ART-VITON: Hard Measurement-Guided Trajectory Alignment in Latent Diffusion for Artifact-Free Virtual Try-On

Anonymous submission

Abstract

Virtual try-on (VITON) aims to generate realistic images of a person wearing a target garment but still faces two key challenges: garment alignment and blending artifacts. GAN-based methods often suffer from training instability and noticeable visual artifacts, whereas diffusion models offer more reliable generation but still struggle with visible seams along the mask boundaries. We model VITON as a linear inverse problem and propose a novel inverse solver integrated into latent diffusion models to enforce measurement consistency during generation. To stabilize sampling, we inject training residuals into the initial latent, reducing semantic drift. Our solver combines latent interpolation and frequency correction to preserve garment semantics and restore fine textures. Our framework is a generalizable, architecture-agnostic method that manipulates latent trajectories externally without modifying the underlying LDM architecture, enabling easy integration and broad applicability. Experiments show significant reductions in artifacts and improved visual realism with faithful adherence to measurement constraints. **(Which datasets??)**

Introduction

Virtual try-on (VITON) aims to synthesize realistic images of a person wearing a target garment, offering a powerful tool for enhancing the online shopping experience. Given a clothing image and a target person image, image-based VITON seeks to generate a new image in which the person appears naturally dressed in the given garment. As e-commerce expands, customers gain a clearer and more intuitive sense of how clothes would look on them, reducing uncertainty in purchases. However, achieving high-quality VITON remains a challenging task due to the need to fit unaligned garments to the person while preserving identity and non-try-on regions (e.g., face, hair, background).

Early VITON methods based on GANs have been widely adopted, typically consisting of a warping module that aligns the garment to the target person and a generator that blends the warped garment into the image. However, these methods suffer from several limitations: the output quality is highly sensitive to warping accuracy, GANs are prone to training instability and mode collapse, and generalization is hindered due to limited garment-person pair diversity in existing datasets (e.g., VITON, VITON-HD, DressCode). Consequently, when blending measurements (i.e., known ground-truth regions) with generated regions to preserve visual quality in reliable areas (e.g., face, hair, background), artifacts such as color, lighting, and texture mismatches frequently appear near boundaries.

Diffusion-based models have recently emerged as a strong alternative, offering better distribution coverage, stable training, and scalability. They can be broadly categorized into two-stage and one-stage pipelines. Two-stage approaches use a separate warping module before generation, structurally resembling GAN-based models. In contrast, one-stage methods eliminate external warping by either encoding garment features into the model (e.g., LoRA, Textual Inversion) or guiding generation through conditional control (e.g., ControlNet, IP-Adapter). While these methods improve garment fidelity, most still fall short in handling background and measurement-constrained regions, often producing artifacts at the boundary between generated and real content. As shown in Fig. 1, these boundary artifacts suggest that conditioning alone is insufficient, and more explicit structural mechanisms are required. Similar issues are also commonly observed in GAN-based methods.

To address boundary artifacts arising from imperfect alignment and measurement inconsistency, we integrate an inverse solver directly into the denoising process. Rather than relying on post-hoc projection—which often introduces abrupt shifts and degrades visual fidelity—the inverse solver iteratively imposes measurement constraints at each generation step, guiding the sampling trajectory to satisfy constraints while maintaining contextual coherence. However, directly altering the sampling path can destabilize generation. To ensure stable and consistent synthesis, we mitigate the training-inference mismatch by injecting residual signals into the initial latent, aligning the inference process with training dynamics, and effectively reducing semantic inconsistencies and visual artifacts.

Furthermore, since simply strengthening constraints can exacerbate semantic drift, we enhance the inverse solver by combining prior-preserving latent interpolation with frequency-level correction. Latent interpolation preserves essential clothing features by reducing semantic drift, while frequency correction restores fine textures lost in pixel-to-latent transformations. This hybrid approach greatly improves boundary sharpness and texture fidelity, producing artifact-free outputs closely aligned with measurements. Im-

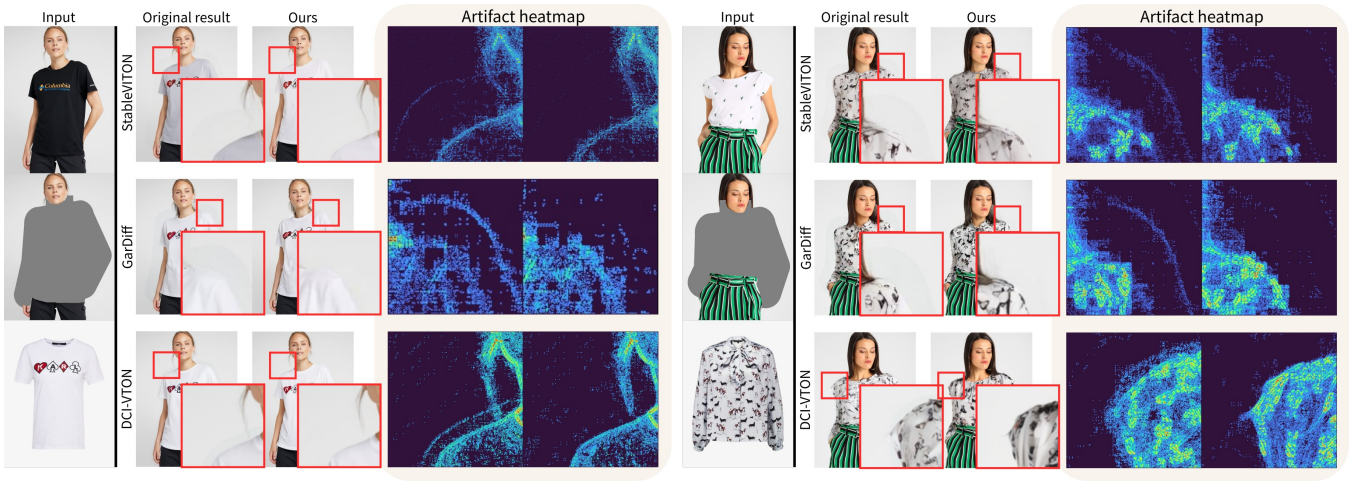

Figure 1: Comparison of artifact heatmaps for three models (DCI-VTON, GarDiff, StableVITON) on the VITON-HD dataset (proposed method vs. baselines). The red box highlights artifact regions, and the heatmap visualizes the gradient magnitude within these areas. Post-processing by projecting measurements onto generated images creates visible boundaries between real and fake regions and degrades visual quality.

portantly, our framework operates externally without modifying the underlying LDM architecture, offering a generalizable solution that markedly enhances VITON's visual realism and structural consistency (Fig. 1).

Related work

Image-Based VITON

Early virtual try-on methods predominantly adopt GAN-based two-stage frameworks, where explicit warping modules align garments to target bodies before synthesis. Pioneering works such as VITON and CP-VTON use TPS transformations or learned geometric matching to improve spatial alignment, while VITON-HD, HR-VITON, and GP-VTON extend this pipeline to high-resolution generation and finer boundary control. Despite their progress, these approaches suffer from a critical reliance on warping quality—where even minor warping errors can lead to unrealistic results—and often struggle to capture fine garment details or natural folds.

More recently, latent diffusion models (LDMs) have emerged as powerful alternatives, offering improved generation quality and flexible conditioning mechanisms. Methods like DCI-VTON and FLDM-VTON feed warped garments into the diffusion process as local conditions, reducing detail loss, while LaDI-VTON employs textual inversion for garment personalization. Approaches such as IDM-VTON, BooW-VTON, and StableVITON utilize adapter modules (e.g., ControlNet, IP-Adapter, Parallel U-Net) to better preserve semantics and style. Others like GarDiff and Diffuse-to-Choose tackle spatial encoding limitations through VAE encoders, GF adapters, and pixel-level hints. Our method complements these advances by introducing a model-agnostic inverse solver that operates independently of the LDM architecture, enabling broad compatibility and effective artifact reduction across various try-on frameworks.

Diffusion Inverse Solvers

Diffusion-based inverse solvers extend classical inverse problem methods by incorporating measurement constraints into the denoising process, aiming to generate outputs consistent with observed measurements. Early approaches enforced these constraints through hard projection. More recent solvers fall into two categories depending on how they impose measurement conditions: Diffusion Posterior Sampling (DPS) adjusts the reverse diffusion trajectory using measurement gradients to maintain consistency between observed and generated regions, while Measurement-Constrained Gradient (MCG) further enforces constraints by projecting outputs onto the measurement subspace. However, MCG's reliance on hard replacement can lead to semantic misalignment and boundary artifacts.

These ideas have been further developed in Latent Diffusion Models (LDMs), which combine latent representations with decoded pixel outputs. PSLD applies pixel-space measurement gradients to latent variables, extending DPS to the latent domain, while Resample replaces decoded measurements and reintroduces noise in an MCG-style manner. TReg and DreamSampler adopt a two-stage strategy that alternates between pixel-space optimization for data consistency and latent-space interpolation for semantic alignment. Building on these insights, we propose a new MCG-based solver tailored for LDMs that enforces pixel-level consistency while refining latent variables to preserve semantic structure and texture. This approach reduces artifacts and achieves better alignment between measured and generated regions.

Preliminaries

Latent Diffusion Model

Latent Diffusion Models (LDMs) apply the diffusion process in a compressed latent space rather than in the pixel

space for efficiency. Given an image $\mathbf{x}_0$, it is first encoded into a latent representation $\mathbf{x}_0 = \mathcal{E}_\varphi(\mathbf{x}_0)$ using a pretrained encoder $\mathcal{E}_\varphi$, and the final output is obtained by decoding $\mathbf{z}_0$ through a decoder $\mathcal{D}_\varphi$.

The forward diffusion gradually perturbs $\mathbf{z}_0$ to $\mathbf{z}_t$ using a predefined noise schedule, and the reverse process is modeled by a denoising network $\epsilon_\theta(\mathbf{z}_t, t)$ that approximates the added noise. Based on Tweedie's formula, the posterior mean $\hat{\mathbf{z}}_0$ can be estimated as:

$$\hat{\mathbf{z}}_0^{(t)} = \mathbb{E}[\mathbf{z}_0|\mathbf{z}_t] = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{z}_t - \sqrt{1-\bar{\alpha}_t}\cdot\epsilon_\theta(\mathbf{z}_t, t)\right) \quad (1)$$

$$\mathbf{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\cdot\hat{\mathbf{z}}_0^{(t)} + \sqrt{1-\bar{\alpha}_{t-1}}\cdot\epsilon_\theta(\mathbf{z}_t, t) \quad (2)$$

The above equation represents a single deterministic update of the DDIM sampling process, which enables efficient and consistent generation in latent diffusion models by eliminating the stochasticity of traditional sampling methods.

Linear Inverse Problem

A linear inverse problem involves recovering an unknown original signal $\mathbf{x} \in \mathbb{R}^n$ from measurements $\mathbf{y} \in \mathbb{R}^m$ obtained by applying a linear operator $\mathcal{A} \in \mathbb{R}^{m\times n}$ to $\mathbf{x}$, possibly with additive noise. This relationship is formally expressed as:

$$\mathbf{y} = \mathcal{A}\mathbf{x} + \mathbf{n}, \quad \mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}),$$

where $\mathbf{n}$ denotes additive Gaussian noise with variance σ^2. The form of the linear operator $\mathcal{A}$ depends on the task. For example, in image inpainting, $\mathcal{A}$ acts as a masking operator selecting a subset of pixels; in super-resolution, it performs downsampling; and in deblurring, it corresponds to convolution with a blur kernel.

In our work on VITON, since the model is trained with noise-free original measurements, we consider a formulation where the generation process is conditioned on these clean measurements.

Method

We tackle the task of reference-based image inpainting for virtual try-on using a Latent Diffusion Model (LDM). The goal is to synthesize a realistic output image $\mathbf{x}_0$ that (i) faithfully preserves the visual attributes of a given reference image $\mathbf{c}$, (ii) adheres to a target measurement $\mathbf{y}$, and (iii) maintains overall visual consistency and plausibility. This task is especially challenging due to the need to harmonize both structure and style while avoiding artifacts, particularly along the masked regions guided by $\mathbf{y}$. To address this, we formulate the problem within both the latent and pixel space of a pretrained LDM, where generation is driven by iterative denoising and refined to enforce measurement constraints without sacrificing semantic alignment.

Prior-Based Initialization

According to prior studies (Choi et al. 2022; Lin et al. 2024), the noise schedule of Stable Diffusion, a representative latent diffusion model, does not guarantee a signal-to-noise ratio (SNR) of zero at the final training timestep, leaving residual signals and causing train-test discrepancies. This issue is further exacerbated in models such as StableVITON, DCI-VTON, and GarDiff, as well as in the official DDIM implementation, which do not start denoising from the final timestep (T=999). To mitigate this mismatch, SeeSR and PASD perform super-resolution by mixing low-quality (LQ) inputs with pure noise at inference. Similarly, DCI-VTON and StableVITON construct $\mathbf{z}_T$ by combining pure noise with either real signals or warped predictions from the warping module. A straightforward approach is to replace the measurement region with noisy observations. However, this fails to preserve semantic coherence with the pure noise outside the measurement region.

To address this, we first generate a prior by applying a single DDPM denoising step to pure Gaussian noise at the final timestep (T=999). This prior aligns with the training distribution and helps bridge semantics between the measurement and masked regions. We then reuse it as the initialization at T=999 to begin the full denoising process. DDPM is used for this initial step because its stochastic nature enables the model to produce a richer and more flexible prior from pure noise, whereas DDIM—faster and deterministic—is used for the subsequent full sampling process.

Model	SSIM ↑	PSNR ↑	LPIPS ↓	FID ↓	KID ↓
DCI-VTON	0.8607	23.6629	0.0852	12.6386	0.0014
+ Prior @ T = 999	**0.8880**	**24.1447**	**0.0782**	**11.4713**	**0.0011**
GarDiff	0.8062	21.1075	0.1016	11.7048	0.0061
+ Prior @ T = 999	**0.8448**	**21.8611**	**0.0864**	**10.5322**	**0.0034**
StableVITON	0.8550	23.1214	0.0835	10.8716	0.0022
+ Prior @ T = 999	**0.8552**	**23.1475**	**0.0833**	**10.4362**	**0.0014**

Table 1: Comparison of prior-based initialization at T=999 across models. Applying the prior improves performance consistently.

(B) and (C)

Characteristics and problems of existing inverse solvers: artifacts and stochastic trajectory

Ours: Measurement constraint → problem: artifact → Interpolation and restoring high-freq

(Requires a linear inverse problem description!!!)

Experiments

z_T, T **configuration for all models**

vs Existing inverse solvers

vs Baselines

vs Ablation 1: various z_T

vs Ablation 2: $pure \rightarrow (A), (B), (C)$

Conclusion

References

Choi, J.; Lee, J.; Shin, C.; Kim, S.; Kim, H.; and Yoon, S. 2022. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11472–11481.

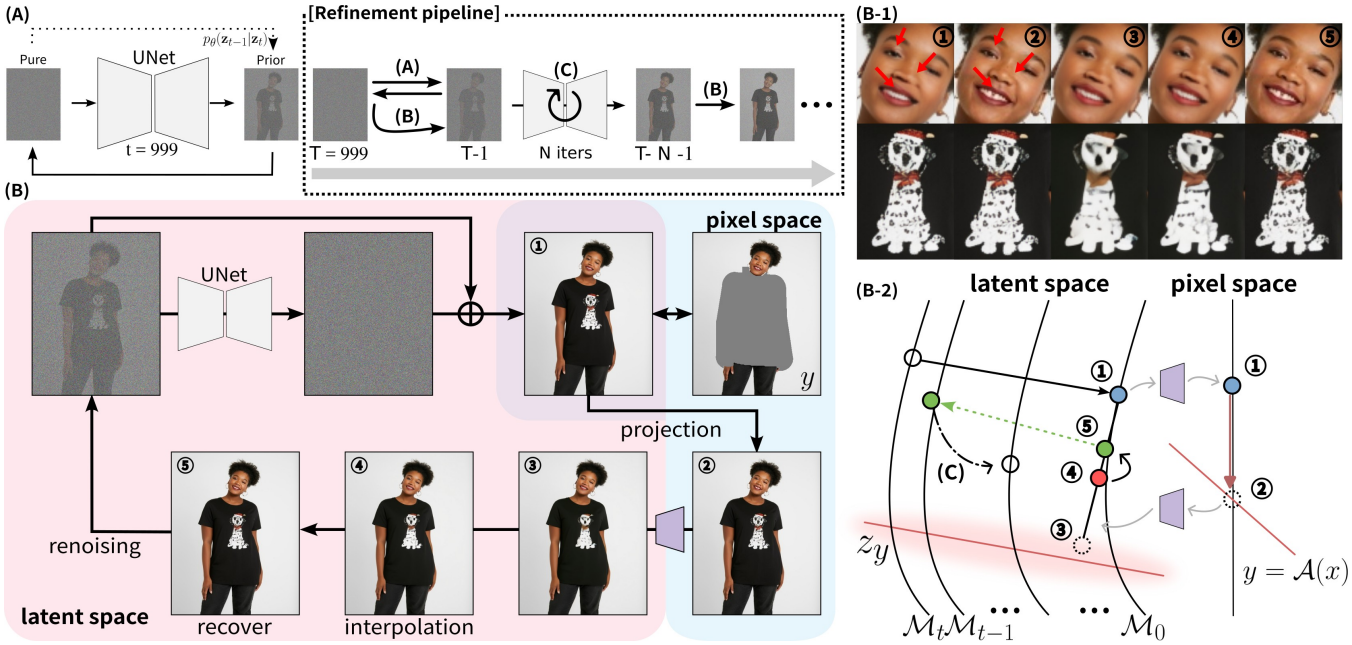

Figure 2: Overview of the proposed inverse-solver-guided refinement pipeline. (A): The latent is initialized with a residual prior at t=999 to reduce training-inference mismatch and align the inference trajectory with training dynamics. (B): During denoising, an inverse solver iteratively enforces measurement constraints while preserving semantic coherence. (B-1): The Tweedie-estimated data (①) contains clothing details but lacks facial fidelity. Projection into the measurement subspace at ② introduces high-frequency details (e.g., face), but re-encoding (②→③) causes detail degradation, which is recovered via latent interpolation (④) and frequency correction (⑤). (B-2): The latent sampling trajectory is gradually aligned with the data manifold $\mathcal{M}_0$ via interpolation (④), frequency correction (⑤), and constraint injection (C), following the degraded encoding step (③). (C): Constraints are injected intermittently to respect LDM priors, allowing natural reconciliation between masked and measured regions.

Method	SSIM ↑	PSNR ↑	LPIPS ↓	FID ↓	KID ↓
HR-VITON	0.8710	22.3368	0.0986	11.7301	0.3926
GP-VTON	0.8718	23.6485	0.0838	12.0564	0.0029
LaDI-VTON	0.8779	22.7451	0.0876	10.5203	**0.0004**
DCI-VTON	0.8871	24.1413	0.0782	11.3634	0.0012
GarDiff	0.8418	21.7263	0.0895	10.5858	0.0042
StableVITON	0.8839	23.5965	0.0757	9.8694	0.0016
DCI-VTON (Ours)	**0.8946**(+0.85%)	**24.6903** (+13.64%)	**0.0722** (+7.67%)	10.5408	0.0005
GarDiff (Ours)	0.8463	21.9647	0.0866	10.3414	0.0036
StableVITON (Ours)	0.8859	23.7027	0.0746	**9.7669**	0.0009

Table 2: Quantitative comparison on various virtual try-on models. Our method is model-agnostic and can be seamlessly applied to diffusion-based architectures (bottom), consistently improving both perceptual and fidelity metrics.

Lin, S.; Liu, B.; Li, J.; and Yang, X. 2024. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 5404–5411.

$\mathbf{z}_T$@ T = 999	SSIM ↑	PSNR ↑	LPIPS ↑	FID ↓	KID ↓
Pure	0.855	23.1214	0.0835	10.4349	0.0014
Pure (51 step)	0.855	23.1363	0.0834	10.4451	**0.0012**
Unmasked	**0.8566**	**23.2551**	0.0834	10.6985	0.0016
Offset noise	0.8425	22.1414	0.0962	**10.3335**	0.0015
Prior (DDIM)	0.855	23.1299	0.0835	10.4631	0.0015
Prior (DDPM) (Ours)	0.8552	23.1475	**0.0833**	10.4362	0.0014

Table 3: Quantitative comparison of $\mathbf{z}_T$ configurations at $T = 999$. Our proposed Prior (DDPM) method achieves a good balance, delivering strong performance across all metrics.