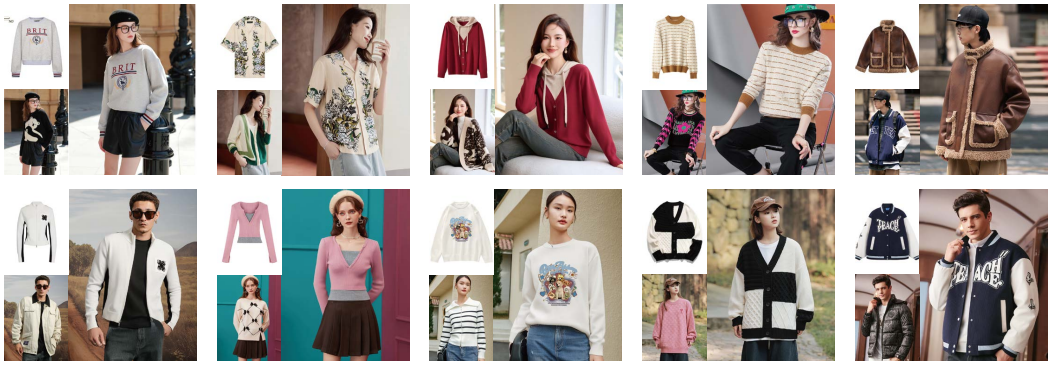
AnyFit: Controllable Virtual Try-on for Any Combination of Attire Across Any Scenario

Yuhan Li[1*], **Hao Zhou**[2], **Wenxiang Shang**[2], **Ran Lin**[2], **Xuanhong Chen**[1], **BingBing Ni**[1†]
[1]Shanghai Jiao Tong University, Shanghai 200240, China
[2]Alibaba
{melodious, nibingbing}@sjtu.edu.cn



(a) High-fidelity virtuil try-on across any scenario

(b) High-fidelity virtuil try-on for any combination of attire

Figure 1: AnyFit shows superior try-ons for any combination of attire across any scenario.

Abstract

While image-based virtual try-on has made significant strides, emerging approaches still fall short of delivering high-fidelity and robust fitting images across various scenarios, as their models suffer from issues of ill-fitted garment styles and quality degrading during the training process, not to mention the lack of support for various combinations of attire. Therefore, we first propose a lightweight, scalable, operator known as Hydra Block for attire combinations. This is achieved through a parallel attention mechanism that facilitates the feature injection of multiple garments from conditionally encoded branches into the main network. Secondly, to significantly enhance the model's robustness and expressiveness in real-world scenarios, we evolve its potential across diverse settings by synthesizing the residuals of multiple models, as well as implementing a mask region boost strategy to overcome the instability caused by information leakage in existing models. Equipped with the above design, AnyFit surpasses all baselines on high-resolution benchmarks and real-world data by a large gap, excelling in producing well-fitting garments replete with photorealistic and rich details. Furthermore, AnyFit's impressive performance on high-fidelity virtual try-ons in any scenario from any image, paves a new path for future research within the fashion community.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-2646

1 Introduction

The dramatic success of e-commerce is steadily demanding a more convenient and personalized customer shopping experience. Among them, image-based virtual try-on (VTON) (23; 47; 22) has emerged as a promising topic in the research community and witnesses rapid advancements (15; 24; 18), whose task is to fit the target garment to the human body with various gestures. However, current methodologies do not meet the high-fidelity and robustness required in real-world applications, often resulting in artifacts or the mismatch of clothing details. Furthermore, the support for a variety of try-on combinations of attire (10) remains an area of ongoing research.

Most prior methodologies (6; 14) utilize a separate warping module to align garments on the human body, subsequently employing a Generative Adversarial Network (GAN) (19) for their integration. This explicit warping process typically yields overly smooth garment transformations and struggles to cope with complex poses and occlusions. (50). While some diffusion-based methods (58; 50) leverage pre-trained diffusion models (43), with structures akin to ReferenceNet (26) to preserve fine-grained garment information; However, these methods encounter difficulties in producing vivid fabric textures and photorealistic lighting and shadows. They also show artifacts in cross-category try-ons, diminishing their inherent text-image capacity when applied to specialized tasks as depicted in Fig. 4. In summary, existing methods still fall short in producing images of high fidelity that exhibit clothing styles rendered with exceptional detail and true-to-life accuracy across scenes. Moreover, these methods are designed solely for trying on individual items of clothing and do not support multi-conditions, thereby failing to facilitate the free combination of tops and bottoms.

As discussed above, we believe that an ideal VTON workflow should exhibit the following properties:

- **Scalability.** The ultimate goal of the VTON model is to enable any free and customizable virtual outfit combination of multiple garments (57). It should support multi-condition injection, allowing for easy expansion to more applications, such as mixing and matching tops and bottoms, layering inner and outer garments, *etc*.
- **Robustness.** Given the diverse scenarios encountered in e-commerce settings (13), the VTON model should generate authentic fabric textures and natural lighting, reproducing the details of the target clothing (*e.g.*, logos, patterns, texts and strips) stably and accurately.

We present the following critical contributions to establish AnyFit as a novel VTON paradigm, which adeptly addresses the challenge of any combination of attire across any conceivable scenario, in Fig. 1. AnyFit mainly consists of two isomorphic U-Nets, namely HydraNet and MainNet. The former is tasked with extracting fine-grained clothing features, while the latter is responsible for generating try-ons. (1) **Scalability**: A hallmark of AnyFit is its innovative introduction of the **Hydra Encoding Block** that only parallelizes attention matrices within a sharing HydraNet, enabling effortless expansion to any quantity of conditions with only 8% increase in parameters for each additional branch. The proposal of parallelizing these blocks is built on the insight that only the self-attention layers are crucial for implicit warping (50), while the remaining components primarily serve as generic feature extractors. We further invent **Hydra Fusion Block** to seamlessly integrate the features of Hydra Encoding into MainNet, with positional embeddings to distinguish encodings from different sources. It is important to note that ReferenceNet (26; 9) or GarmentNet (50) could be seen as specific instances of HydraNet when limited to a single condition. (2) **Robustness**: Observations indicate a noticeable reduction in the robustness and quality of images generated by existing Virtual Try-On (VTON) works, in comparison to the original stable diffusion performances. Inspired by discussions in the community (2), we present the **Prior Model Evolution** strategy. This innovative approach involves merging parameter variations within a model family (*e.g.*, a collection of fine-tuned versions of SDXL (41)), enabling the independent evolution of multiple capabilities of the base model. This strategy emerges as an intuitively logical and highly effective method for amplifying the model's innate potential prior to training. This is particularly relevant when contending with the significant escalation in training costs associated with dual U-Nets—an aspect that is overlooked in previous research. Furthermore, we introduce the **Adaptive Mask Boost** to further enhance the fit of the attire as a bonus. It requires length augmentation of *parsing-free* mask regions during the training phase, allowing the model to autonomously understand the overall shape of the clothing, which emancipates the model from previous reliance on hints of masks derived from garments. During inference, we adapt the shape of the mask area based on the aspect ratio of the target garment, thereby markedly encouraging the generation of well-fitted try-ons, particularly for long garments (*e.g.*, windbreaker).

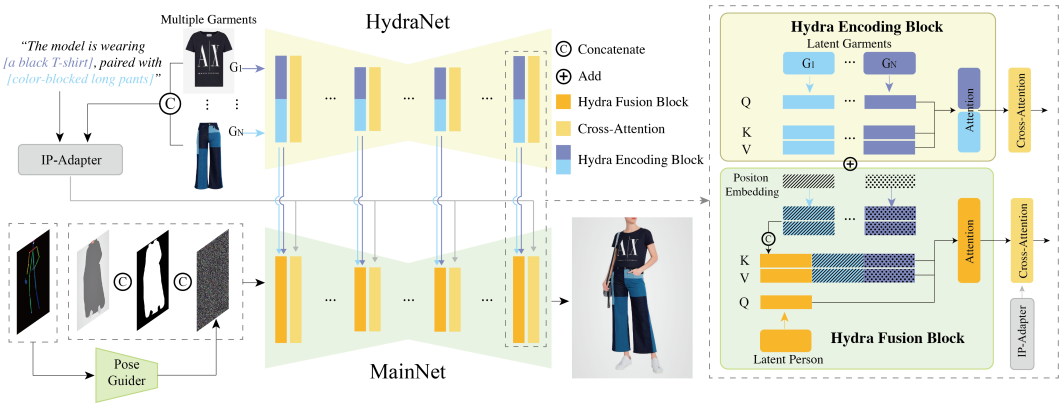


Figure 2: Overall framework of our method.

To the best of our knowledge, AnyFit stands as a pioneering VTON method to fulfill scalability and robustness requirements. Our innovative HydraNet and Prior Model Evolution strategies have the potential to transform not just the domain of VTON, but to catalyze advancements across a broader spectrum of conditional generation applications. Finally, we have carried out comprehensive experiments on try-on benchmarks (12; 16) and engaged in challenging validation using in-the-wild sets. These experiments demonstrate that our model shows exceptional performance that eclipses current methods by a substantial margin, in terms of garment fidelity and robustness when addressing street-captured scenarios. In addition, our method has realized a formidable capability for multi-garment try-ons, culminating in results that exhibit strikingly harmonized upper and lower styles.

2 Related works

GAN-based virtual try-on. The virtual try-on task is concerned with synthesizing images of a person donning the designated garment with appropriate fit (23), while retaining salient characteristics of the original garment and person, given a pair of images depicting a person and a target garment. To execute this task, numerous works (17; 30; 34; 6; 14; 22) have utilized Generative Adversarial Networks (GANs) (19) with two-stage strategy (35; 18; 52): (1) warping the clothing to the desired shape (7; 36) and (2) fusing the deformed clothing via try-on generator based on GAN. HR-VITON (35) conducts both warping and segmentation concurrently to address issues related to body occlusion and misfit of garments. GP-VTON (49) introduces localized warping along with global parsing to independently simulate the deformation of different regions of clothing, aiming to achieve a more form-fitting result. However, these existing approaches that rely on an explicit warping module are incapable of supplementing the sides of the clothing and the natural lighting and shadows (50).

Diffusion-based virtual try-on. As significant progress in Text-to-Image diffusion models (25; 39; 27; 11) is witnessed in recent years, some works (10) have been motivated to incorporate pre-trained diffusion models (43; 41) as priors into virtual try-on task. LADI-VTON (38) and DCI-VTON (20) explicitly deform the clothing to achieve pixel-level alignment with the human body, followed by a diffusion model to blend the clothing with the human body as refinement. StableVITON (31) introduces an end-to-end approach that injects intermediate feature maps from a spatial encoder into the U-Net decoder via a zero cross-attention block, akin to the ControlNet (55) structure. Most recently, OOTDiffusion (50) and IDM (13) achieve garment feature extraction with a parallel U-Net and feed them through self-attention for enhanced integration. Unfortunately, these methods intrinsically lack support for try-ons that involve multiple garments. Moreover, they exhibit artifacts and unstable garment fits for arbitrary images, which leads to a degradation of performance on out-of-distribution images in complex backgrounds and poses.

3 Method

3.1 Model overview

An overview of the AnyFit is presented in Fig. 2. The backbone of AnyFit employs the SDXL (43), with the preliminary detailed in Appendix C.2. Given a human image $x_h \in \mathbb{R}^{H\times W\times 3}$ and a target garment image $x_g \in \mathbb{R}^{H\times W\times 3}$, AnyFit is aimed to generate an authentic try-on image x_{tr}. We employ OpenPose (8; 53) to obtain clothing-agnostic mask x_m and masked person image x_{ag}

https://doi.org/10.52202/079017-2646

adjusting for the size of different garments, as detailed in Sec. 3.3. We treat VTON as a specific case of image inpainting (51), endeavoring to fill the masked person x_{ag} with the cloth x_g. The main inpainting U-Net (MainNet) inputs 3 concatenated components with 9 channels: the noisy image z_t, the latent agnostic image $E(x_{ag})$ and the resized agnostic mask x_m, where $E(\cdot)$ represents VAE (32) encoding. A Pose Guider (26) with 4 convolution layers (4×4 kernels, 2×2 strides, $16, 32, 64, 128$ channels) is incorporated to align the pose image $E(x_p)$ with noise z_t.

Scalability: To preserve the fine details of the clothing, as well as to support both single and multiple garment VTONs, we employ a HydraNet that mirrors the MainNet in encoding clothing information. It shares the same weight initialization as the MainNet and innovatively parallelizes attention metrics based on the number of conditions to create Hydra Encoding Blocks for different conditional encodings. **Robustness:** During training, issues such as mask information leakage and quality degradation were observed. To address these issues, we adopt Adaptive Mask Boost and Prior Model Evolution, respectively, which significantly bolster the model's robustness across different scenarios cost-effectively and straightforwardly.

3.2 HydraNet for multi-condition VTON

HydraNet. Inspired by successful practice in human editing (26; 9), we introduce a garment encoding network isomorphic to the main generative network (MainNet), that precisely preserves the details of clothing. When dealing with multi-garment VTON, a direct method might involve replicating multiple garment encoding nets to manage different conditions. This approach, however, would lead to a significant increase in the number of parameters, rendering it computationally prohibitive. Experimentally we discover that for conditions with similar content (such as different types of clothing), the self-attention module plays a vital role in the latent warping of the garments, aligning them with the locations requiring inpainting. Conversely, other network architectures, which typically tasked with general feature extraction, can be shared across different condition encoding branches without compromising the model's performance. In view of this, we innovatively propose HydraNet for multi-condition encoding. It operates based on a shared Unet structure, while parallelizing the attention modules according to the number of input conditions, thereby constructing Hydra Encoding Blocks. Specifically, we parallelize the self-attention matrices with identical initial weights, and feed the multi-condition *key* and *value* features $\left\{z^i_{hk}, z^i_{hv}\right\}$ into MainNet, which encode the fine-grained details of the clothing. It's notable that ReferenceNet (26; 9) or GarmentNet (50) could be seen as specific instances of HydraNet limited to a single condition. HydraNet requires only one forward pass (timestep $t = 0$) to encode clothing before the multiple denoising steps in MainNet, with the additional temporal and parameter overheads being minimal for each added condition, in Tab. 2.

Hydra Fusion. We propose a highly efficient and easily scalable Hydra Fusion Block to replace the self-attention layers in MainNet, accomplishing the feature injection from HydraNet to MainNet for any length via concatenation. Specifically, given the *key* and *value* features $\left\{z^i_{hk}, z^i_{hv}\right\} \in \mathbb{R}^{b\times l\times c}$ from the HydraNet, we introduce learnable position embeddings to distinguish features from different source conditions. The superscript i denotes different input conditions. Subsequently, we concatenate the *key* and *value* along the l dimension to obtain the final $\left\{z^{all}_{hk}, z^{all}_{hv}\right\} \in \mathbb{R}^{b\times Nl\times c}$ as:

$$z^{all}_{hk} = (z^1_{hk} + PE^1(z^1_{hk})) \oplus (z^2_{hk} + PE^2(z^2_{hk})) \oplus \cdots \oplus (z^N_{hk} + PE^N(z^N_{hk})), \tag{1}$$

where N denotes the total number of input conditions, PE represents positional encoding, and $\oplus$ signifies concatenation. z^{all}_{hv} follows a similar formulation. Facing the *key* and *value* features $\{z_{mk}, z_{mv}\} \in \mathbb{R}^{b\times l\times c}$ from the MainNet and concatenated HydraNet features $\left\{z^{all}_{hk}, z^{all}_{hv}\right\}$, we once again concatenate the corresponding features along the l dimension into $\{z_{ck}, z_{cv}\} \in \mathbb{R}^{b\times (N+1)l\times c}$, which are then used in subsequent attention calculations with z_{mq}. It is noteworthy that, with the lightweight design leveraging parallelization and concatenation, HydraNet can be effortlessly extended to perform injections with any number of conditions, thereby possessing a more expansive application potential within the domain of generative models.

3.3 Model evolution and mask boost for robust VTON

Prior Model Evolution. A diminution in image generation performance with the SDXL-inpainting model compared to the SDXL base model is noted (1). We attribute this degradation to the disruption of previously well-aligned correspondences between text and images during the inpainting pre-training phase. Drawing inspiration from the open-source community (2), we develop a Prior Model

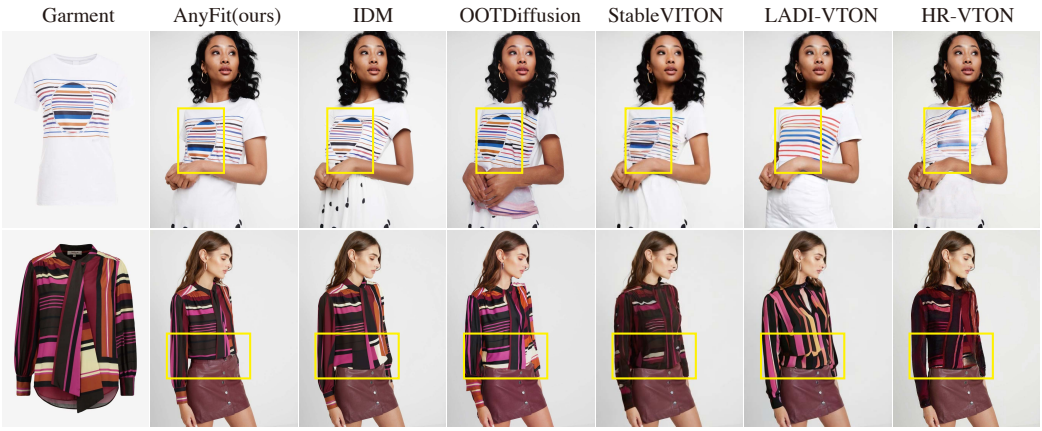


Figure 3: Visual comparisons on VITON-HD. AnyFit displays superior details and outfit styling.

Evolution strategy, which enhances the model's strength and adaptability in generating outfit images at a very low cost, even without training. Specifically, we meticulously amalgamate the weights from three distinct, powerful models to evolve the initial weights for our model. These models comprise: SDXL-base-1.0 (41), SDXL-inpainting-0.1 (4) with inpainting capabilities, and DreamshaperXL alpha2 (3), which demonstrates superior performance in generating clothing and human figures. Their weights are denoted by $\mathbf{W}_{base}, \mathbf{W}_{inp}, \mathbf{W}_{ds}$. The evolution formula is as follows:

$$\mathbf{W}_{new} = \mathbf{W}_{base} + \alpha \cdot \underbrace{(\mathbf{W}_{inp} - \mathbf{W}_{base})}_{\text{Inpainting Increment}} + \beta \cdot \underbrace{(\mathbf{W}_{ds} - \mathbf{W}_{base})}_{\text{Outfitting Increment}}, \tag{2}$$

where α and β are the balancing coefficients that account for the capability enhancements from SDXL-inpainting and DreamshaperXL. It is important to note that we directly copy the extra 5 channels in the *conv in* layer of the SDXL-inpainting into the merged model, multiplying them by α.

However, the optimal values of α and β are not apparent. We hope to determine the optimal α and β to ensure that the initial weight $\mathbf{W}_{new}$ achieve the best evaluation performance, i.e.

$$\underset{(\alpha,\beta)\in[0,2]^2}{\arg\min} \; f(\alpha,\beta) = \Phi\left(\mathbf{W}_{base} + \alpha \cdot (\mathbf{W}_{inp} - \mathbf{W}_{base}) + \beta \cdot (\mathbf{W}_{ds} - \mathbf{W}_{base})\right), \tag{3}$$

where Φ is a non-differentiable evaluation function. Empirically, we assume that f exhibits monotonic or convex properties with respect to the balancing coefficients (α, β) in most regions. Therefore, we discretize the continuous domain $[0, 2]^2$ into a grid with $\delta = 0.1$ as the step size and design the discrete greedy algorithm 1, to search for the optimal (α, β). In our algorithm, we selecte the CLIP score (42) on 20 fixed inpainting image-text pairs as the evaluation function Φ. The optimal solution obtained is $(\alpha, \beta) = (1.0, 1.1)$. Please refer to the Appendix C.1 for more explanations.

Algorithm 1 Discrete greedy algorithm

Require: Evaluation function f, step size δ.
Initialize $(\alpha, \beta) \leftarrow (0.5, 0.5)$
while True **do**
  $f_{\text{current}} \leftarrow f(\alpha, \beta)$
  $N \leftarrow \{(\alpha+\delta, \beta), (\alpha-\delta, \beta), (\alpha, \beta+\delta), (\alpha, \beta-\delta)\}$
  $N \leftarrow \{(\alpha', \beta') \in N \mid 0 \le \alpha' \le 2 \text{ and } 0 \le \beta' \le 2\}$
  $F_N \leftarrow \{f(\alpha', \beta') \mid (\alpha', \beta') \in N\}$
  if $\min(F_N) \ge f_{\text{current}}$ **then**
    break
  else
    $(\alpha, \beta) \leftarrow \arg\min_{(\alpha', \beta') \in N} f(\alpha', \beta')$
  end if
end while
return (α, β)

Adaptive Mask Boost. Previous works generally exhibit limited robustness in cross-category try-on scenarios, resulting in inaccurately rendered clothing styles as shown in Fig. 6 and 10. This is largely due to a dependence on agnostic masks derived from clothing parsing, which tends to leak the edges of the clothing shape during training. This leakage may cause the generated garments to almost entirely cover the agnostic mask region. In response to these limitations, we employ an intuitive and effective approach that significantly enhances the model's robustness about cross-category try-on, *i.e.*, the Adaptive Mask Boost strategy, which primarily comprises mask augmentation during training and

https://doi.org/10.52202/079017-2646

adaptive elongation during inference. Specifically, during training, the agnostic mask is extracted solely using OpenPose body joint detections *without leveraging human parsing*. We perform random elongation of the mask by a factor $f \sim \text{Uniform}(1.2, 1.5)$ with a probability of $P = 0.5$. This training setting forces the model to autonomously determine the optimal cloth length. During inference, we assess the aspect ratio σ of the bounding box of the laid-out garment. If $\sigma > 1.2$, we proportionally extend the agnostic area to match σ, creating an adaptive agnostic mask that conforms to the garment's style. Experiments have validated that AnyFit with Adaptive Mask Boost autonomously determines the appropriate garment length, yielding robust try-on results across different clothing categories.

Figure 4: Visual results on proprietary and in-the-wild data. Best viewed when zoomed in.

4 Experiments

4.1 Experimental setup

Datasets. Our experiments are carried out on two publicly available datasets, VITON-HD (12) (11647 training pairs) and DressCode (16) (48392 training pairs) using the official splits for training and testing, as well as an additional proprietary e-commerce dataset. This proprietary dataset contains 50602 training pairs and 2500 testing pairs of mainly upper-body person and upper garment images, featuring complex patterns, backgrounds, and postures, alongside a rich variety of styles including layered garment ensembles, which present a more challenging scenario. As for multi-garment try-on, we utilized a HumanParsing model (28; 29) to extract clothing items from DressCode and constructed triplets consisting of (upper-body garment, lower-body garment, model image). In these triplets, one garment is an original laid-out image, while the other is a cropped image from the person's image. Finally, we construct 24314 publicly available upper-lower triplets by garment crop from the 15363 upper-body pairs and 8951 lower-body pairs in DressCode, called DressCode-multiple. A subset of 1800 triplets is reserved as the test set. Please refer to the Appendix B for more details.

Figure 5: Visual comparisons on the DressCode-multiple. AnyFit exhibits an elegant integration between upper and lower garments, accurate length control, and appropriate overall styling.

Figure 6: Visual validation about model evolution and mask boost in (a), (c), (d). We also provide visual results about mask reliance in (b) found in previous work.

Implementation details. We initialize the AnyFit with Prior Model Evolution strategy described in Sec. 3.3, and fine-tune it using an AdamW optimizer (37) with a constant learning rate of $5e-5$. We train three variants of models on the VITON-HD, DressCode and proprietary dataset at a resolution of 1024×768, independently. Subsequently, we extend the model trained on DressCode to enable multi-garment try-ons by training on the DressCode-multiple dataset. All the models are trained for 150 epochs on 8 NVIDIA A100 GPUs with DeepSpeed (5) ZeRO-2 to reduce memory usage, at a batch size of 15. At inference time, we run AnyFit on a single NVIDIA RTX 3090 GPU for 30 sampling steps with the DDIM sampler (44). Outfitting dropout (50) is used with a guidance scale $s_g = 1.3$. The data augmentation follows the same protocol as in StableVTON (31). We also employ pretrained IP-Adapter (54) for SDXL. Please refer to the Appendix D for more details.

https://doi.org/10.52202/079017-2646

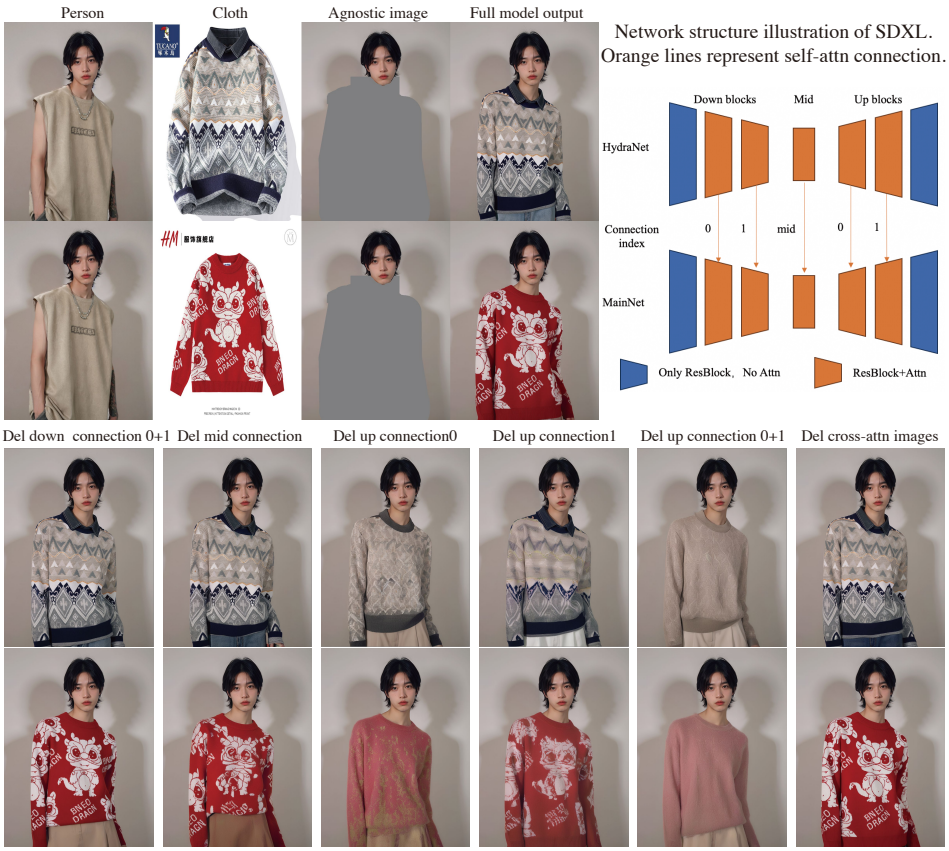


Figure 7: We separately cut off the self-attention injections between different blocks of HydraNet and MainNet, as well as the image features from IP-Adapter in cross-attention layers. The results show that the self-attention layers between the up blocks are the decisive factor affecting the performance.

Table 1: Quantitative comparisons on the VITON-HD (12) and DressCode (16).

Dataset	VITON-HD (12)				DressCode (16)			
Method	LPIPS ↓	SSIM ↑	FID ↓	KID ↓	LPIPS ↓	SSIM ↑	FID ↓	KID ↓
HR-VTON (35)	0.097	0.878	12.31	3.86	-	-	-	-
DCI-VTON (20)	**0.072**	0.892	8.76	0.92	-	-	-	-
StableVTON (31)	0.076	0.891	9.35	1.51	-	-	-	-
OOTDiffusion (50)	0.093	0.856	9.16	0.68	-	-	-	-
GP-VTON* (49)	0.083	0.892	9.17	0.93	0.051	0.921	5.88	1.28
LADI-VTON (38)	0.091	0.875	9.42	1.63	0.072	0.902	6.94	2.33
IDM (13)	0.078	0.881	9.12	1.03	0.046	**0.923**	5.32	1.24
AnyFit(ours)	0.075	**0.893**	**8.60**	**0.55**	**0.044**	0.904	**4.51**	**0.48**

Evaluation protocols. We measure reconstruction accuracy by LPIPS (56) and SSIM (48) in a paired setting provided ground truth images, and authenticity of unpaired synthesized images by FID (40) and KID (45) without ground truth. All evaluations are conducted at a resolution of 512×384.

Baselines. We compare our model on single try-on tasks on VITON-HD, DressCode and our proprietary dataset with previous baselines including HR-VTON (35), LADI-VTON (38), DCI-VTON (20), StableVITON (31), OOTDiffusion (50), and the state of the art IDM (13). We directly utilize their released pre-trained models. As for multi-garment try-on, we developed a two-stage IDM model as a strong baseline, referred to as IDM-2Stage, which dresses the upper and lower garments sequentially. Inspired by (10), we concatenate the upper and lower garments on width spatially and feed them to a single-conditional HydraNet for training as another baseline, termed VTON-concat. Finally, we compared AnyFit with Paint by Example (51), IDM-2Stage, and VTON-concat.

Table 2: Quantitative comparisons on the DressCode-multiple. The "Time" represents the inference time increase compared to its single-garment try-on.

Method	FID ↓	KID ↓	Time ↓
Paint-by-Example (51)	35.17	13.12	95%
IDM-2Stage (13)	21.47	7.85	93%
VTON-concat (10)	<u>21.11</u>	<u>7.30</u>	**8%**
AnyFit (ours)	**20.43**	**7.10**	<u>9%</u>

Table 3: Comparisons on proprietary dataset. *AnyFit (xxx)* is trained only on *xxx* dataset.

Method	FID ↓	KID ↓
LADI-VTON (38)	52.24	6.51
DCI-VTON (20)	57.96	12.35
StableVTON (31)	53.80	8.13
IDM (13)	48.76	4.35
AnyFit (VITON-HD)	<u>46.95</u>	<u>2.73</u>
AnyFit (proprietary)	**43.97**	**0.69**

4.2 Qualitative results

Single-garment try-on. Fig.3 and 4 provide a qualitative comparison between AnyFit and the baselines on VITON-HD, the more challenging proprietary and in-the-wild data, covering open-garment and layering rendering scenarios. For a fair comparison with the baselines, we include results of AnyFit trained on VITON-HD. AnyFit excels in retaining intricate pattern details, owing to the effective collaboration between HydraNet and the IP-Adapter. It also maintains the correct silhouette of the clothing at a semantic level. This suggests that, through Mask Boost, AnyFit enhances the recollection of the original shape of the clothing, while other models, influenced by the mask, tend to generate incorrect appearances. The Prior Model Evolution further strengthens the texture representation of the apparel. Notably, when trained on the proprietary dataset, AnyFit automatically fills in inner garments or unzips clothing based on posture, a capability absent in the version trained on VITON-HD due to the lack of such training data.

Multi-garment try-on. Fig. 5 offers a qualitative comparison for multi-garment try-ons using the compiled DressCode-multiple dataset. Firstly, AnyFit demonstrates high-fidelity cloth preservation. Importantly, thanks to the distinct and individual Hydra-Blocks situated in different condition branches, AnyFit accurately depicts the demarcation between the upper and lower garments, showcasing a reasonable transition at the interconnection. In contrast, VTON-concat mishandles the relative clothing sizes after concatenation, leading to garment distortion and blurring. Meanwhile, IDM-2Stage faces artifacts at the juncture of the upper and lower garments, because it obscures parts of one garment while trying on another. Remarkably, despite training with one garment presented as a flat lay image and the other as a warped cloth cropped from a person image, AnyFit remains strikingly robust when faced with both garments presented as flat lays during inference.

4.3 Quantitative results

As indicated in Tab. 1 2 3, extensive experiments conducted on VITON-HD (12), DressCode (16), the proprietary dataset, and DressCode-multiple consistently prove that AnyFit significantly surpasses all baselines. This confirms AnyFit's capability to deliver superior try-on quality in both single-garment and multi-garment tasks across various scenes. Moreover, we note that AnyFit shows considerable improvement in unpaired settings in terms of the FID and KID metrics, demonstrating our model's robustness for cross-category try-ons. For more results, please refer to the Appendix E.

4.4 Ablation study

Hydra Blocks. To validate our proposed Hydra Blocks, we directly employ a singular conditioned HydraNet (which degenerates to ReferenceNet (26) actually) as the baseline "w/o Hydra Block" to encode both the top and bottom garment conditions concurrently, and then concatenate them into MainNet. As illustrated in Tab. 4, Fig. 8 and 11, a model lacking the Hydra Block tends to produce artifacts at the junction of the top and bottom garments. Such models also frequently allow the features of one garment to influence the other, leading to incorrect clothing styles. However, with the introduction of the Hydra Block, AnyFit consistently exhibits more stable results.

Prior Model Evolution. We qualitatively demonstrate the effects of the Prior Model Evolution in Fig. 13 and 6 (a). The SDXL-evolved model reduces artifacts and enhances robustness significantly, while outputs without Prior Model Evolution typically feature oversaturated colors as well as lighting and shadows that do not harmonize with the background. The gradual enhancement of model capabilities is visualized in Fig. 6(c). We also empirically and quantitatively validate the effectiveness

https://doi.org/10.52202/079017-2646

Table 4: Quantitative ablation study.

Dataset	Method	LPIPS ↓	SSIM ↑	FID ↓	KID ↓
DressCode-multiple	- w/o Hydra Blocks	-	-	22.48	8.02
	- w/o Prior Model Evolution	-	-	21.35	7.58
	Full AnyFit	-	-	**20.43**	**7.10**
the proprietary dataset	- w/o Adaptive Mask Boost	0.183	**0.748**	44.75	1.44
	- w/o Prior Model Evolution	0.192	0.740	45.01	1.52
	Full AnyFit	**0.181**	0.743	**43.97**	**0.69**

Figure 8: Visual ablation study. Without Prior Model Evolution, AnyFit suffers reduced fabric detail and less realistic textures. While Hydra Blocks improve intersections of upper and lower garments.

of the Prior Model Evolution strategy after training on the Virtual Try-On (VTON) task in Fig. 8 and Tab. 4. The Prior Model Evolution, by improving the model's initial capabilities, lessens the difficulty of learning and fosters a dramatic improvement in outfitting capacity and logo fidelity.

Adaptive Mask Boost. We illustratively showcase the issues of information leakage and mask reliance found in previous methods in Fig. 6 (b) and Fig. 10. Additionally, we empirically and quantitatively validate the effectiveness of the Adaptive Mask Boost strategy in Table 4 and Fig. 10. This strategy significantly heightens the model's robustness towards different categories of clothing, enabling the autonomous determination of appropriate garment length rather than relying on masks. Furthermore, we manually adjust the aspect ratios σ in Fig. 6 (d), which demonstrates the positive impact of adaptive elongation during inference. More ablation studies are detailed in the Appendix A.

5 Conclusion

We introduce AnyFit, a novel and robust VTON pipeline suitable for any combination of attire across any imaginable scenario, offering a revolutionary leap in realistic try-on effects. To support multi-garment try-ons, AnyFit constructs HydraNet with lightweight and scalable parallelized attention that facilitates the feature injection of multiple garments. Observing artifacts in real-world scenarios, we evolve its potential by synthesizing the residuals of multiple models, as well as implementing a mask region boost strategy. Comprehensive experiments on high-resolution benchmarks and real-world data have demonstrated that AnyFit significantly surpasses all baselines by a large gap.

Broader impacts. With the ability to synthesize images, arises the risk that AnyFit might be used for inappropriate purposes such as producing media that breaches intellectual property rights or privacy norms. Because of these risks, we strongly advocate for the conscientious use of this technology.

Limitation and future work. Our approach exhibits excellent performance in single-garment and multi-garment virtual try-on applications. However, it still faces some limitations. Firstly, it shares the shortcomings of large text-image models, sometimes showing instability in generating hands with complex structures. Secondly, our model offers initial but not yet fully mature text control capabilities (for details, please refer to the Appendix A), providing opportunities for future enhancements.

https://doi.org/10.52202/079017-2646

Acknowledgements

This work was supported by National Science Foundation of China (U20B2072, 61976137). This work was also partly supported by SJTU Medical Engineering Cross Research Grant YG2021ZD18.

References

[1] Stable diffusion xl shows worse inpainting. https://www.reddit.com/r/StableDiffusion/comments/166bz7b/sdxl_base_model_for_inpainting_way_worse_than_15/

[2] Discussions about model synthesis. https://www.reddit.com/r/StableDiffusion/comments/zcby0o/you_can_now_merge_inpainting_and_regular_models/

[3] Dreamshaper xl. https://civitai.com/models/112902/dreamshaper-xl

[4] Sdxl-inpainting-0.1. https://huggingface.co/diffusers/stable-diffusion-xl-1.0-inpainting-0.1

[5] Deepspeed. https://github.com/microsoft/DeepSpeed

[6] 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)

[7] Bookstein, F.L.: Principal warps: Thin-plate splines and the decomposition of deformations. IEEE Transactions on pattern analysis and machine intelligence (1989)

[8] Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2017)

[9] Chang, D., Shi, Y., Gao, Q., Fu, J., Xu, H., Song, G., Yan, Q., Yang, X., Soleymani, M.: Magicdance: Realistic human dance video generation with motions & facial expressions transfer. arXiv preprint arXiv:2311.12052 (2023)

[10] Chen, M., Chen, X., Zhai, Z., Ju, C., Hong, X., Lan, J., Xiao, S.: Wear-any-way: Manipulable virtual try-on via sparse correspondence alignment. arXiv preprint (2024)

[11] 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)

[12] Choi, S., Park, S., Lee, M., Choo, J.: Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2021)

[13] Choi, Y., Kwak, S., Lee, K., Choi, H., Shin, J.: Improving diffusion models for virtual try-on. arXiv preprint arXiv:2403.05139 (2024)

[14] Chopra, A., Jain, R., Hemani, M., Krishnamurthy, B.: Zflow: Gated appearance flow-based virtual try-on with 3d priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)

[15] 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)

[16] Davide, M., Matteo, F., Marcella, C., Federico, L., Fabio, C., Rita, C.: Dress code: High-resolution multi-category virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)

[17] Dong, H., Liang, X., Zhang, Y., Zhang, X., Shen, X., Xie, Z., Wu, B., Yin, J.: Fashion editing with adversarial parsing learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8120–8128 (2020)

 https://doi.org/10.52202/079017-2646

[18] 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)

[19] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems (2014)

[20] 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)

[21] Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018)

[22] Han, X., Hu, X., Huang, W., Scott, M.R.: Clothflow: A flow-based model for clothed person generation. In: Proceedings of the IEEE/CVF international conference on computer vision (2019)

[23] Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018)

[24] He, S., Song, Y.Z., Xiang, T.: Style-based global appearance flow for virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)

[25] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems (2020)

[26] Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., Bo, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. arXiv preprint arXiv:2311.17117 (2023)

[27] 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)

[28] Jin, Z.: Sssegmenation: An open source supervised semantic segmentation toolbox based on pytorch. arXiv preprint arXiv:2305.17091 (2023)

[29] Jin, Z., Hu, X., Zhu, L., Song, L., Yuan, L., Yu, L.: Idrnet: Intervention-driven relation network for semantic segmentation. Advances in Neural Information Processing Systems **36** (2024)

[30] Jo, Y., Park, J.: Sc-fegan: Face editing generative adversarial network with user's sketch and color. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1745–1753 (2019)

[31] Kim, J., Gu, G., Park, M., Park, S., Choo, J.: Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)

[32] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

[33] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)

[34] Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5549–5558 (2020)

[35] 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)

[36] Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2017)

[37] Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam (2018)

[38] 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)

[39] Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)

[40] Parmar, G., Zhang, R., Zhu, J.Y.: On aliased resizing and surprising subtleties in gan evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)

[41] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)

[42] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning (2021)

[43] 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)

[44] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

[45] Sutherland, J., Arbel, M., Gretton, A.: Demystifying mmd gans. In: International Conference for Learning Representations (2018)

[46] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems (2017)

[47] Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., Yang, M.: Toward characteristic-preserving image-based virtual try-on network. In: Proceedings of the European conference on computer vision (ECCV) (2018)

[48] 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 (2004)

[49] 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)

[50] 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)

[51] 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)

[52] 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)

[53] Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4210–4220 (2023)

[54] Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)

 https://doi.org/10.52202/079017-2646

[55] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)

[56] 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 (2018)

[57] 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. arXiv preprint arXiv:2405.00448 (2024)

[58] 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)

Appendix

In Section A, we provide more ablation studies about the role of text prompts in AnyFit, Hydra Block, Adaptive Mask Boost strategy and Prior Model Evolution. In Section B, we provide additional details about the proprietary dataset and the multi-garment VTON dataset. In Section C, we provide additional details about discrete greedy algorithm during Prior Model Evolution, and preliminary knowledge about stable diffusion. In Section D, we provide additional details about experimental setup, including data augmentation and outfitting dropout. Lastly, in Section E, we present a multitude of AnyFit generated images, including single-garment visual try-ons and multi-garment virtual try-ons, with additional results displayed in challenging scenarios.

A More ablation study

A.1 The role of text prompt in AnyFit

In fact, we have discovered that text plays a certain role in controlling the overall try-on style. As show in Fig. 9, by adjusting the prompt, AnyFit is able to achieve variations in Virtual Try-On (VTON) apparel styles. However, this form of control is unstable, leaving space for further exploration.

Figure 9: By adjusting the prompt, AnyFit is able to achieve variations in VTON apparel styles.

A.2 More ablation study about Adaptive Mask Boost strategy

The Adaptive Mask Boost strategy primarily comprises mask augmentation during training and adaptive elongation during inference. The role of adaptive elongation has been thoroughly illustrated in Fig. 6 in the main body, where it addresses the challenge of try-ons for long garments. Here, we predominantly discuss the function of parsing-free mask augmentation during training.

Specific to the implementation, during training, the agnostic mask is extracted solely using OpenPose body joint detections to occlude the original clothing, without leveraging human parsing, and random elongation of the mask is performed. Consequently, the shape of the mask is correlated only with the human pose and not with the clothing. **The significance of this procedure lies in eliminating the possibility of the mask leaking clothing information, ensuring that the model cannot cheat from the shape of the mask**. From Fig. 10, we observe that previous models often suffer from leakage of clothing information, where the models tend to generate garments that fill the entire mask areas. In real-world inference scenarios, it is challenging to provide an entirely accurate mask region.

https://doi.org/10.52202/079017-2646

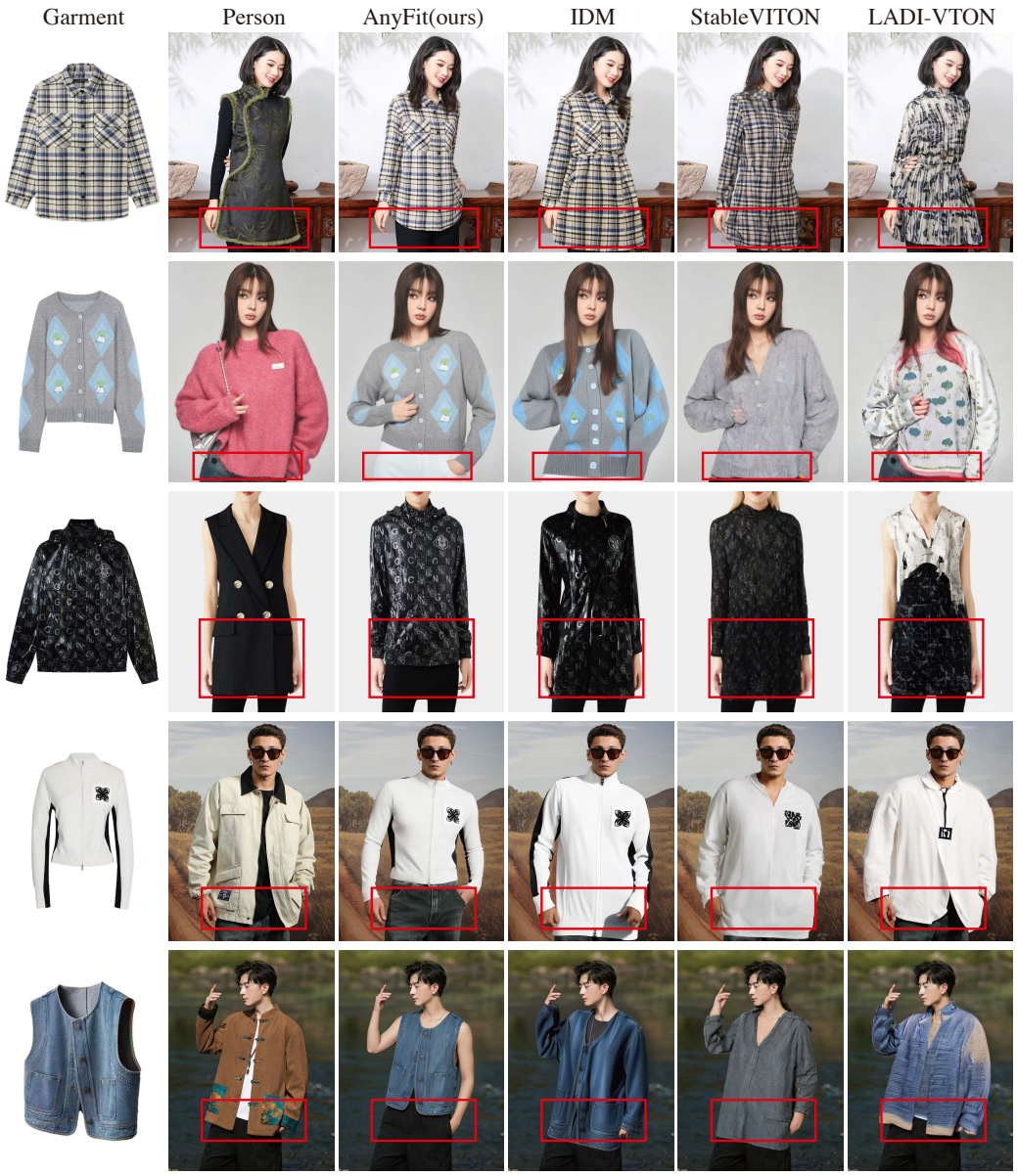


Figure 10: Visual validation of the role of parsing-free mask augmentation during training within Adaptive Mask Boost strategy. AnyFit is only trained on VITON-HD for a fair comparison. Previous methods often suffer from leakage of clothing information, where the model tends to generate garments that fill the entire mask area, while AnyFit autonomously determines the correct length of the garments, producing attractive try-on results.

In contrast, our model is free from this predicament. Through mask augmentation during training, AnyFit autonomously determines the correct length of the garments, producing attractive try-on results.

A.3 More ablation study about Hydra Block

In Fig. 11, we present additional visual results to validate the effectiveness of the Hydra Block. "w/ Hydra Block" represents our fully proposed AnyFit, whereas the version "w/o Hydra Block" omits the Hydra Encoding Block as well as the positional embeddings within the Hydra Fusion. It directly

Figure 11: Visual comparisons on the DressCode-multiple. Model lacking the Hydra Block is more prone to producing artifacts at the junction of the top and bottom garments.

employs a singular conditioned HydraNet (which, in this scenario, degenerates to ReferenceNet) to encode both the top and bottom garment conditions concurrently, and then injects them into MainNet.

As illustrated in Fig. 11, model lacking the Hydra Block is more prone to producing artifacts at the junction of the top and bottom garments. They also tend to allow the features of the bottom garment to influence the top, resulting in incorrect clothing styles, such as overly long tops or erroneous stripe patterns. After equipped with the Hydra Block, AnyFit noticeably exhibits more stable results.

https://doi.org/10.52202/079017-2646

A.4 More ablation study about Prior Model Evolution

In the main body of the paper, we have demonstrated the effects of Prior Model Evolution. However, we also notice that the capabilities of large-scale text-to-image models are closely related to the value of their classifier-free guidance (CFG). For simplicity, in the main body of the paper, all images are generated using a CFG value of 3.0 in Fig. 6 when discussing the comparisons related to Prior Model Evolution. Here, we provide a more granular comparison concerning the presence or absence of the Prior Model Evolution strategy. The results presented here are all based on models directly initialized with synthetic weights, without any training.

As depicted in Fig. 13, an increase in the CFG value leads to a generalized oversaturation in the images. Within a reasonable range of CFG values, models incorporating Prior Model Evolution exhibit more realistic fabric textures and more plausible inpainting results. In contrast, outputs without Prior Model Evolution typically feature oversaturated colors and the absence of detailed wrinkles, as well as lighting and shadows that do not harmonize with the background. This illustrates the SDXL-inpainting-0.1 (41) model diminishes the exemplary text-to-image capabilities of the original SDXL-BASE model, resulting in more mediocre outcomes. We attribute this degradation to the disruption of the previously well-aligned correspondences between text and images during the inpainting pre-training phase. However, our model with Prior Model Evolution, significantly ameliorates this issue, enhancing its overall robustness. We actually find that **after a minimal amount of fine-tuning, the model's synthetic outputs become substantially more powerful**.

B Datasets

In this section, we provide a detailed description of the proprietary dataset and the multi-garment VTON dataset, *i.e.*, DressCode-multiple, which is constructed based on the publicly available DressCode (16) dataset.

B.1 The proprietary dataset

Our proprietary dataset comprises 50,602 training image pairs and 2,500 testing pairs. Each pair consists of a flat-laid garment image and a frontal upper body model image, with most model images featuring complex backgrounds. The proprietary dataset is collected from e-commerce websites, with attention paid to achieving a balanced distribution of clothing categories. We gathered apparel from 12 different categories, including both menswear and womenswear, across various seasons such as spring, summer, autumn, and winter. As illustrated in Fig. 12, the person images display diverse backgrounds. We standardized the image resolution to 1024x768 and conducted preprocessing on this basis to test different models. This preprocessing involved using Densepose (21) and OpenPose (8) to extract human features, employing the SAM (33) model to extract the main garment body, and constructing the gnostic mask images, *etc*.

Figure 12: Examples of the proprietary dataset.

Figure 13: Visual validation of the role of Prior Model Evolution in various CFG weights without any training. Outputs without Prior Model Evolution typically feature oversaturated colors and the absence of detailed wrinkles, as well as lighting and shadows that do not harmonize with the background. Best viewed when zoomed in.

https://doi.org/10.52202/079017-2646

Figure 14: Examples of the DressCode-multiple dataset.

B.2 The DressCode-multiple dataset

To facilitate research on multi-garment virtual try-on, we require a dataset composed of image triplets, each containing an upper garment image, a lower garment image, and a model image wearing the corresponding garments. However, obtaining such data with strict alignment is challenging. Leveraging the DressCode dataset (16), which includes upper and lower garment data as well as full-body model images, we set out to construct the DressCode-multiple dataset, consisting of triplets, as illustrated in Fig. 14. Assuming we start with the upper garment data, where we already have a flat lay upper garment image and a model image wearing the corresponding upper garment, we use human parsing techniques (28; 29) to roughly segment the lower garment portion and extract it from the model image to serve as the corresponding lower garment image. At this point, the triplet consists of (flat lay upper, cropped lower, model image). Similarly, triplets derived from the lower garment data result in a composition of (cropped upper, flat lay lower, model image). Using this approach, we construct 24,314 publicly available upper-lower triplets by cropping garments from the 15,363 upper-body pairs and 8,951 lower-body pairs for training. For testing, we take the flat-laid garments from the upper and lower garment test sets, shuffle them, and combine them randomly, pairing them

with the test models from the upper garment data to create (flat lay upper, flat lay lower, model image) configurations for unpaired, real-world multi-garment virtual try-on tests. The paired testing is not conducted due to the lack of ground truth triplets.

It is worth mentioning that our model has not encountered triplets in the form of (flat lay upper garment, flat lay lower garment, model image) during training. However, it has already learned the correct way to wear upper and lower garments through training on cropped images and demonstrates good robustness. We believe that if real triplets consisting of (flat lay upper garment, flat lay lower garment, model image) were available for training, the model would exhibit even better performance.

C Method supplement

C.1 Discrete greedy algorithm

Our objective is to determine the optimal balancing coefficients α and β to ensure that the initial weight $\mathbf{W}_{new}$ achieve the best evaluation performance, i.e.

$$\underset{(\alpha,\beta)\in[0,2]^2}{\arg\min}\ f(\alpha,\beta) = \Phi\left(\mathbf{W}_{base} + \alpha\cdot(\mathbf{W}_{inp} - \mathbf{W}_{base}) + \beta\cdot(\mathbf{W}_{ds} - \mathbf{W}_{base})\right)$$

Note that the function Φ is non-differentiable, which precludes the use of gradient-based optimization algorithms to find the minimum point. Empirically, we find that the evaluation function f exhibits monotonic or convex properties with respect to the balancing coefficients (α, β) in most regions. Therefore, we discretize the continuous domain $[0, 2]^2$ into a grid with δ as the step size and design the following algorithm, inspired by the greedy method, to search for the optimal (α, β).

Algorithm 2 Discrete greedy algorithm

Require: Evaluation function f, step size δ.
Initialize $(\alpha, \beta) \leftarrow (0.5, 0.5)$
while True **do**
  $f_{\text{current}} \leftarrow f(\alpha, \beta)$
  $N \leftarrow \{(\alpha+\delta, \beta), (\alpha-\delta, \beta), (\alpha, \beta+\delta), (\alpha, \beta-\delta)\}$
  $N \leftarrow \{(\alpha', \beta') \in N \mid 0 \le \alpha' \le 2 \text{ and } 0 \le \beta' \le 2\}$ ▷ Filter out-of-bound neighbors
  $F_N \leftarrow \{f(\alpha', \beta') \mid (\alpha', \beta') \in N\}$ ▷ Compute f values for valid neighbors
  if $\min(F_N) \ge f_{\text{current}}$ **then**
    break
  else
    $(\alpha, \beta) \leftarrow \arg\min_{(\alpha', \beta')\in N} f(\alpha', \beta')$
  end if
end while
return (α, β)

In this algorithm, we initialize (α, β) at $(0.5, 0.5)$. During each iteration, we calculate the evaluation value f at the current point and four adjacent points in the directions up, down, left, and right. The algorithm then updates the current point to the one with the minimum evaluation value among these adjacent points. The termination condition of the algorithm is met when the evaluation values of all neighbors are greater than that of the current point. This method can be viewed as a discrete version of gradient descent method. Please note that under the assumption that f is a convex or monotonic function, the algorithm is guaranteed to converge to the global optimal solution within the discrete parameter space.

C.2 Preliminary

Stable Diffusion. Our AnyFit is an extension of Stable Diffusion (43), which is one of the most commonly used latent diffusion models. Stable Diffusion employs a variational autoencoder (32) (VAE) that consists of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$ to enable image representations in the latent space. And a UNet ϵ_θ is trained to denoise a Gaussian noise ϵ with a conditioning input encoded by a CLIP text encoder (42) τ_θ. Given an image $\mathbf{x}$ and a text prompt $\mathbf{y}$, the training of the denoising UNet ϵ_θ is performed by minimizing the following loss function:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathcal{E}(\mathbf{x}),\mathbf{y},\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \tau_\theta(\mathbf{y}))\|_2^2\right], \tag{4}$$

https://doi.org/10.52202/079017-2646

Table 5: Full quantitative comparisons on the proprietary dataset. AnyFit *(xxx)* represents AnyFit trained on the corresponding *xxx* dataset.

Method	LPIPS ↓	SSIM ↑	FID ↓	KID ↓
LADI-VTON (38)	0.252	0.734	52.24	6.51
DCI-VTON (20)	0.264	0.734	57.96	12.35
StableVTON-base (31)	0.245	0.694	54.70	8.44
StableVTON-repainting (31)	0.242	0.720	53.80	8.13
IDM (13)	0.247	0.701	48.76	4.35
AnyFit (VITON-HD)	<u>0.200</u>	<u>0.740</u>	<u>46.95</u>	<u>2.73</u>
AnyFit (proprietary)	**0.181**	**0.743**	**43.97**	**0.69**

where $t \in \{1, ..., T\}$ denotes the time step of the forward diffusion process, and $\mathbf{z}_t$ is the encoded image $\mathcal{E}(\mathbf{x})$ with the added Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$ (*i.e.*, the noise latent). Note that the conditioning input $\tau_\theta(\mathbf{y})$ is correlated with the denoising UNet by the cross-attention mechanism (46).

D Experimental details

Data augmentation. We have implemented data augmentation techniques that could potentially enhance the model's generalization ability as well as its color accuracy performance. Specifically, the data augmentation operations include (a) horizontal flipping of images, (b) resizing garments and human figures through padding (up to 10% of the image size), (c) randomly adjusting the image's hue within a range of -5 to +5, and (d) randomly adjusting the image's contrast within a specified range (between 0.8 and 1.2 times the original contrast). Each of these operations occurs independently with a 50% probability. Moreover, these operations are simultaneously applied to both the garment and model images.

Outfitting dropout. Following the OOTDiffusion approach, we applied Outfitting Dropout, which essentially acts as a form of image-conditioned classifier-free guidance. It enhances the contrast and sharpness of the generated images. Specifically, during the training process of our MainNet, we randomly drop the input garment latent as $E(g) = \varnothing$, where $\varnothing \in \mathbb{R}^{4\times h\times w}$ refers to an all-zero latent. In this way, the denoising UNet is trained both conditionally and unconditionally. Then at inference time, we simply use a guidance scale $s_g \geq 1$ to adjust the strength of conditional control over the predicted noise:

$$\hat{\epsilon}_\theta(\mathbf{z}_t, \omega_{\theta'}(\mathcal{E}(\mathbf{g}))) = \epsilon_\theta(\mathbf{z}_t, \varnothing) + s_\mathbf{g} \cdot (\epsilon_\theta(\mathbf{z}_t, \omega_{\theta'}(\mathcal{E}(\mathbf{g}))) - \epsilon_\theta(\mathbf{z}_t, \varnothing)). \tag{5}$$

In practice, we empirically set the outfitting dropout ratio to 10% in training, and the guidance scale $s_\mathbf{g}$ to 1.2 as default. We exclusively employ Outfitting Dropout within the Hydra Fusion Block that bridges HydraNet and MainNet. We have observed that applying outfitting dropout elsewhere introduces pixel-level artifacts.

E More experiment results

E.1 Full comparison on the proprietary dataset

Due to space constraints, we have not presented quantitative comparisons on the proprietary dataset in a paired setting within the main body of the paper; they are additionally reported in Tab. 5. Furthermore, we showcase additional results in Fig. 15, which include outcomes from state-of-the-art baselines and AnyFit trained on both the proprietary dataset and VITON-HD. Our method exhibits a considerable lead in performance.

E.2 More visual results

We provide more visual results on VITON-HD, DressCode, and proprietary dataset for inspection in Fig. 16, 17 and 18. Fig. 11 provides more results about multi-garment try-ons.

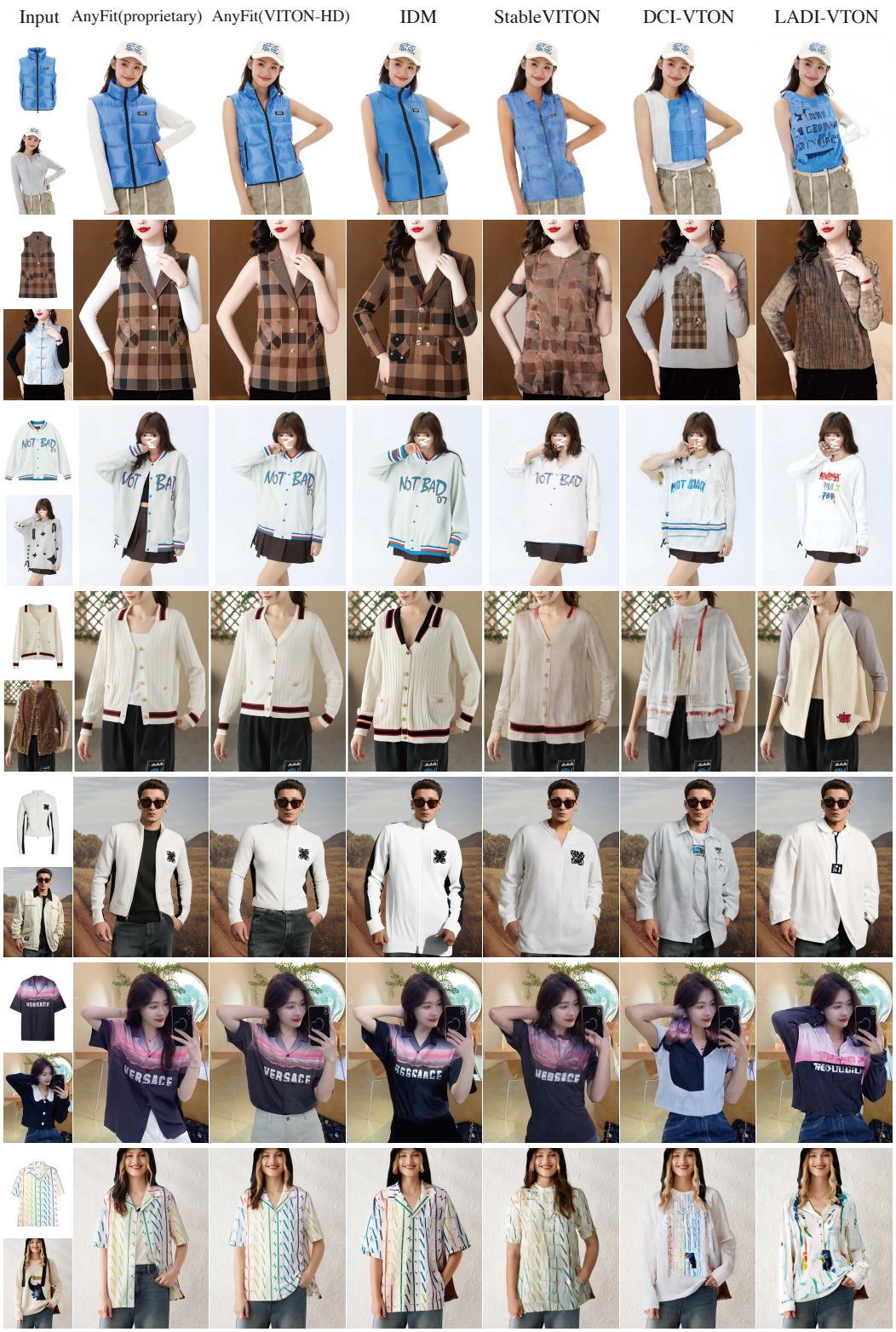


Figure 15: More visual comparisons on the proprietary dataset. AnyFit displays superior garment details and outfit styling. Best viewed when zoomed in.

https://doi.org/10.52202/079017-2646

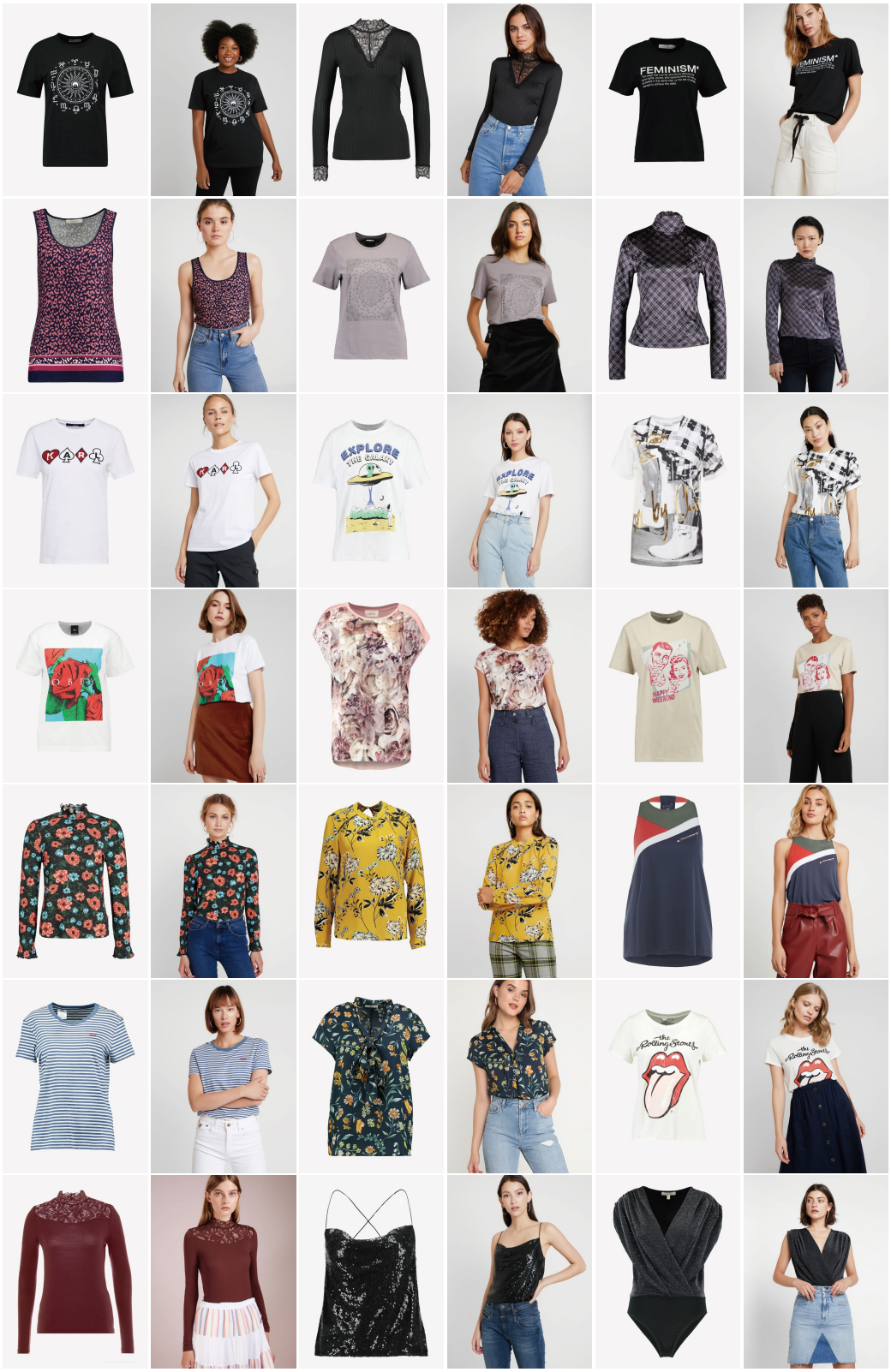


Figure 16: More visual results on the VITON-HD test data by AnyFit trained on VITON-HD training data. Best viewed when zoomed in.

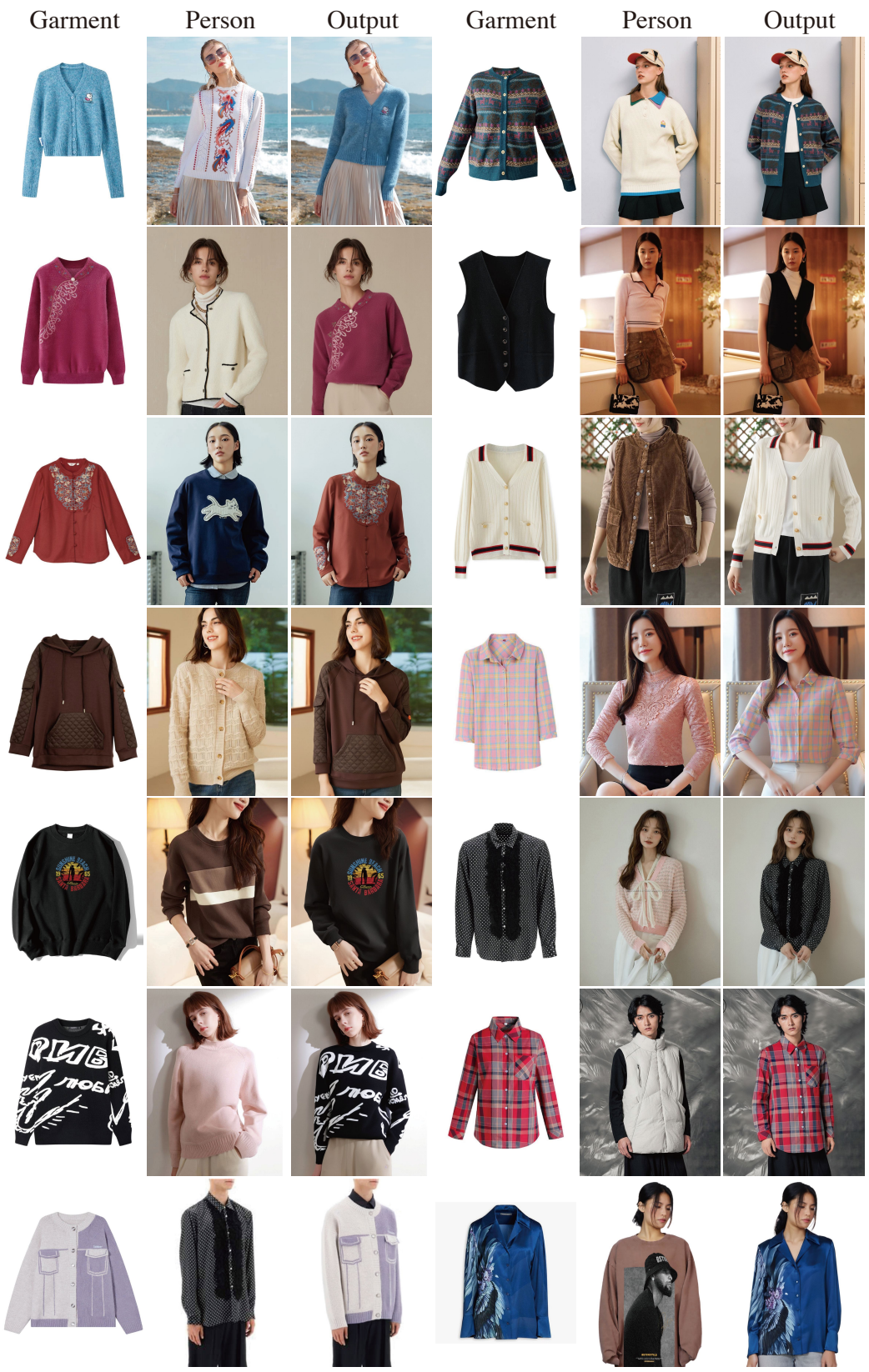


Figure 17: More visual results on the proprietary test data by AnyFit trained on proprietary training data. Best viewed when zoomed in.

https://doi.org/10.52202/079017-2646

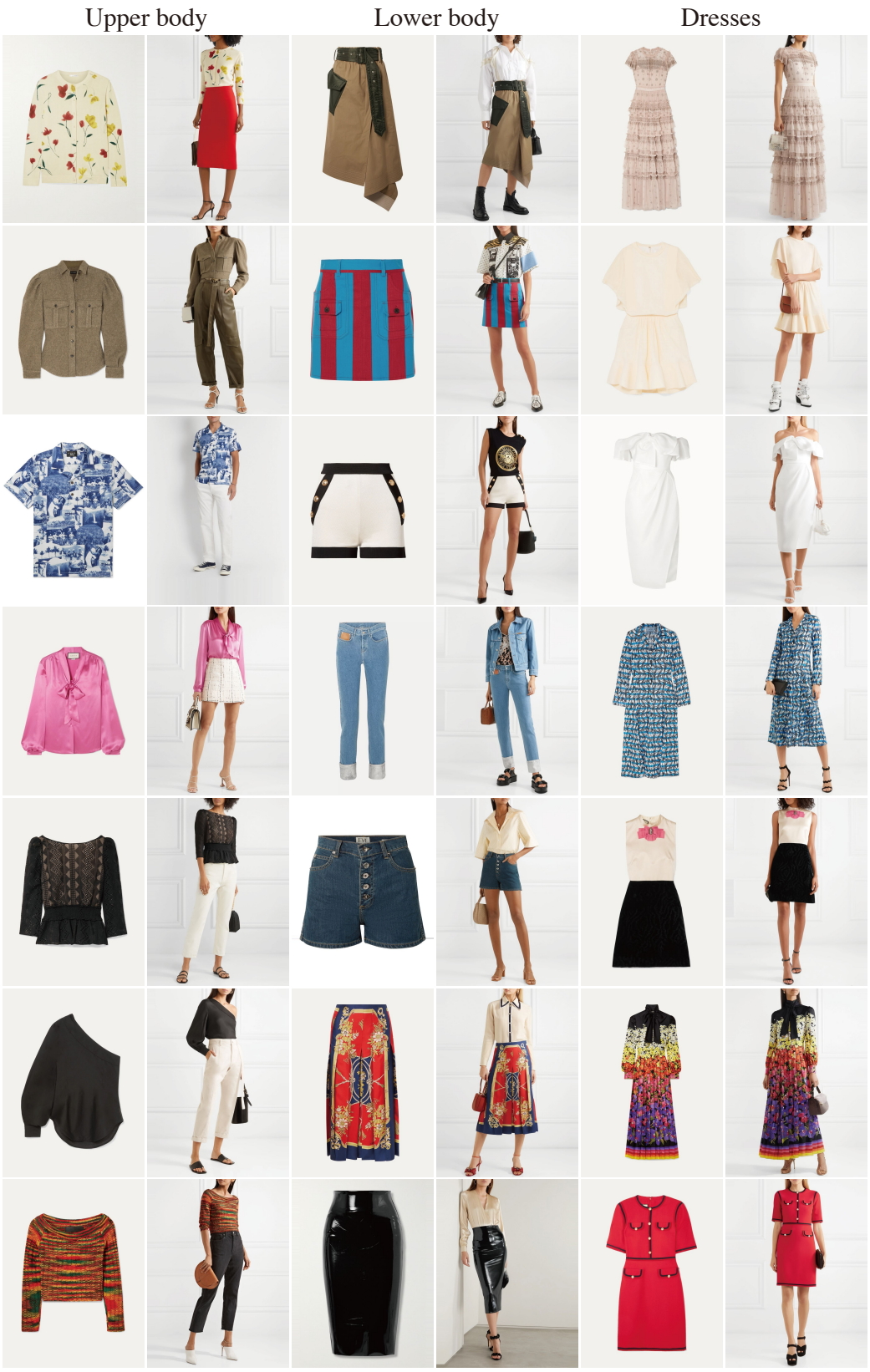


Figure 18: More visual results on the DressCode test data by AnyFit trained on DressCode training data. Best viewed when zoomed in.

https://doi.org/10.52202/079017-2646

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made. And the claims match theoretical and experimental results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to the limitations part in the main paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

https://doi.org/10.52202/079017-2646

Justification: All theoretical results are accompanied by appropriate proofs and proper citations. Every claim is substantiated by empirical evidence or supported by referenced literature.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Our contribution is a novel model architecture along with associated model enhancement strategies. All parameters and operational steps are detailed in the Experimental Section and the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We utilized two open-source datasets in addition to one proprietary dataset. Should our paper be accepted, we plan to make efforts to release a portion of the proprietary data to the public. While the code for the paper is not open-sourced, we have provided comprehensive instructions necessary for replication.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have provided all the training and test details. Please refer to the "Experimental Setup" part in the main paper and other related sections in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We adhered to the common evaluation practices established by prior work in our field. Error bars are not reported because it would be too computationally expensive. We believe that the metrics reported in main paper prove the efficacy of the model we proposed.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

https://doi.org/10.52202/079017-2646

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the sufficient information on the computer resources for each experiment in "Experimental Setup" part.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper fully conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to the broader impacts part in the main paper.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no risks for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We follow the instructions by the creators of each asset. We also cite the original paper that produced the code package or dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

https://doi.org/10.52202/079017-2646

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-2646