan, Muyang Li, Philip Torr, Adel Bibi, and Tongliang Liu. Force: Transferable visual jailbreaking attacks via feature over-reliance correction. *arXiv preprint arXiv:2509.21029*, 2025. 2

[9] Yongli Xiang, Ziming Hong, Zhaoqing Wang, Xiangyu Zhao, Bo Han, and Tongliang Liu. When safety collides: Resolving multi-category harmful conflicts in text-to-image diffusion via adaptive safety guidance. *arXiv preprint arXiv:2602.20880*, 2026.

[10] Bowen Zheng, Yongli Xiang, Ziming Hong, Zerong Lin, Chaojian Yu, Tongliang Liu, and Xinge You. Vii: Visual instruction injection for jailbreaking image-to-video generation models. *arXiv preprint arXiv:2602.20999*, 2026. 2

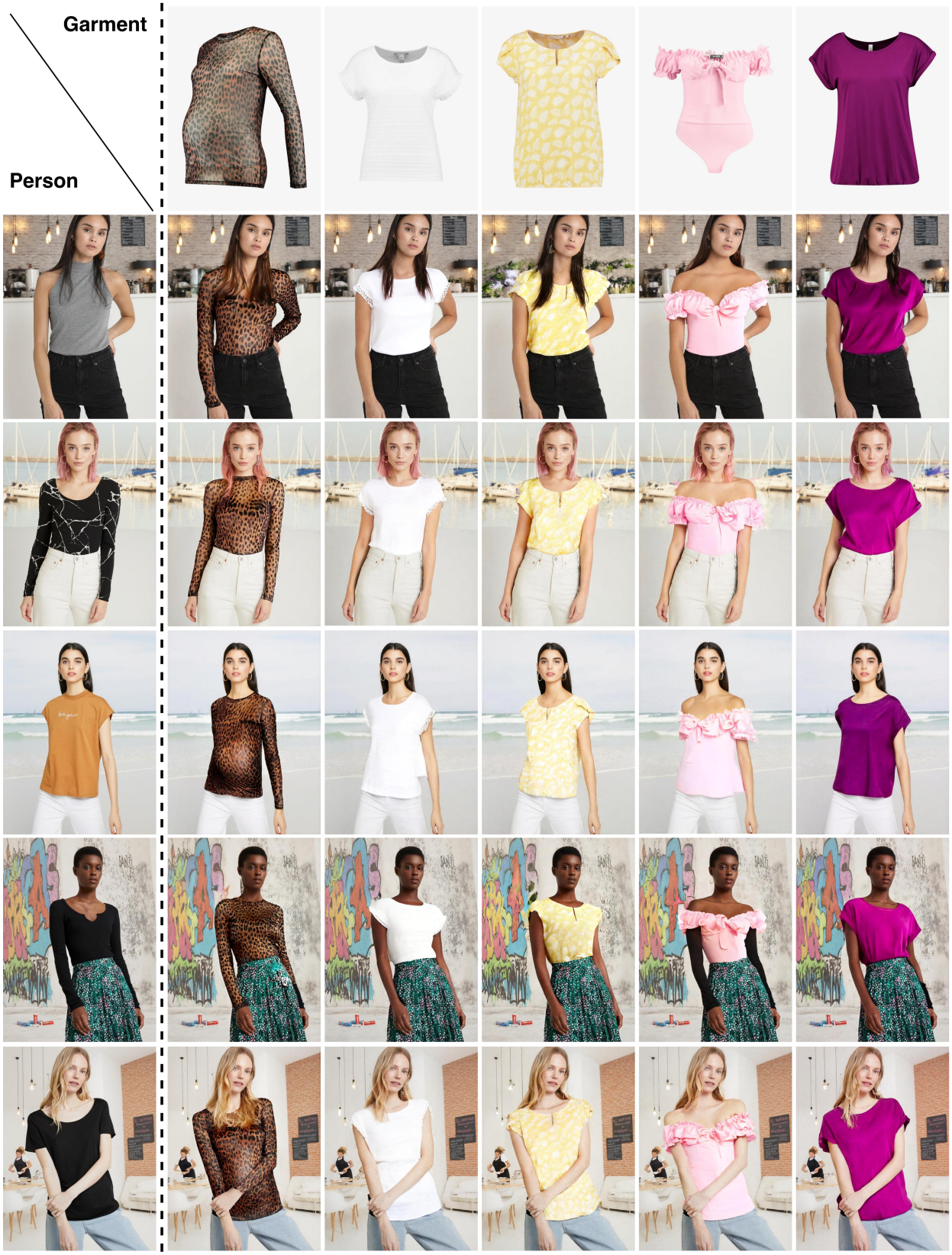

Figure 3. Try-on results on the VITON-HD In-the-Wild test set, produced by MOBILE-VTON. Zoom in for finer details.

Figure 4. Try-on results on the DressCode test set, produced by MOBILE-VTON. Zoom in for finer details.

Figure 5. Try-on results on the VITON-HD test set, produced by MOBILE-VTON. Zoom in for finer details.