PG-VTON: Single-Pass Training-Free Virtual Try-On via Patch-Guided Reference Alignment

Supplementary Material

A. Detailed Ablation Studies

A.1. Effect of PIP Duration α

To understand how sensitive PG-VTON is to the length of the priming phase, we vary the PIP duration ratio α over a range of values and report FID_p, SSIM_p, and LPIPS_p on VITON-HD at 1024×768 resolution (Table 4). Concretely, we sweep $\alpha \in \{0, 0.06, 0.1, 0.2, 0.3\}$, where $\alpha = 0$ corresponds to disabling PIP entirely. We observe a clear trend: very short priming (α close to 0) does not give the model enough time to absorb fine garment attributes such as logos, stripes, and small prints, so identity- and garment-centric metrics (e.g., FID_p, LPIPS_p) remain noticeably worse than the best setting; increasing α into a moderate range yields the strongest quantitative performance, as the garments are firmly anchored early while the later steps are still free to refine pose, shading, and background; however, pushing α too high makes the "hard-pasted" patches persist too far along the trajectory, so that the subsequent denoising cannot fully smooth them out, occasionally leading to slightly stiff silhouettes or faint boundary artifacts. Overall, these results support our design intuition that PIP should dominate only the early portion of the flow, leaving the backbone inpainting model to take over in the later steps; in the main experiments we therefore adopt a conservative default of $\alpha = 0.1$, which already provides a substantial gain over the no-PIP baseline while remaining robust across datasets and resolutions.

Table 4. Ablation on the PIP duration ratio α.

α	$\text{FID}_p \downarrow$	$\text{SSIM}_p \uparrow$	$\text{LPIPS}_p \downarrow$
0.0 (w/o PIP)	8.099	0.849	0.106
0.06	6.854	0.874	0.086
0.1	**6.749**	**0.877**	**0.086**
0.2	7.178	0.873	0.082
0.3	7.738	0.860	0.084

A.2. Effect of RAA Scaling Factor γ

Reference-Aware Attention (RAA) modulates the cross-attention between person queries and garment keys via a scalar factor γ that rescales the corresponding attention logits block S_{PG} before the softmax, and we now study how this scaling affects performance. We vary γ over a broad range, $\gamma \in \{1.0, 2.0, 3.0, 5.0, 7.0, 10.0\}$, while keeping all other hyperparameters fixed ($\alpha = 0.1$, $K = 3$, $\beta = 0.15$, $T = 50$), and report results on VITON-HD. When $\gamma = 1.0$, RAA effectively collapses to the DiT backbone with PIP only. Gradually increasing γ from 1.0 to around 3.0 consistently reduces both FID_u and FID_p and improves SSIM_p/LPIPS_p, indicating that a moderate amplification of S_{PG} helps the person tokens attend more strongly to garment features and better propagate fine textures and logos into the try-on region. Beyond this range, however, the gains saturate: very large values such as $\gamma \geq 5.0$ start to over-emphasize the garment branches, occasionally leading to over-transfer of background patterns or mild suppression of local facial and hair details. In practice, we find that PG-VTON is not overly sensitive within a reasonable range of γ, and that $\gamma = 3$ offers a good trade-off between garment fidelity and overall stability. Combined with the presence ablation in Table 3, these trends confirm that the majority of the overall improvement is attributable to PIP, while RAA serves as a complementary refinement module that further sharpens logos, prints, and edges without fundamentally changing the behavior of the system.

Table 5. Ablation on the RAA scaling factor γ for garment-to-person attention.

γ	$\text{FID}_p \downarrow$	$\text{SSIM}_p \uparrow$	$\text{LPIPS}_p \downarrow$
1 (w/o RAA)	7.382	0.852	0.100
2	6.914	0.873	0.089
3	**6.749**	**0.877**	**0.086**
5	6.897	0.875	0.086
7	7.361	0.874	0.088
10	8.902	0.871	0.094

A.3. Effect of Patch Count and Patch Scale in PIP

A natural question is whether the improvement brought by PIP mainly depends on the exact number or size of the pasted garment patches. To examine this, we vary both the patch count K and the patch scale β, while fixing the remaining hyperparameters to $\alpha = 0.1$, $\gamma = 3$, and $T = 50$. Patch centers are sampled uniformly at random within the garment mask. Specifically, we consider $K \in \{0, 1, 2, 3, 4, 5\}$ and $\beta \in \{0.10, 0.15, 0.20\}$, where the patch side length is defined as $s_{\text{patch}} = \beta \cdot S_{\text{short}}$. The results are reported in Table 6.

The case $K = 0$ corresponds to the RAA only baseline without explicit patch anchoring. Compared with this baseline, introducing PIP consistently yields a clear improvement on all garment-centered metrics, confirming that the key benefit comes from injecting localized garment identity

Table 6. **Ablation on patch count K and patch scale β in PIP.** Once patch anchoring is introduced, the performance remains stable across a reasonable range of K and β, indicating that PIP is not sensitive to the exact patch number or size.

K	β	$\text{FID}_p \downarrow$	$\text{SSIM}_p \uparrow$	$\text{LPIPS}_p \downarrow$
0 (RAA only)	–	8.099	0.849	0.106
1	0.15	6.869	0.875	0.087
2		6.821	0.875	0.088
3		6.749	0.877	0.086
4		6.841	0.874	0.089
5		6.946	0.874	0.089
3	0.10	6.751	0.855	0.088
	0.15	6.749	0.877	0.086
	0.20	6.884	0.853	0.088

cues at an early stage. By contrast, once PIP is enabled, the performance differences across different values of K and β remain relatively small. This suggests that the effectiveness of PIP is not tied to a particular patch count or patch size, but rather to the existence of a lightweight spatial cue that guides the model toward the correct garment identity and local appearance.

A.4. Effect of Sampling Steps in a Single-Pass Trajectory

One practical advantage of PG-VTON is that it achieves strong visual quality in a single diffusion trajectory, without relying on unusually long sampling schedules or multiple passes. To verify that our method does not secretly depend on an excessive number of steps, we vary the number of denoising steps S used at inference time while keeping the controller configuration fixed ($\alpha = 0.1$, $\gamma = 3$, $K = 3$, $\beta = 0.15$) and evaluate on VITON-HD. We sweep $S \in \{10, 20, 30, 40, 50\}$ and summarize the results in Table 7. Thanks to the efficiency of the underlying rectified-flow backbone, PG-VTON already reaches competitive or near-saturating performance at relatively low step counts, the garment-aware metrics are close to those of the default $S = 50$ setting, and visual inspections show that garment identity, pose consistency, and background coherence are largely preserved. Increasing S beyond 30 brings only marginal improvements, with diminishing returns in both FID_p and LPIPS_p, while naturally incurring higher runtime. These results confirm that PG-VTON's gains are not a by-product of very long sampling trajectories and that the proposed patch-guided controller can effectively steer the generation process even under tight step budgets, making the method attractive for practical applications where inference cost is a concern.

Table 7. Ablation on sampling steps S.

step S	$\text{FID}_p \downarrow$	$\text{SSIM}_p \uparrow$	$\text{LPIPS}_p \downarrow$
10	7.260	0.877	0.087
20	6.965	0.876	0.087
30	6.914	0.876	0.086
40	6.824	0.875	0.086
50	**6.749**	**0.877**	**0.086**

B. Extended Comparisons

B.1. Runtime and Memory Analysis

We compare the computational efficiency of PG-VTON against the training-free OmniVTON baseline on a single NVIDIA A100 GPU. As summarized in Table 8, both methods generate 1024×768 outputs, but PG-VTON requires only a *single* diffusion trajectory, whereas OmniVTON relies on a multi-stage pipeline with multiple diffusion calls and compositing steps. This single-pass design leads to a substantially lower average per-image latency (39.5 s vs. 70.8 s) and reduces peak VRAM usage from 45 GB to 34 GB, while using the same backbone resolution and similar sampler settings. In practice, this translates into noticeably faster inference and a smaller memory footprint, making PG-VTON easier to deploy in resource-constrained environments and more scalable to high-resolution generation.

Table 8. **Efficiency Analysis.** Comparison of inference stages, time, and computational cost.

Method	Stages	Time (s)	VRAM (GB)
OmniVTON	Multi	70.8	45
PG-VTON (Ours)	**Single**	**39.5**	**34**

B.2. Quantitative Evaluation on Subject Insertion

We further evaluate our method on the COCO-EE benchmark to assess its generalization beyond standard virtual try-on settings. We compare against representative methods reported on the public leaderboard, including PBE, ObjectStitch, AnyDoor, and ControlCom. As reported in Table 9, despite requiring no task-specific training, our method achieves competitive zero-shot performance, with an FID of 3.32 and a CLIP Score of 85.14. This result is notable given that the compared baselines are specialized methods designed for image editing or compositional generation.

B.3. Low-Resolution Results

We further verify the robustness of PG-VTON at lower resolutions by evaluating on VITON-HD at 512×384, following

Table 9. Quantitative comparison on the COCO-EE benchmark.

Method	**FID** ↓	**CLIP Score** ↑
PBE	3.18	84.84
ObjectStitch	3.35	85.97
AnyDoor	3.60	89.70
ControlCom	3.19	88.31
Ours (Training-free)	3.32	85.14

the standard setting used by OmniVTON and prior VTON baselines. Table 10 reports a comparison with exemplar-based editing methods (PBE, AnyDoor, TIGIC, Cross-Image), supervised VTON systems (GP-VTON, CAT-DM, D^4-VTON, IDM-VTON), and the training-free OmniVTON. PG-VTON achieves the best garment-centric scores (lowest FID_p and $LPIPS_p$, highest $SSIM_p$) while maintaining competitive holistic quality as measured by FID_u, despite using a single inference pass and no task-specific finetuning. These results show that the proposed patch-guided controller remains effective under reduced spatial resolution, and that its advantages over both supervised and training-free baselines are not limited to a particular scale or evaluation protocol.

Table 10. Quantitative comparison with state-of-the-art methods. The best results are highlighted in **bold**.

Method	FID_u ↓	FID_p ↓	$SSIM_p$ ↑	$LPIPS_p$ ↓
PBE	19.230	17.649	0.784	0.227
AnyDoor	14.830	9.922	0.796	0.164
TIGIC	90.338	88.900	0.613	0.422
Cross-Image	62.614	57.286	0.760	0.256
GP-VTON	51.566	49.196	0.810	0.249
CAT-DM	28.869	26.339	0.775	0.229
D^4-VTON	25.299	23.914	0.790	0.250
IDM-VTON	23.035	20.460	0.812	0.147
OmniVTON	**9.621**	7.758	0.832	0.145
Ours	9.967	**7.296**	**0.853**	**0.082**

C. Robustness and Randomness Analysis

PIP relies on a simple bounding-box-based mapping from garment patches to the person canvas (Sec. 3.2), which is intentionally coarse and does not attempt to perfectly align fine-grained geometry. As a result, at very early time steps (e.g., $t \approx 0.1T$) the pasted patches may appear slightly misaligned or "floating" with respect to the underlying body structure, as shown in the timestep visualizations of Fig. 7. However, once the priming window ends (i.e., $t > \alpha T$) and PIP is turned off, the subsequent evolution is dominated by the strong inpainting prior of FLUX.1-Fill together with the cross-attention modulation from RAA. These components gradually diffuse and transport the injected textures towards semantically appropriate regions (e.g., from a roughly placed torso patch to the actual clothing area defined by the person mask), while suppressing inconsistent artifacts along the way. Empirically, we observe that even when we deliberately enlarge the garment bounding box or jitter the mapping, the final outputs remain visually plausible and free of obvious patch boundaries, indicating that the backbone-plus-controller system exhibits a self-correcting behavior: coarse geometric hints from PIP are sufficient to lock in garment identity early, and the later denoising steps refine geometry, shading, and boundaries so that the final try-on result is not overly sensitive to small errors in the bounding-box mapping.

Figure 7. Self-Correcting Behavior.

D. Additional Visual Results and Limitations

Additional qualitative results. We provide more qualitative comparisons in Figs. 8 and 9.

Failure cases and limitations. Despite these strong results, our method still exhibits several limitations. As shown in Figs. 10 In some cases with weak or ambiguous reference signals—such as highly reflective fabrics, very subtle prints, or poor segmentation around the clothing boundary—the mask-adjacent garment features can leak and dominate the synthesized region, causing slight bleeding of colors or textures outside the intended clothing area and occasionally over-smoothing local body details. Moreover, because PIP relies on coarse bounding-box mappings rather than precise flow fields, the patch locations can be imperfect when the pose is extremely articulated or when the garment has complex topology, leading to rare but noticeable artifacts around sleeves, hems, or layered garments. These issues suggest promising directions for future work, including more adaptive and fine-grained mask refinement, learning stronger spatial priors for patch placement, and combining PIP with structure-aware correspondence cues to further improve robustness in the most challenging real-world settings.

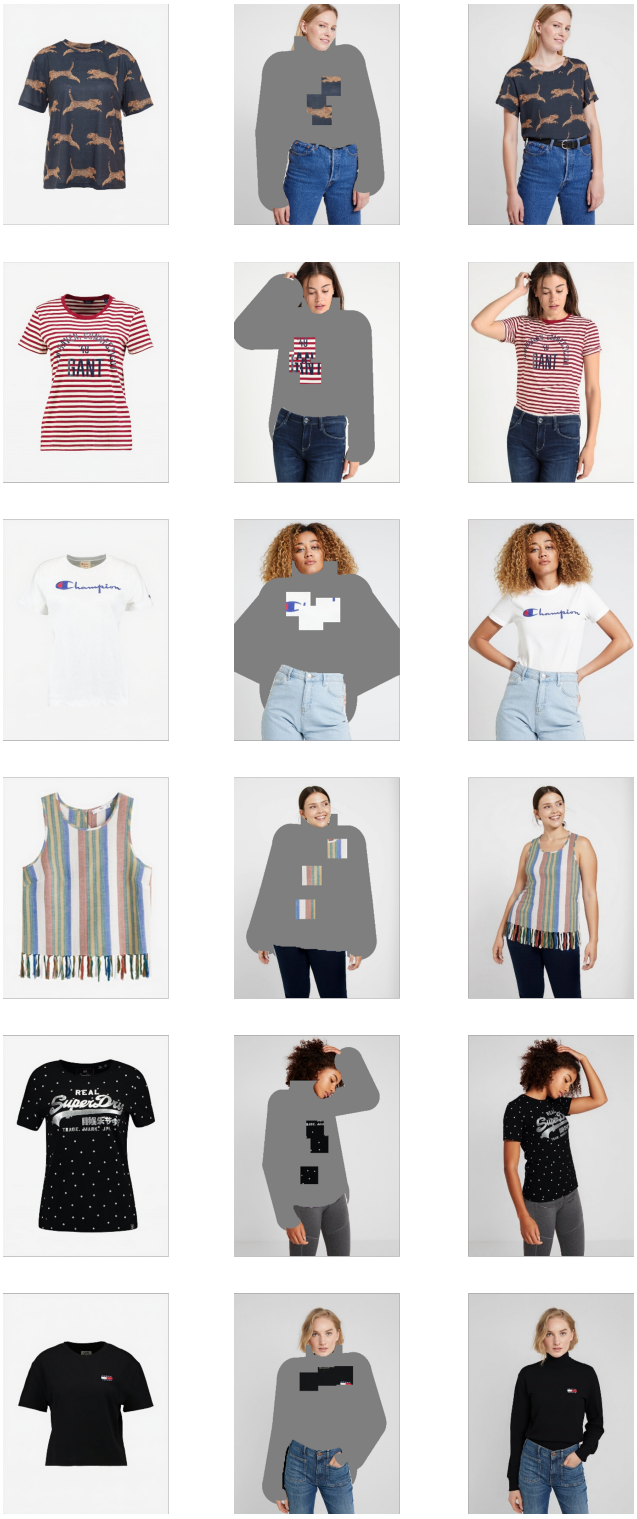

Figure 8. Additional Qualitative Results on VITON-HD.

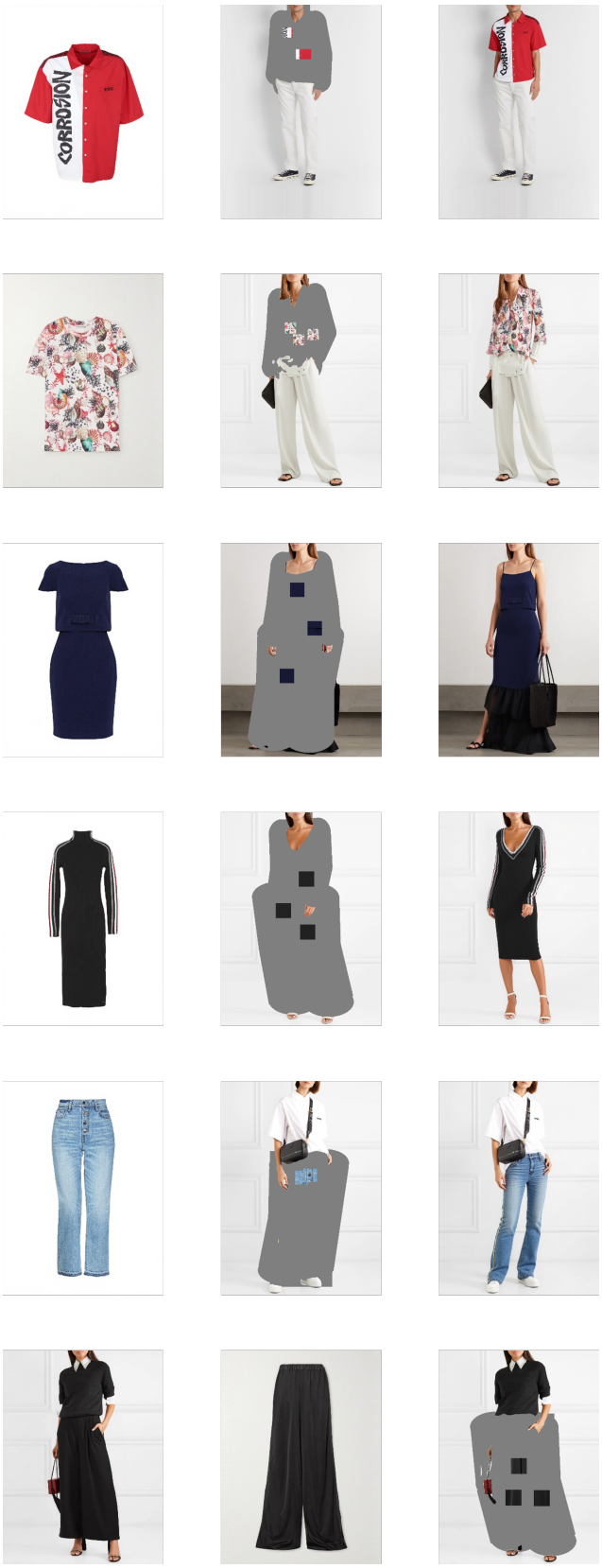

Figure 9. Additional Qualitative Results on Dress Code.

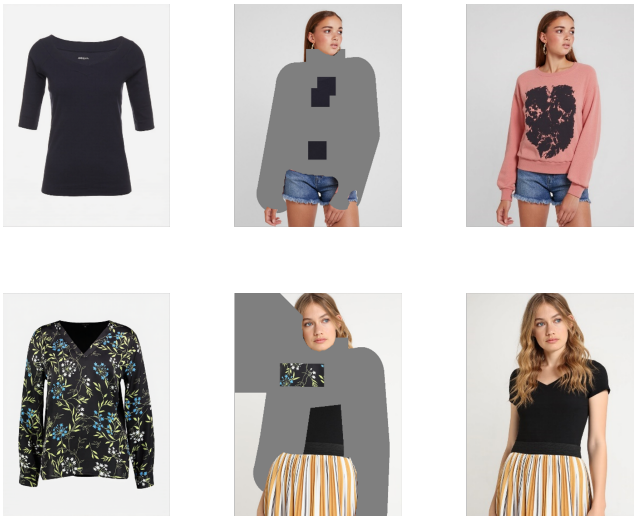

Figure 10. Failure Cases