# WearWow: Native 2K Multi-Garment Virtual Try-On via Adaptive Token Packing and Preference Alignment

Xujie Zhang[1⋆], Runyan Du[2⋆], Song Chang[2], Jiang Li[2], Dongliang Shao[2], Liping Wu[2], Luo Wei[2], Xiaochao Qu[2], Luoqi Liu[2], and Xiaodan Liang[1†]

[1] Shenzhen Campus of Sun Yat-sen University, China
zhangxj59@mail2.sysu.edu.cn, liangxd9@mail.sysu.edu.cn
[2] Meitu Lab, China
{dry, cs2, lj28, sdl, wlp1, lw20, qxc, llq5}@meitu.com

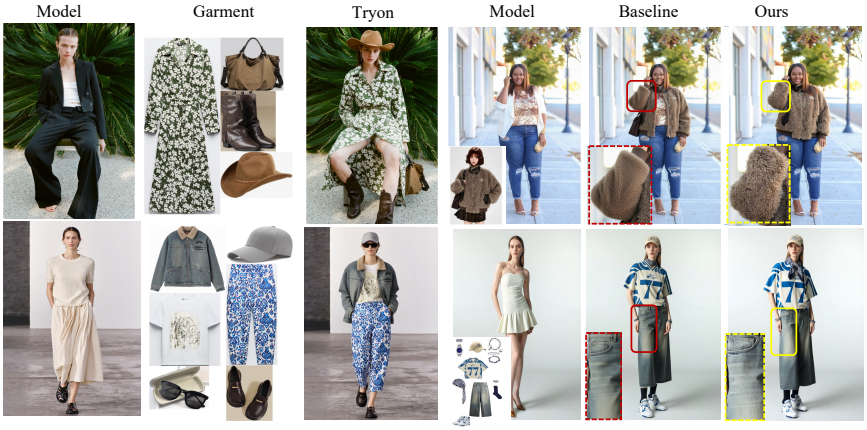


**Fig. 1: Native 2K Multi-Garment Try-On with WearWow.** (Left) Mask-Free Joint Synthesis: WearWow seamlessly composes diverse categories (garments, footwear, accessories) with strict topological alignment. (Right) Tactile Realism: While standard baselines suffer from severe over-smoothing, our framework explicitly restores micro-level high-frequency textures (e.g., curly fleece, denim), overcoming plastic-like degradation to achieve unprecedented material fidelity.

**Abstract.** Synthesizing native 2K multi-garment virtual try-on is a formidable frontier in digital fashion, critically bottlenecked by two fundamental limitations: the $\mathcal{O}(N^2)$ memory explosion induced by 2k conditions, and the spectral bias of diffusion models that over-smooths high-frequency fabric details. We present WearWow, an end-to-end, mask-free generative framework that pioneers ultra-high-resolution multi-garment synthesis. To mitigate the memory explosion , we propose Adaptive 2D Token Packing (ATP). ATP leverages inherent garment sparsity to algorithmically pack heterogeneous items onto a unified 2D canvas and

⋆ Equal contribution.
† Corresponding author.

prune uninformative background tokens, minimizing the effective sequence length and subsequent memory overhead while rigorously preserving 2D spatial priors. To rectify texture degradation, we introduce the Multi-dimensional Try-on Reward (MTR) system. MTR synergizes a Semantic Guidance Reward to explicitly drive tactile restoration with a Cloth Distribution Reward to implicitly anchor the physical distribution, a joint formulation that effectively mitigates the severe reward hacking. Furthermore, we curate WearWow-2K, an extreme-quality dataset comprising native 2K triplets, providing physically correct spatial interactions that naturally empower the model's mask-free generation. Extensive experiments demonstrate that WearWow establishes a new state-of-the-art, exceeding existing commercial baselines in native 2K multi-garment synthesis.

**Keywords:** Multi-garment Try-on · Native 2K Resolution · Mask-free Generation · Preference Optimization

## 1 Introduction

Image-based Virtual Try-On (VTON) [2, 13, 14, 24, 42] is undergoing a profound paradigm shift: evolving from the macroscopic pursuit of "structural alignment"—typically constrained to 512p or 1K resolutions to ensure correct pose warping—towards the industrial-grade demand for "tactile realism" at ultra-high resolutions (2K and beyond). Driven by the rapid evolution of generative diffusion models, recent state-of-the-art architectures have demonstrated remarkable success in preserving complex human poses and single-garment semantics. However, when transitioning to authentic commercial deployments, which strictly necessitate native 2K resolution combined with multi-garment joint generation (e.g., simultaneously synthesizing inner-wear, outerwear, and bottoms) and mask-free inference, existing end-to-end frameworks encounter severe systemic bottlenecks. As the generative objective fundamentally shifts from merely "wearing it correctly" to "rendering authentic physical textures," current paradigms face a significant degradation, bottlenecked by both physical hardware limitations and theoretical generative flaws.

Specifically, scaling diffusion-based VTON to the 2K multi-garment regime is obstructed by two primary scaling barriers. The first is the computational bottleneck driven by sequence length. In modern U-Net or Diffusion Transformer backbones, the computational complexity of self-attention mechanisms scales quadratically with the total number of visual tokens. Alongside the fixed base tokens $N_{\text{base}}$ required for the target and model images, naively concatenating $K$ high-definition reference garments (each containing $N$ tokens) directly expands the total sequence length to $N_{\text{base}} + K \cdot N$. This triggers an intractable computational cost of $\mathcal{O}((N_{\text{base}} + K \cdot N)^2)$ and a catastrophic spatial memory explosion, rendering native 2K multi-garment training fundamentally infeasible. Consequently, existing methods are forced to rely on suboptimal cascaded super-resolution pipelines, which inevitably introduce structural hallucinations and disrupt native spatial coherence.

The second barrier is the generative quality degradation caused by high-frequency attenuation. Even if memory constraints are artificially circumvented, generative models face a fundamental signal processing flaw at extreme spatial scales. The conventional $L_2$-norm (MSE) denoising objective exhibits an inherent spectral bias and a mean-seeking property. While relatively imperceptible at 1K resolution, this bias acts as an aggressive low-pass filter at the 2K spatial dimension. It systematically smooths out the microscopic high-frequency spatial components crucial for rendering complex textiles. As a result, materials heavily dependent on high-frequency details, such as coarse wool, heavy knits, and rugged denim, suffer from severe material degradation, deteriorating into overly smoothed, plastic-like synthetic surfaces that fail to capture tactile authenticity.

To systematically address these computational and theoretical bottlenecks, we propose WearWow, a unified framework specifically engineered for mask-free, native 2K multi-garment virtual try-on. The core of our framework is driven by a sequential design philosophy: first, making extreme-resolution training computationally feasible; and second, ensuring the generated textures achieve industrial-grade tactile realism. To achieve the first objective and fundamentally mitigate the computational explosion associated with multi-condition processing , we introduce Adaptive 2D Token Packing (ATP). By exploiting the inherent spatial sparsity of flat-lay garment images, ATP dynamically prunes uninformative background regions and packs heterogeneous visual tokens into a highly compact 2D latent matrix. Crucially, ATP preserves the original 2D spatial topology through rigorously mapped 2D positional encodings. By tightly packing valid regions and aggressively pruning background voids, this spatial reorganization drastically slashes the effective sequence length and subsequent memory footprint, enabling end-to-end joint training on standard hardware while mathematically guaranteeing spatial inductive biases.

To achieve the second objective and counteract the spectral attenuation inherent in standard diffusion objectives, we formulate the Multi-dimensional Try-on Reward (MTR) pipeline. MTR systematically shifts the optimization paradigm from purely latent-space regression to fine-grained preference alignment. Specifically, MTR introduces a Cloth Distribution Reward (CDR) to implicitly anchor the generative trajectory to high-fidelity physical data distributions, and a Semantic Guidance Reward (SGR) to explicitly drive tactile restoration via contrastive textual anchors. This dual-dimensional alignment effectively neutralizes the spectral bias of MSE, physically restoring the micro-level structural integrity of complex fabrics without succumbing to severe reward hacking.

The advancement of ultra-high-resolution digital fashion is further hindered by the severe scarcity of high-fidelity, multi-garment paired data at extreme spatial scales. To establish a rigorous benchmark and enable true mask-free generative capabilities, we curate and release WearWow-2K. This large-scale dataset features native 2K paired images with synthesized agnostic models, providing flawless high-frequency texture gradients and physically correct spatial interactions that naturally empower the model's target person mask-free generation. In summary, our core contributions are four-fold:

- We propose **WearWow**, the effective end-to-end multi-garment virtual try-on framework operating natively at 2K resolution, effectively advancing the task from macroscopic structural warping to micro-level tactile rendering.
- We introduce **Adaptive 2D Token Packing** , a spatial representation mechanism that mitigates the computational bottleneck by minimizing the effective sequence length while drastically slashing the sequence footprint and strictly preserving 2D spatial priors.
- We formulate the **Multi-dimensional Try-on Reward pipeline**. It effectively counteracts the spectral bias of diffusion models to physically rescue high-frequency fabric details and overcome plastic-like degradation.
- We release the **WearWow-2K dataset** featuring synthesized agnostic multi-garment models to unlock mask-free generation, and establish 2K resolution evaluation dimensions for ultra-HD digital fashion research.

## 2 Related Work

**Single-Garment Virtual Try-on.** Early virtual try-on paradigms predominantly relied on Generative Adversarial Networks (GANs) [10,11,16,20,23,38,42–44, 48], typically employing a two-stage pipeline consisting of an explicit warping module followed by a blending generator. While effective for macroscopic pose alignment, GAN-based architectures inherently struggle with synthesizing complex, high-frequency fabric textures and scaling to ultra-high resolutions.

Recently, Diffusion Models [29] have revolutionized the field by treating try-on as a conditional generation or inpainting task [5, 18, 28, 47]. Methods such as TryonDiffusion [52] introduced implicit warping mechanisms to better handle fabric deformations. Subsequent state-of-the-art frameworks—including OOTDiffusion [45], IDM-VTON [7], Any2AnyTryOn [15], IMAGDressing [35], Leffa [51], PromptDresser [17], cascade [21]and CatVTON [8]—have progressively refined appearance preservation through advanced cross-attention strategies and tailored condition encoders. Despite their remarkable success in preserving complex human poses and single-garment semantics, these architectures are intrinsically designed for single-item scenarios. When naively scaled to native 2K resolutions, they inevitably encounter the spectral bias of the conventional MSE objective, leading to over-smoothed, "plastic-like" textures that fail to satisfy industrial-grade tactile realism.

**Multi-Garment Joint Synthesis.** Compared to single-item VTON, multi-garment synthesis presents a significantly more daunting challenge due to the complex spatial occlusions and the exponential surge in condition tokens. While some early works explored compositional try-on via cyclical or sequential generation [9, 22, 43, 50], they inherently suffer from severe error accumulation and unnatural layer blending. More recently, OminiTry [12] pioneered the exploration of multi-garment inputs within a unified diffusion framework. However, due to the quadratic ($\mathcal{O}(N^2)$) memory explosion of self-attention mechanisms, there remains a severe lack of dedicated, open-source vertical baselines capable of executing native 2K multi-garment joint inference without catastrophic memory failures (OOM).

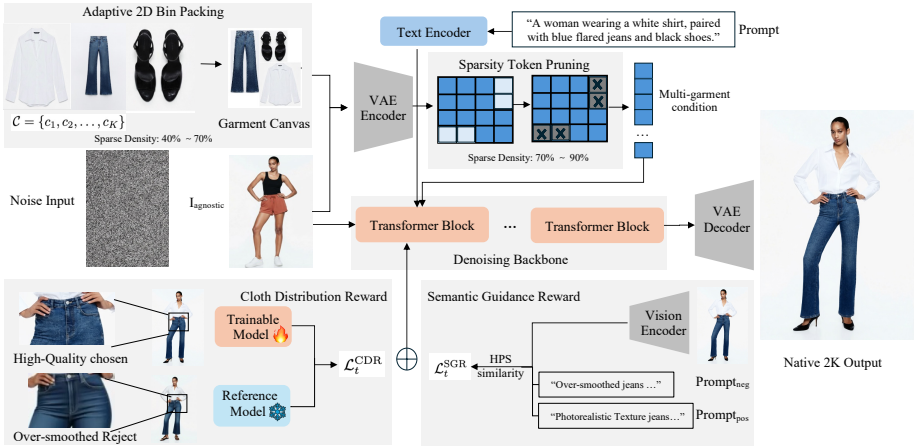


**Fig. 2:** The overall framework of WearWow. Our pipeline enables mask-free, native 2K multi-garment virtual try-on through two core mechanisms. **(Top) Adaptive 2D Token Packing:** Given a set of reference garments $\mathcal{C}$, ATP algorithmically packs them onto a unified spatial canvas. Following VAE encoding, a Sparsity Token Pruning mechanism aggressively discards background void tokens. This drastically minimizes the effective sequence length to mitigate the memory explosion, while strictly preserving 2D positional encodings. **(Bottom) Multi-dimensional Try-on Reward:** To overcome the spectral bias of standard SFT, MTR fine-tunes the generative trajectory. It synergizes a Cloth Distribution Reward (CDR) that penalizes over-smoothed artifacts to anchor physical distributions, and a Semantic Guidance Reward (SGR) that leverages HPS-driven contrastive textual anchors ($Prompt_{pos}, Prompt_{neg}$) to explicitly restore high-frequency tactile realism.

Consequently, the community has often looked toward state-of-the-art Large Vision-Language Models (LVLMs) and commercial generative APIs (e.g., QWen-image [40], Seedream4.0 [33], GPT-image 1.5 [1], KeLingv3 [37]) to handle extreme combinatorial complexity. While these foundation models have demonstrated astonishing semantic understanding and zero-shot compositional capabilities, they remain generalized text-to-image engines. They frequently suffer from cross-garment feature leakage, struggle to maintain spatial coherence under complex occlusion, and fail to render authentic material fidelity.

In stark contrast, our proposed WearWow is specifically engineered for this extreme frontier. By fundamentally resolving the multi-condition memory bottleneck via Adaptive 2D Token Packing (ATP) and explicitly rescuing high-frequency tactile textures through the Multi-dimensional Try-on Reward , WearWow establishes a robust, highly efficient vertical solution for ultra-high-resolution digital fashion.

# 3 Methodology

## 3.1 Overview and Problem Formulation

Native 2K multi-garment virtual try-on demands a unified generative paradigm capable of disentangling complex spatial interactions while strictly bounding computational overhead. Formally, WearWow accepts a triplet of conditional inputs: (1) a clothing-agnostic target person image $I_{\text{agnostic}}$; (2) a heterogeneous set of $K$ high-resolution reference garments $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$; and (3) a layer-aware textual prompt $\mathcal{P}$. To enable mask-free generation, we construct the WearWow-2K data engine using a "Real Ground-Truth, Synthesized Condition" paradigm. We utilize authentic 2K multi-garment photographs as the absolute target $\hat{I}_{\text{target}}$ to preserve pristine high-frequency textures, while inversely synthesizing $I_{\text{agnostic}}$ via a robust try-off pipeline. This formulation forces the model to implicitly learn physical depth ordering without relying on precise manual masking prior to inference (dataset details are deferred to the supplementary).

As illustrated in Figure 2, successfully scaling this formulation into the native 2K regime requires addressing two sequential hurdles: computational feasibility and generative quality. Our framework tackles these through two core components. First, Adaptive 2D Token Packing (ATP, Section 3.2) effectively mitigates the spatial memory bottleneck, providing a practical solution for how to train at extreme resolutions. Second, building upon this computationally feasible foundation, the Multi-dimensional Try-on Reward (MTR, Section 3.3) counteracts the inherent spectral bias of diffusion models, addressing how to train effectively to restore authentic tactile realism.

## 3.2 Adaptive 2D Token Packing (ATP)

In authentic virtual try-on scenarios, high-resolution reference garments inherently contain a massive proportion of uninformative background pixels. In a standard Transformer-based generative backbone, the input sequence consists of the fixed-length base tokens $N_{\text{base}}$ (derived from the noisy target and the clothing-agnostic model) and the conditional garment tokens. Naively injecting $K$ reference garments, each producing $N$ visual tokens, extends the total sequence length to $N_{\text{base}} + K \cdot N$. Given the inherent quadratic complexity of self-attention mechanisms, this naive concatenation triggers a prohibitive computational cost of $\mathcal{O}((N_{\text{base}} + K \cdot N)^2)$. As $K$ increases, this geometric explosion leads to catastrophic VRAM consumption and severe cross-garment spatial interference. To mitigate this without compromising high-frequency structural integrity, we propose the Adaptive 2D Token Packing (ATP) mechanism.

**Adaptive 2D Bin Packing.** Crucially, ATP does not alter the inherent quadratic nature of self-attention but strategically minimizes the conditional sequence length. Rather than concatenating conditions in 1D—which arbitrarily destroys spatial priors—ATP formulates reference injection as an algorithmic 2D bin packing optimization. The $K$ heterogeneous items are dynamically scaled based on geometric proportions and tightly packed onto a unified 2D spatial

canvas with a fixed token capacity $L$ (where $L \leq K \cdot N$). This spatial recombination instantly decouples the overall complexity from the quadratic scaling of $K$, strictly bounding the theoretical cost to $\mathcal{O}((N_{\text{base}} + L)^2)$. Empirically, this packing dramatically increases effective foreground token utilization from a baseline of approximately 40% (inherent in standard padded batching) to nearly 70%, while strictly isolating spatial coordinates to prevent feature leakage.

**Sparsity Token Pruning and Fallback.** To push inference efficiency to the physical limit, ATP further exploits the inherent spatial sparsity of the unified canvas. Despite optimal 2D bounding-box packing, the canvas inevitably retains uninformative blank padding due to the irregular topology of garments. Utilizing precise binary valid-region masks, ATP explicitly prunes and discards $m$ visual tokens corresponding to these background void spaces post-patchification. Consequently, the actual conditional sequence length processed by the transformer is strictly reduced to $(L - m)$, bounding the final computational complexity to $\mathcal{O}((N_{\text{base}} + L - m)^2)$. This precise mask-guided pruning elevates the conditional token utilization from 70% to over 90%, ensuring that the computational budget is entirely dedicated to informative garment features.

### 3.3 Multi-dimensional Try-on Reward (MTR)

While the proposed ATP mechanism successfully mitigates memory explosion—making native 2K multi-garment training computationally feasible—merely scaling the resolution does not automatically guarantee photorealistic textures. In practical, standard Supervised Fine-Tuning for virtual try-on encounters a fundamental quality bottleneck rooted in the inherent spectral bias of the MSE objective. The model tends to prioritize low-frequency structural alignment while critically neglecting the high-frequency details essential for fine-grained tactile realism. At native 2K resolution, this spectral bias leads to suboptimal material fidelity, causing complex textiles (e.g., wool, denim) to deteriorate into over-smoothed, "plastic-like" surfaces. To systematically restore these high-frequency components and ensure the rendering of authentic fabrics, we formulate the Multi-dimensional Try-on Reward (MTR), a post-training preference alignment pipeline that explicitly refines tactile realism via dual-dimensional objectives. Theoretically, MTR inherits the concepts of Direct Preference Optimization [27] (DPO), fundamentally bypassing the computationally expensive policy sampling required by traditional reinforcement learning algorithms like PPO [32].

**Cloth Distribution Reward (CDR)** The objective of CDR is to leverage the distributional margin between high-fidelity texture patterns and the over-smoothed artifacts inherent in the SFT stage. By shifting the generative trajectory toward the high-quality data manifold, CDR implicitly penalizes the low-pass filtering effect.

Specifically, we curate a comparative subset based on the physical material fidelity of generated results. At each noisy timestep $t$, a preference pair $(x_t^w, x_t^l)$ is constructed by diffusing a high-quality (chosen) sample $x_0^w$ and a low-quality

(rejected) sample $x_0^l$ to the same noise level. The alignment loss is formulated as:

$$\mathcal{L}_t^{\mathrm{CDR}} = -\mathbb{E}_{t,x_t^w,x_t^l}\left[\log\sigma\left(\mathbf{R}_\theta(x_t^w,\mathcal{C},\mathcal{P},v_{\mathrm{GT}}^w) - \mathbf{R}_\theta(x_t^l,\mathcal{C},\mathcal{P},v_{\mathrm{GT}}^l)\right)\right], \quad (1)$$

where $\sigma$ denotes the sigmoid function, $\mathcal{P}$ is the styling prompt, and $\mathcal{C}$ is the reference garment set. To avoid training an explicit reward model, the implicit reward $\mathbf{R}_\theta$ is defined directly via the velocity-field prediction errors between the trainable model $v_\theta$ and the frozen reference model $v_{\mathrm{ref}}$:

$$\mathbf{R}_\theta(x_t^*,\mathcal{C},\mathcal{P}) = \beta\left(\|v_{\mathrm{ref}}(x_t^*,\mathcal{C},\mathcal{P}) - v_{\mathrm{GT}}^*\|_2^2 - \|v_\theta(x_t^*,\mathcal{C},\mathcal{P}) - v_{\mathrm{GT}}^*\|_2^2\right), \quad (2)$$

where $\beta$ is a scaling hyper-parameter controlling the reward margin. $v_\theta$ and $v_{\mathrm{ref}}$ denote the velocity fields predicted at timestep $t$ by the trainable model and the frozen reference model, respectively. $v_{\mathrm{GT}}$ represents the corresponding ground-truth velocity target analytically computed for the current timestep. $*$ could be $w$ or $l$, which is the chosen or rejected samples at the same time.

**Semantic Guidance Reward (SGR)** While the Cloth Distribution Reward (CDR) successfully calibrates the velocity field to restore fine-grained fabric textures, relying exclusively on visual preference optimization suffers from sluggish convergence. We attribute this bottleneck to the rigid image-text prior established during the prolonged SFT phase. Since the underlying foundation model strongly couples visual distributions with textual prompts, optimizing the visual branch in isolation creates significant optimization inertia. To explicitly break this barrier and accelerate multimodal alignment, we introduce a complementary textual objective inspired by Semantic Relative Preference Optimization [36].

SGR widens the semantic similarity gap between the generated image and high-quality material prompts versus low-quality ones. First, given the clothing-agnostic model $I_{\mathrm{agnostic}}$, the visual conditions $\mathcal{C}$, and the baseline prompt $\mathcal{P}$, the diffusion backbone predicts the velocity field $v_\theta(x_t,\mathcal{C},\mathcal{P})$ at the noisy timestep $t$. Utilizing single-step denoising approximation, we project the current latent state directly to the pristine $t=0$ space, obtaining the native 2K spatial observation $\hat{I}_0$. To establish explicit contrastive semantic guidance, we construct paired semantic anchors: a positive prompt $\mathcal{P}_{\mathrm{pos}}$ (e.g., highly detailed, authentic material descriptions) and a negative prompt $\mathcal{P}_{\mathrm{neg}}$ (e.g., over-smoothed or plastic-like instructions). A frozen multi-modal reward model (e.g., CLIP [26]) then computes their respective cosine similarities with the generated image at each timestep $t$, denoted as $\mathbf{S}_t^{\mathrm{pos}}$ and $\mathbf{S}_t^{\mathrm{neg}}$.

$$\mathbf{S}_t^{\mathrm{pos}} = \cos(f_{\mathrm{img}}(\hat{I}_0), f_{\mathrm{txt}}(\mathcal{P}_{\mathrm{pos}})), \quad \mathbf{S}_t^{\mathrm{neg}} = \cos(f_{\mathrm{img}}(\hat{I}_0), f_{\mathrm{txt}}(\mathcal{P}_{\mathrm{neg}})), \quad (3)$$

where $f_{\mathrm{img}}$ and $f_{\mathrm{txt}}$ denote the vision and text encoders, respectively. The Semantic Guidance Loss is formulated using a standard margin-based ReLU activation:

$$\mathcal{L}_t^{\mathrm{SGR}} = \mathrm{ReLU}\left(\tau - [(1+\omega)\mathbf{S}_t^{\mathrm{pos}} - \mathbf{S}_t^{\mathrm{neg}}]\right), \quad (4)$$

where $\omega > 0$ acts as the relative CFG scaling coefficient, and $\tau$ is the predefined semantic margin.

A well-documented vulnerability of purely semantic-driven optimization is "reward hacking", where the model uncontrollably over-optimizes for textual alignment at the severe expense of structural integrity (e.g., inducing chromatic shifts or topological fragmentation). To effectively circumvent this degenerate behavior, we do not rely on SGR in isolation. Instead, our complete Multi-dimensional Try-on Reward (MTR) pipeline synergizes both dimensional objectives:

$$\mathcal{L}_t^{\mathrm{MTR}} = \alpha_t \mathcal{L}_t^{\mathrm{CDR}} + \lambda \mathcal{L}_t^{\mathrm{SGR}}, \tag{5}$$

where $\lambda$ is a balancing coefficient, and $\alpha_t$ is a timestep-dependent weighting function designed to dynamically dampen the CDR gradient at extreme noise levels.

This dual-dimensional balancing mechanism acts as a robust defense against reward hacking. The CDR implicitly anchors the generative trajectory within the physically correct high-frequency data distribution, preventing structural collapse, while the SGR provides explicitly accelerated semantic guidance for tactile restoration. Together, they enable the model to simultaneously fine-tune its outputs from both visual and textual perspectives.

## 4 Experiment

### 4.1 Datasets

To support native 2K multi-garment synthesis and empower mask-free training, we curate the WearWow-2K dataset, comprising approximately 100,000 training triplets and 2,000 test triplets at a pristine $2048 \times 1536$ resolution. To guarantee extreme visual quality and complex spatial diversity, we employ a hybrid curation strategy. Unlike existing single-item benchmarks, WearWow-2K features sophisticated layer-aware styling combinations ranging from 1 to 6 concurrent items per subject. Its highly diverse sartorial ontology extends beyond basic apparel to include footwear and fine-grained accessories (e.g., earrings), providing the physically correct spatial interactions necessary for multi-condition learning. Detailed data distributions and qualitative visual examples are deferred to the supplementary material.

### 4.2 Implementation Details

We build the WearWow generative backbone upon the pre-trained Qwen-Image-Edit foundation model, extracting hierarchical features from reference garments and textual prompts via a Qwen2.5VL [46] vision-language encoder coupled with our Adaptive 2D Token Packing (ATP) module. During the initial Supervised Fine-Tuning (SFT) phase at native 2K resolution ($2048 \times 1536$). And a global batch size of 8 for 80K steps. Following SFT convergence, we execute the Multi-dimensional Try-on Reward (MTR) alignment to explicitly restore tactile realism. The SGR text-image similarities are computed using a frozen HPSv2 [41].


**Fig. 3:** Qualitative comparisons on VITON-HD and DressCode in the single try-on. Compared with other methods, WearWow produces more texture-consistent images.

MTR fine-tuning runs for an additional 800 steps at a reduced learning rate of $1 \times 10^{-5}$, with hyperparameters empirically set to CFG scale $\omega \in [0.1, 0.6]$ (sampled as the timesteps changing), safety margin $\tau = 0.5$, CDR scale $\beta = 0.5$, and balancing weight $\lambda = 0.65$. All training is conducted in *bf16* on 8 $\times$ NVIDIA H20 GPUs.

### 4.3 Baselines and Experimental Setup

To comprehensively evaluate the robustness and scalability of WearWow, we design a dual-track evaluation protocol: a standard single-garment setting and an ultra-high-resolution multi-garment setting.

**Single-Garment Setup.** To ensure a rigorous and comprehensive evaluation, we conduct our single-garment experiments on a unified test dataset that includes the widely adopted VITON-HD [6] and DressCode [25] benchmarks. Since existing open-source baselines strictly limit their scope to this single-item paradigm, we evaluate this specific task at the standard 1024 $\times$ 768 resolution. We benchmark WearWow against a comprehensive suite of state-of-the-art diffusion-based models, including Any2AnyTryOn [15], IDM-VTON [7], IMAGDressing [35], Leffa [51], OOTDiffusion [45], PromptDresser [19], and CatVTON [8].

**Multi-Garment Setup.** To evaluate the extreme combinatorial complexity of multi-garment synthesis, we utilize our proposed WearWow-2K dataset at its native 2048 $\times$ 1536 resolution. Because traditional open-source try-on models mathematically collapse (OOM) under multi-condition 2K inputs, we benchmark WearWow against the most powerful recent generative foundation models and commercial APIs. The baselines include local multi-modal models (OminiTry [12], QWen-image [40]) and top-tier closed-source APIs (Nano Banana Pro [30], KeLingv3 [37], Seedream4.0 [33] and GPT-image 1.5 [3]).

**Evaluation Metrics.** We employ a robust set of quantitative metrics to assess generative quality and structural alignment. To explicitly quantify our core contribution in restoring high-frequency tactile realism and perceptual quality, we prioritize the Learned Perceptual Image Patch Similarity (LPIPS [49]) and


**Fig. 4:** Qualitative comparisons on WearWow-2k,showcasing our superior photorealism on complex multi-item combinations and challenging woolen textures.

Fréchet Inception Distance (FID [34]). We also report Kernel Inception Distance (KID [4]) for unbiased distribution measurement and Structural Similarity Index (SSIM [39]) for spatial alignment. For ablation study, CLIP score is also computed by measuring the similarity between textual description and the result.

**Human Evaluation (HE).** To complement quantitative metrics that occasionally miss micro-level perceptual nuances at 2K resolution, we conduct a rigorous Human Evaluation with 100 diverse participants. Evaluators blindly score randomized outputs from WearWow and baselines across three dimensions: (1) Material Fidelity (tactile realism without plastic artifacts), (2) Spatial Coherence (correct multi-garment occlusion), and (3) Overall Photorealism.The averaged Human Evaluation (HE) scores are reported to provide a definitive subjective benchmark.

### 4.4 Quantitative Analysis

**Single-Garment Evaluation.** As reported in Table 1, WearWow demonstrates highly competitive performance in perceptual realism. Specifically, WearWow achieves strong results across distribution and perceptual metrics, showing notable improvements over recent baselines like IDM-VTON [7] and Leffa [51] in terms of FID and LPIPS. It is worth noting that while CatVTON [8] achieves a marginally higher SSIM, this metric heavily relies on pixel-wise mean error, which can inherently favor models that produce overly smoothed, low-frequency outputs. In contrast, WearWow's improvements in FID, KID, and LPIPS indicate that our MTR pipeline effectively restores high-frequency tactile realism. Furthermore, our Human Evaluation (HE) supports this alignment with human visual preferences; evaluators generally favored WearWow in terms of material fidelity and overall photorealism.

**Table 1:** Quantitative comparisons on the VITON-HD and DressCode dataset.

| Method | SSIM ↑ | FID ↓ | LPIPS ↓ | KID ↓ | HE ↑ |
|---|---|---|---|---|---|
| Any2AnyTryOn [15] | 0.8460 | 6.7285 | 0.1292 | 1.4906 | 0.1138 |
| IDM-VTON [7] | 0.8779 | 5.9893 | 0.0720 | 1.7433 | 0.0863 |
| IMAGDressing [35] | 0.8298 | 9.0388 | 0.1120 | 2.8076 | 0.0379 |
| Leffa [51] | 0.8582 | 5.9421 | 0.0825 | 0.9899 | 0.1389 |
| OOTDiffusion [45] | 0.8565 | 5.9640 | 0.0782 | 0.9249 | 0.0505 |
| PromptDresser [19] | 0.8513 | 6.5618 | 0.0973 | 1.3420 | 0.0905 |
| CatVTON [8] | **0.8950** | 4.5698 | 0.0746 | 1.0298 | 0.0737 |
| **WearWow** | 0.8815 | **4.2475** | **0.0688** | **0.6503** | **0.4084** |

**Table 2:** Quantitative comparisons on WearWow-2K .

| Method | FID ↓ | KID ↓ | LPIPS ↓ | SSIM ↑ | HE ↑ |
|---|---|---|---|---|---|
| OminiTry [12] | 12.4164 | 2.0228 | 0.1444 | 0.8002 | 0.0257 |
| QWen-image [40] | 14.3538 | 2.3196 | 0.2175 | 0.7516 | 0.0286 |
| Seedream4.0 [33] | 14.3389 | 4.7620 | 0.2175 | 0.7699 | 0.1514 |
| Nano Banana Pro [31] | 8.1823 | 0.8285 | 0.1114 | 0.8215 | 0.2229 |
| KeLingv3 [37] | 8.9624 | 1.0006 | 0.1108 | 0.8255 | 0.1314 |
| GPT-image 1.5 [3] | 10.7218 | 1.3345 | 0.1724 | 0.7705 | 0.1857 |
| **WearWow (ours)** | **7.5400** | **0.2643** | **0.0773** | **0.8461** | **0.2543** |

**Multi-Garment Evaluation.** The quantitative advantages of WearWow are further evident in the native 2K multi-garment regime (Table 2). Commercial APIs like Nano Banana Pro [31] and KeLingv3 [37], despite their large parameter counts, often encounter difficulties with strict spatial control and structural consistency, reflecting in their FID and KID scores. WearWow effectively addresses these challenges, achieving consistent improvements across key metrics (FID, KID, SSIM, and LPIPS). This stable performance indicates that our Adaptive 2D Token Packing successfully mitigates the spatial interference of multi-garment conditions, while MTR maintains the generative quality at extreme spatial scales. Consistent with the objective metrics, our Human Evaluation (HE) shows a clear user preference for WearWow, particularly highlighting its spatial coherence and mask-free compositional stability compared to closed-source foundation models.

### 4.5 Qualitative Analysis

**Single-Garment Tactile Realism and Consistency.** As illustrated in Figure 3, existing baselines struggle severely when rendering garments with complex physical properties. Models like IDM-VTON [7] and CatVTON [8] tend to act as low-pass filters, degrading intricate materials into synthetic, plastic-like flat surfaces. Conversely, empowered by the dual-dimensional MTR alignment,


**Fig. 5:** We visualize the impact of our proposed modules. our full MTR pipeline (**Ours**) robustly restores micro-level tactile realism without sacrificing structural integrity.

WearWow exhibits superior capability in restoring high-frequency details. Notably, when processing challenging fabrics like rugged denim or complex texture garment , our framework demonstrates exceptional stability and consistency.

**Multi-Garment Stability and Material Fidelity.** Figure 4 demonstrates the mask-free joint synthesis capabilities on WearWow-2K. When prompted to simultaneously wear multiple items (e.g., inner-shirts, thick outerwear, denim pants, and fine accessories), recent open-source models and baseline APIs often exhibit severe compositional instability. They either hallucinate bizarre merged textures (feature leakage), entirely ignore specific clothing items, or suffer from catastrophic texture degradation at native 2K resolution. In stark contrast, WearWow operates with impeccable generative stability. Driven by the layer-aware annotations and the ATP module's strict 2D positional preservation, WearWow accurately renders correct depth orderings (e.g., a jacket explicitly open over a tucked-in shirt) with flawless occlusion boundaries, while simultaneously maintaining extreme material fidelity across all concurrent garments.

### 4.6 Ablation study

To rigorously validate our core contributions, we evaluate the spatial representation strategy (ATP) and preference alignment pipeline (MTR) against various baselines (Table 3 and Figure 5). To validate ATP, we compare WearWow against two scaling paradigms. The **w/ Cascaded SR** variant processes inputs at a lower resolution before employing an external Super-Resolution model, which inevitably introduces structural hallucinations. Alternatively, maintaining native 2K training but naively downscaling reference items to satisfy memory

constraints (**w/o ATP**) severely destroys high-frequency visual priors. Consequently, both approaches suffer from noticeable texture degradation and catastrophic over-smoothing, exhibiting inferior scores compared to our ATP-equipped model. This demonstrates that algorithmically packing and pruning background tokens is essential for preserving exact foreground resolution and spatial inductive biases.

Next, we dissect the MTR pipeline to validate the synergy between its physical and semantic components. Applying only the Cloth Distribution Reward (**w/ CDR**) anchors the generative trajectory to the data distribution, yielding slight visual improvements. However, its training convergence is notably slow, and it ultimately struggles to fully overcome the SFT optimization inertia, leaving the tactile realism of complex high-frequency textures insufficient. Conversely, relying solely on the Semantic Guidance Reward (**w/ SGR**) provides a strong textual push but exhibits critical vulnerabilities. Pure semantic guidance initially forces an over-alignment with textual prompts, introducing unnatural chromatic shifts and an excessively rough, frizzy appearance to the fabrics. More critically, as training progresses, it inevitably induces severe "reward hacking". By synergizing both objectives, our full MTR pipeline successfully resolves all aforementioned issues; CDR provides the essential physical anchor to ensure stability and prevent hacking, while SGR effectively accelerates convergence and restores flawless micro-level tactile realism.

**Table 3:** Comparison of our full WearWow framework against its ablative variants.

| Method | SSIM ↑ | LPIPS ↓ | CLIP Score ↑ |
|---|---|---|---|
| w/ cascade SR | 0.8502 | 0.1237 | 0.2651 |
| w/o ATP | 0.8468 | 0.0923 | 0.2654 |
| w/ SGR | 0.8599 | 0.0874 | 0.2678 |
| w/ CDR | 0.8596 | 0.0961 | 0.2664 |
| **Ours** | **0.8677** | **0.0870** | **0.2690** |

## 5 Conclusion

In this paper, we introduced **WearWow**, an effective generative framework that scales virtual try-on to the native 2K multi-garment regime. To dismantle the prohibitive memory barriers associated with extreme-resolution condition injection, we proposed Adaptive 2D Token Packing . By exploiting the inherent spatial sparsity of garment images, ATP algorithmically packs and prunes heterogeneous visual tokens into a unified 2D layout, effectively mitigate the computational explosion while rigorously preserving essential 2D spatial priors. Furthermore, to address the inherent spectral bias and over-smoothing degradation of standard diffusion models, we formulated the Multi-dimensional Try-on Reward. By synergizing a Cloth Distribution Reward to anchor the physical distribution with a Semantic Guidance Reward to provide explicit textual pushes, MTR aligns the generative trajectory to penalize plastic-like artifacts and robustly restore micro-level tactile realism. Supported by the WearWow-2K dataset, our frame-

work achieves both precise spatial composition and authentic fabric details and sets a state-of-the-art for 2K multi-garment virtual try-on.
**Social Impact**: While WearWow significantly advances digital fashion, we acknowledge the inherent risks of ultra-high-fidelity human generation. We emphasize the critical need for robust deepfake detection algorithms and strictly regulated deployment to prevent the malicious misuse of generated identities.
**Limitations and Future Work** While WearWow successfully scales multi-garment virtual try-on to native 2K resolution and establishes a new standard for tactile realism, it possesses certain limitations that present exciting avenues for future research. Handling extreme complex back-view occlusions remains a common challenge. Future work will explore integrating explicit 3D human body priors to guarantee rigorous, view-consistent multi-garment synthesis across arbitrary camera angles. Furthermore, building upon the robust spatial representation and tactile foundation established by WearWow, a natural and highly promising progression is extending this framework into the temporal domain. We aim to explore temporal consistency mechanisms for high-resolution video virtual try-on, coupled with dynamic garment relighting under complex environmental illuminations.

## Acknowledgements

This work is supported by National Key Research and Development Program of China2024YFE0203100)Scientific Research Innovation Capability Support Project for Young Faculty (No.ZYGXQNJSKYCXNLZCXM-I28), Shenzhen Science and Technology ProgramGrant No. QNXMA20250701093544048 and General Embodied AI Center of Sun Yat-sen University.

## References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 5
2. Bai, S., Zhou, H., Li, Z., Zhou, C., Yang, H.: Single stage virtual try-on via deformable attention flows. In: European Conference on Computer Vision (2022) 2
3. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023) 10, 12
4. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans (2021), https://arxiv.org/abs/1801.01401 11
5. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. arXiv preprint arXiv:2307.09481 (2023) 4
6. Choi, S., Park, S., Lee, M., Choo, J.: Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR) (2021) 10
7. Choi, Y., Kwak, S., Lee, K., Choi, H., Shin, J.: Improving diffusion models for authentic virtual try-on in the wild. arXiv preprint arXiv:2403.05139 (2024) 4, 10, 11, 12

8. Chong, Z., Dong, X., Li, H., Zhang, S., Zhang, W., Zhang, X., Zhao, H., Liang, X.: Catvton: Concatenation is all you need for virtual try-on with diffusion models (2024), https://arxiv.org/abs/2407.15886 4, 10, 11, 12
9. Cui, A., McKee, D., Lazebnik, S.: Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In: Proceedings of the IEEE/CVF international conference on computer vision (2021) 4
10. Dong, H., Liang, X., Shen, X., Wang, B., Lai, H., Zhu, J., Hu, Z., Yin, J.: Towards multi-pose guided virtual try-on network. In: Proceedings of the IEEE/CVF international conference on computer vision (2019) 4
11. Dong, X., Zhao, F., Xie, Z., Zhang, X., Du, D.K., Zheng, M., Long, X., Liang, X., Yang, J.: Dressing in the wild by watching dance videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022) 4
12. Feng, Y., Zhang, L., Cao, H., Chen, Y., Feng, X., Cao, J., Wu, Y., Wang, B.: Omnitry: Virtual try-on anything without masks. arXiv preprint arXiv:2508.13632 (2025) 4, 10, 12
13. Ge, Y., Song, Y., Zhang, R., Ge, C., Liu, W., Luo, P.: Parser-free virtual try-on via distilling appearance flows. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2021) 2
14. Gou, J., Sun, S., Zhang, J., Si, J., Qian, C., Zhang, L.: Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In: Proceedings of the 31st ACM International Conference on Multimedia (2023) 2
15. Guo, H., Zeng, B., Song, Y., Zhang, W., Liu, J., Zhang, C.: Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19085–19096 (2025) 4, 10, 12
16. He, S., Song, Y.Z., Xiang, T.: Style-based global appearance flow for virtual try-on. In: CVPR (2022) 4
17. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022) 4
18. Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J.: Composer: Creative and controllable image synthesis with composable conditions. arXiv preprint arXiv:2302.09778 (2023) 4
19. Kim, J., Jin, H., Park, S., Choo, J.: Promptdresser: Improving the quality and controllability of virtual try-on via generative textual prompt and prompt-aware mask (2024), https://arxiv.org/abs/2412.16978 10, 12
20. Lee, S., Gu, G., Park, S., Choi, S., Choo, J.: High-resolution virtual try-on with misalignment and occlusion-handled conditions. In: European Conference on Computer Vision (2022) 4
21. Li, G., Wang, Y., Luan, J., Zhao, L., Xing, W., Lin, H., Ou, B.: Cascaded diffusion models for virtual try-on: Improving control and resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4689–4697 (2025) 4
22. Lin, E., Zhang, X., Zhao, F., Luo, Y., Dong, X., Zeng, L., Liang, X.: Dreamfit: Garment-centric human generation via a lightweight anything-dressing encoder. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). pp. 5218–5226 (2025), aAAI 2025 4
23. Minar, M.R., Tuan, T.T., Ahn, H., Rosin, P., Lai, Y.K.: Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2020) 4

24. Morelli, D., Baldrati, A., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R.: LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On. In: Proceedings of the ACM International Conference on Multimedia (2023) 2
25. Morelli, D., Fincato, M., Cornia, M., Landi, F., Cesari, F., Cucchiara, R.: Dress code: High-resolution multi-category virtual try-on (2022) 10
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021) 8
27. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems **36**, 53728–53741 (2023) 7
28. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022) 4
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2022) 4
30. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding (2022) 10
31. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022) 12
32. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017) 7
33. Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., et al.: Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint arXiv:2509.20427 (2025) 5, 10, 12
34. Seitzer, M.: pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid (August 2020), version 0.3.0 11
35. Shen, F., Jiang, X., He, X., Ye, H., Wang, C., Du, Xiaoyu, L.Z., Tang, J.: Imagdressing-v1: Customizable virtual dressing (2024) 4, 10, 12
36. Shen, X., Li, Z., Yang, Z., Zhang, S., Zhang, Y., Li, D., Wang, C., Lu, Q., Tang, Y.: Directly aligning the full diffusion trajectory with fine-grained human preference. arXiv preprint arXiv:2509.06942 (2025) 8
37. Team, K., Chen, J., Ci, Y., Du, X., Feng, Z., Gai, K., Guo, S., Han, F., He, J., He, K., et al.: Kling-omni technical report. arXiv preprint arXiv:2512.16776 (2025) 5, 10, 12
38. Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L.: Toward characteristic-preserving image-based virtual try-on network. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 589–604 (2018) 4
39. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004) 11
40. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025) 5, 10, 12

41. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023) 9
42. Xie, Z., Huang, Z., Dong, X., Zhao, F., Dong, H., Zhang, X., Zhu, F., Liang, X.: Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 2, 4
43. Xie, Z., Huang, Z., Zhao, F., Dong, H., Kampffmeyer, M., Dong, X., Zhu, F., Liang, X.: Pasta-gan++: A versatile framework for high-resolution unpaired virtual try-on. arXiv preprint arXiv:2207.13475 (2022) 4
44. Xie, Z., Zhang, X., Zhao, F., Dong, H., Kampffmeyer, M.C., Yan, H., Liang, X.: Was-vton: Warping architecture search for virtual try-on network. In: Proceedings of the 29th ACM International Conference on Multimedia (2021) 4
45. Xu, Y., Gu, T., Chen, W., Chen, C.: Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. arXiv preprint arXiv:2403.01779 (2024) 4, 10, 12
46. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 9
47. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 4
48. Yang, H., Zhang, R., Guo, X., Liu, W., Zuo, W., Luo, P.: Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2020) 4
49. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 10
50. Zhang, X., Lin, E., Li, X., Luo, Y., Kampffmeyer, M., Dong, X., Liang, X.: Mmtryon: Multi-modal multi-reference control for high-quality fashion generation (2024), https://arxiv.org/abs/2405.00448 4
51. Zhou, Z., Liu, S., Han, X., Liu, H., Ng, K.W., Xie, T., Cong, Y., Li, H., Xu, M., Pérez-Rúa, J.M., Patel, A., Xiang, T., Shi, M., He, S.: Learning flow fields in attention for controllable person image generation. arXiv preprint arXiv:2412.08486 (2024) 4, 10, 11, 12
52. Zhu, L., Yang, D., Zhu, T., Reda, F., Chan, W., Saharia, C., Norouzi, M., Kemelmacher-Shlizerman, I.: Tryondiffusion: A tale of two unets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 4