PROXYFIT: POINT CORRESPONDENCES AS PROXY REPRESENTATIONS FOR WEARING-STYLE CONTROL IN VIRTUAL TRY-ON

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion-based virtual try-on transfers garment appearance with high fidelity, but offers limited control over *how* a garment is worn. Text instructions leave exact target positions ambiguous, while existing spatial interfaces restrict garment-part coverage or provide insufficient control adherence. We introduce **ProxyFit**, which represents the intended wearing state through editable *proxy point mappings* from a flat garment to the body, enabling precise interaction by dragging target points. Our first contribution is an automated data pipeline that matches mesh-derived garment and person points, filters unreliable mappings and images, and constructs *contrastive try-on pairs* with the same person and garment but different wearing states. Our second contribution is a dedicated proxy representation branch: Fold extracts person-point features, attention retrieves local garment features, and ReFold writes them back at the specified body locations. This explicit routing improves adherence to user-specified point targets while preserving local garment appearance. Experiments demonstrate strong performance in both precise point control and standard virtual try-on. Across front opening, sleeve placement, hem placement, and fit, ProxyFit achieves 82.8% mean spatial control accuracy, exceeding WAW-style by 9.5 percentage points. It reduces point-placement error by 32.1% relative to Qwen-Image-Edit, while achieving LPIPS of 0.201 and FID of 44.35 on the standard VTON task.

1 INTRODUCTION

Image-based virtual try-on (VTON) renders a target person wearing a garment specified by a product image (Han et al., 2018; Choi et al., 2021; Morelli et al., 2022). Diffusion-based systems increasingly preserve garment texture, person identity, and scene appearance (Zhu et al., 2023; Kim et al., 2024; Choi et al., 2024; Xu et al., 2025), yet leave users with limited ability to precisely control the wearing state. We pursue *highly controllable virtual try-on*: beyond selecting a garment, users should be able to specify and edit the intended placement of its parts on the body while preserving garment fidelity and person identity. We formulate this task through *proxy point mappings* that link garment parts to requested body locations. The model must interpret these spatial relations and faithfully realize the edited targets. Front opening, sleeve roll height, hem tucking, and garment fit are representative instances of this broader task.

Existing interfaces only partially support this task (Figure 1). Text-conditioned VTON systems describe categorical styles (Zhang et al., 2024; Kim et al., 2025), and general editors such as Qwen-Image-Edit accept open-ended instructions (Wu et al., 2025). However, describing a desired change does not uniquely specify the target location of each garment part, and text-guided editing can introduce undesired changes to otherwise unedited content (e.g., the pose changes produced by Qwen-Image-Edit in Figure 4). MOFA-VTON derives upper- and lower-body regions from a user-drawn curve and adjusts their layouts through specialized branches (Han et al., 2026); this two-region design limits control beyond the upper/lower garment boundary. Wear-Any-Way offers a more general spatial interface through sparse garment–person correspondences (Chen et al., 2024). These points guide attention, but conditioning on a correspondence does not ensure that the associated garment features are placed at its target. The remaining challenge is to combine flexible point interaction with reliable, spatially precise execution.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

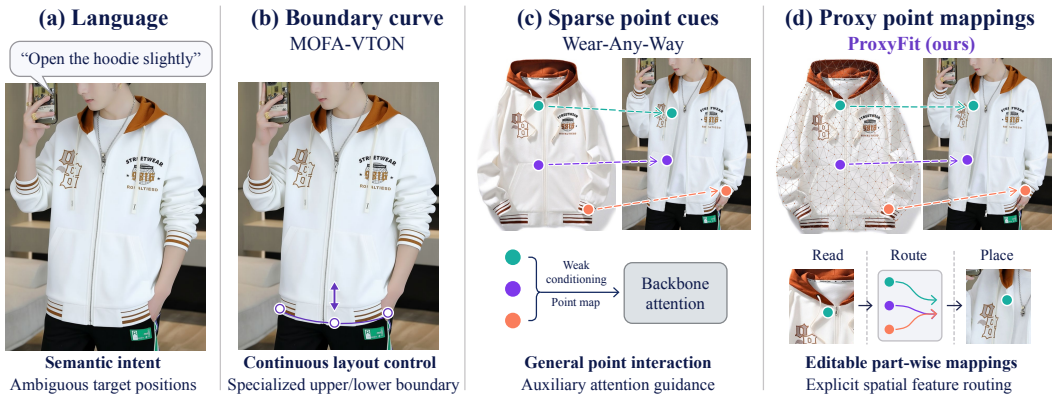


Figure 1: Interfaces for wearing-style control. (a) Language expresses semantic intent but leaves target positions ambiguous. (b) MOFA-VTON controls the upper/lower garment boundary. (c) Wear-Any-Way supplies sparse correspondences as attention guidance. (d) ProxyFit uses editable proxy point mappings to explicitly read, route, and place garment features.

We address this challenge with **ProxyFit**. An automated data pipeline constructs contrastive try-on pairs that share the person and garment but differ in wearing style. It triangulates the flat and worn garment regions, matches mesh-derived candidate points with DINO features, and filters unreliable mappings and image pairs. A dedicated proxy representation branch then executes the retained mappings inside the denoising network. Person-point features retrieve information from local neighborhoods around the selected garment points, and the retrieved features are written back at the requested body locations. Training on paired states with shared noise and timesteps encourages the model to respond to changes in these mappings under the standard denoising objective.

To our knowledge, ProxyFit is the first VTON framework to utilize a dedicated proxy branch that explicitly reads garment features around source points and writes them to user-specified target locations. This design enables precise, coordinated editing of multiple garment parts through a single point-based interface. Our results show more accurate target placement than same-backbone point conditioning, while the proxy branch outperforms RGB- and embedding-based injection in spatial control accuracy and point-placement precision (Tables 2 and 4). Our contributions are twofold:

- We develop an automated proxy-data pipeline that combines mesh-based correspondence matching, rule-based mapping filters, and multimodal pair review. It produces quality-filtered proxy point mappings and contrastive try-on pairs for learning wearing-state changes while preserving shared try-on conditions.
- We introduce a proxy representation branch that explicitly routes garment features through Fold–Attention–ReFold operations. Coordinate-based feature extraction and placement provide precise part-level control, supported by matched-noise paired training and compatible with ordinary VTON through mapping dropout.

Experiments show that ProxyFit improves spatial control accuracy and point-placement precision over same-backbone point guidance, while retaining strong standard try-on quality on the standard VTON task.

2 Related Work

Diffusion-Based Virtual Try-On. Diffusion models (Ho et al., 2020) provide the generative foundation for recent VTON systems. Latent diffusion and SDXL improve the efficiency and fidelity of high-resolution synthesis (Rombach et al., 2022; Podell et al., 2024), while DiT and Stable Diffusion 3 extend scalable image generation through Transformer backbones (Peebles & Xie, 2023; Esser et al., 2024). Within VTON, LaDI-VTON and DCI-VTON combine garment alignment with diffusion-based synthesis (Morelli et al., 2023; Gou et al., 2023), whereas StableVITON learns semantic correspondence through latent cross-attention (Kim et al., 2024). TryOnDiffusion unifies garment



Figure 2: Proxy point mappings represent wearing-state changes. Selected mesh-derived source points identify garment parts; their person-side targets determine the intended placement in each state. Matching colors denote the same part, and arrows illustrate joint edits to the front, sleeves, and hem.

and person processing in a Parallel-UNet (Zhu et al., 2023); IDM-VTON, OOTDiffusion, and OutfitAnyone further develop reference-feature transfer with dual-stream diffusion architectures (Choi et al., 2024; Xu et al., 2025; Sun et al., 2024). CatVTON simplifies conditioning through spatial concatenation (Chong et al., 2025), and FitDiT emphasizes garment-detail preservation with a diffusion Transformer (Jiang et al., 2024). MMTryon and AnyFit extend conditioning to multiple garment references and outfit composition (Zhang et al., 2024; Li et al., 2024). These developments improve appearance transfer, efficiency, and garment coverage; ProxyFit focuses on the complementary problem of explicitly specifying where individual garment parts should appear.

Wearing-Style and Spatial Control. Text-conditioned VTON systems express wearing styles through semantic instructions: MMTryon supports multimodal outfit control, while PromptDresser uses generated descriptions and prompt-aware masks (Zhang et al., 2024; Kim et al., 2025). General editors such as Qwen-Image-Edit accept broader free-form requests, but language alone leaves exact part locations unspecified. Spatial conditioning methods offer more explicit guidance: ControlNet and T2I-Adapter inject structural conditions through additional branches (Zhang et al., 2023; Mou et al., 2024), while GLIGEN introduces gated layers for grounded generation (Li et al., 2023). For direct manipulation, DragGAN and DragDiffusion move selected image points toward user-specified targets (Pan et al., 2023; Shi et al., 2024). In VTON, MOFA-VTON converts a curve into upper/lower regions for layout adjustment (Han et al., 2026); its two-region design is specialized to the garment boundary. Wear-Any-Way provides a more general garment–person point interface through sparse correspondence alignment (Chen et al., 2024). ProxyFit builds on the flexibility of point interaction with an explicit source-read and target-write pathway for proxy point mappings.

Correspondence and Feature Routing. Self-supervised visual representations, including DINO and DINOv2, encode transferable semantic information (Caron et al., 2021; Oquab et al., 2024); DINOv3 further strengthens dense feature quality at scale (Siméoni et al., 2025). Such features support part-level matching, as demonstrated by dense ViT descriptors, while diffusion features also exhibit emergent correspondence (Amir et al., 2022; Tang et al., 2023). Complementary work controls generation by manipulating attention: Prompt-to-Prompt edits cross-attention maps, MasaCtrl retrieves source-image content through mutual self-attention, and IP-Adapter adds decoupled attention for image prompts (Hertz et al., 2022; Cao et al., 2023; Ye et al., 2023). ProxyImg explores editable image representations through disentangled proxy embeddings (Chen et al., 2026). ProxyFit uses garment–person proxy point mappings to jointly define which local garment features are available for retrieval and where the retrieved features are written back, connecting correspondence-based supervision to explicit spatial control.

3 METHOD

ProxyFit combines an automated proxy-point data pipeline with an explicit proxy representation branch (Figure 3). The pipeline constructs quality-filtered proxy point mappings and contrastive

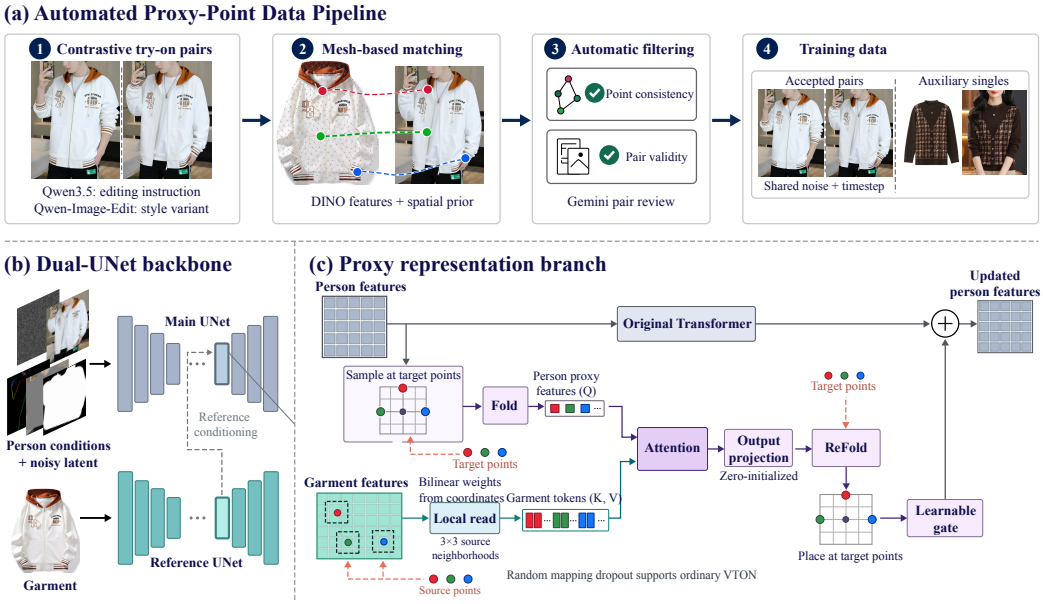


Figure 3: ProxyFit framework. (a) The automated pipeline generates contrastive try-on pairs, matches mesh-derived points, and filters mappings with rules and pairs with Gemini 3.5 Flash. (b) A dual-UNet backbone transfers garment appearance. (c) The proxy branch folds person features into point queries, reads garment-local features, and refolds the projected residual at the target points through a learnable gate.

try-on pairs. The branch executes these mappings through Fold–Attention–ReFold feature routing, trained with matched-noise paired supervision and ordinary single-example denoising.

3.1 Problem Formulation and Proxy Representation

Let I_g be a flat garment image, and let C_p collect the standard person conditions: the person representation, pose, agnostic image, and inpainting mask. A *proxy point mapping* links a source point $\mathbf{c}_i$ on the garment to a target point $\mathbf{p}_i$ on the person. The source identifies *which garment part* to transfer; the target specifies *where it should appear*. Together, these mappings express the intended wearing state without predefined style labels:

$$\mathcal{P} = \{(\mathbf{c}_i, \mathbf{p}_i, m_i)\}_{i=1}^{K}, \qquad \hat{I} = G(I_g, C_p, \mathcal{P}). \tag{1}$$

Here, $\mathbf{c}_i, \mathbf{p}_i \in [0,1]^2$ are normalized coordinates and $m_i \in \{0,1\}$ indicates validity; invalid entries pad the set to a maximum of K mappings. Keeping source points fixed while dragging target points precisely expresses the desired part placement, including simultaneous edits to several regions (Figure 2).

3.2 Automated Proxy-Point Data Pipeline

Contrastive try-on pairs. For each source garment–person example, Qwen3.5 (Qwen Team, 2026) identifies the current front-opening, sleeve, hem, and fit states, selects an applicable alternative, and generates an editing instruction. Qwen-Image-Edit (Wu et al., 2025) is instructed to change the wearing style while preserving person identity, pose, garment appearance, and background. The original and edited targets form a candidate *contrastive try-on pair*. Both share the garment, pose, agnostic image, and garment-agnostic mask. The mask is dilated to remove original clothing cues and cover the new wearing state, without using the edited target image at inference.

Mesh-based proxy-point matching. We extract the flat and worn garment regions using human parsing (Jin, 2023), triangulate their contours and interiors, and use mesh vertices as candidate proxy points following ProxyImg (Chen et al., 2026). Initial DINOv3 feature matches (Siméoni et al., 2025)

estimate an affine transformation that coarsely aligns the flat garment with its worn counterpart. We then combine feature similarity with this spatial prior to select a visually similar, spatially plausible person-side match for each source vertex. Repeating the process for both targets maps the same garment parts to their locations in the two wearing states.

Automatic filtering of mappings and pairs. Filtering checks both correspondence reliability and image-pair validity. For a mapping to be retained, the garment point must select the person point as its best feature match, and the reverse search must recover the garment point among its top-3 candidates; cosine similarity must exceed 0.5. We remove duplicate locations and retain at most $K = 128$ mappings. Their order is shuffled during training so that tensor slots do not acquire fixed point identities. Gemini 3.5 Flash (Google, 2026) then reviews human-structure realism, garment identity, style-change success, and proxy-point accuracy. This review filters out Qwen-Image-Edit samples with instruction-following failures or undesired changes to content outside the intended edit. A pair is retained only when both targets pass. If only one passes, it becomes an *auxiliary single example*: a valid standalone garment–person sample used for ordinary denoising training rather than paired supervision.

3.3 Proxy Representation Branch

The dual-UNet backbone (Hu et al., 2023; Li et al., 2024), detailed in Appendix A.1, retains its original reference-conditioning path. An independent proxy residual branch augments the Main UNet's spatial Transformer blocks, consuming coordinates directly to establish part-wise garment–person relations (input fields are listed in Table 6 of Appendix A.2). We randomly drop all proxy point mappings for a subset of training samples, allowing the same model to support ordinary VTON when no points are provided.

Fold: extracting person proxy features. Fold extracts a feature at each requested body location. Let $F_p \in \mathbb{R}^{B\times C\times H_p\times W_p}$ and $F_g \in \mathbb{R}^{B\times C\times H_g\times W_g}$ denote matched person and garment feature maps. Bilinear sampling (Jaderberg et al., 2015) produces the person proxy feature

$$\mathbf{f}_i^p = \sum_{\mathbf{u}\in\mathcal{N}(\mathbf{p}_i)} \omega_{i\mathbf{u}} F_p[\mathbf{u}], \qquad \sum_{\mathbf{u}} \omega_{i\mathbf{u}} = 1. \tag{2}$$

$\mathcal{N}(\mathbf{p}_i)$ contains the four neighboring grid locations. The bilinear weights $\omega_{i\mathbf{u}}$ are computed from coordinates, not learned. These features supply the person-point queries.

Attention: retrieving paired garment features. Each person point $\mathbf{p}_i$ queries only the 3×3 garment window $\mathcal{G}_i$ at its paired source $\mathbf{c}_i$. Layer normalization and learned projections map $\mathbf{f}_i^p$ to query Q_i, and $\mathcal{G}_i$ to keys K_i and values V_i. For each valid mapping, each head computes (Vaswani et al., 2017):

$$Z_i = \mathrm{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_h}}\right) V_i, \tag{3}$$

where d_h is the head dimension and softmax runs over the nine tokens of $\mathcal{G}_i$. The shared index i enforces source–target pairing; no attention crosses between pairs. Invalid mappings contribute zero residual.

ReFold: placing features at target points. ReFold projects each retrieved feature Z_i and distributes it to the target's four neighboring locations using the same coordinate-derived weights as Fold. A learnable gate scales the spatial residual:

$$\Delta F_p[\mathbf{u}] = \sum_{i:m_i=1,\ \mathbf{u}\in\mathcal{N}(\mathbf{p}_i)} \omega_{i\mathbf{u}} W_O Z_i, \qquad F_p' = F_p + \tanh(g)\Delta F_p. \tag{4}$$

Each layer has independent projections and a gate. We initialize the output projection W_O and its bias to zero, and sample g from a zero-mean Gaussian with standard deviation 10^{-2}. The branch therefore leaves the pretrained backbone unchanged initially, then learns to write garment information at the specified body locations.

3.4 Training and Inference

Let $C = (I_g, C_p)$ denote the shared try-on conditions. The pipeline yields two data formats:

$$\begin{aligned} d_{\text{pair}} &= \big(C, (I^a, \mathcal{P}^a), (I^b, \mathcal{P}^b)\big), \\ d_{\text{single}} &= \big(C, I, \mathcal{P}\big). \end{aligned} \tag{5}$$

A contrastive pair contains two wearing states and their mappings under the same garment and person conditions; a single example contains one valid target and its mappings. We interleave pair batches and single-example batches. Within each pair, the two targets share noise ϵ and timestep t, reducing stochastic differences unrelated to the edited mappings. Their standard diffusion denoising losses (Ho et al., 2020; Rombach et al., 2022) are averaged:

$$\mathcal{L}_{\text{pair}} = \mathbb{E}_{d_{\text{pair}},t,\epsilon}\left[\frac{1}{2}\sum_{s\in\{a,b\}} \|f_\theta(x_t^s, t, C, \mathcal{P}^s) - \epsilon\|_2^2\right]. \tag{6}$$

Here, x_t^s is the noisy latent of I^s. Auxiliary single examples use the ordinary loss

$$\mathcal{L}_{\text{single}} = \mathbb{E}_{d_{\text{single}},t,\epsilon}\left[\|f_\theta(x_t, t, C, \mathcal{P}) - \epsilon\|_2^2\right]. \tag{7}$$

The two batch types combine supervision for mapping-driven edits with broader standalone-example coverage; *contrastive* describes the paired data, not an additional loss.

Inference-time control. Users keep garment-side points fixed and drag person-side targets to precisely express the intended edit. The Reference UNet runs once, and its features are reused throughout denoising. Ordinary VTON is supported by training-time mapping dropout.

4 Experiments

4.1 Experimental Setup

Datasets. Our pipeline retains 17,704 contrastive try-on pairs, split into 16,504 training pairs, 200 validation pairs, and 1,000 held-out test pairs. Training also uses 27,075 auxiliary single examples. All variants of an original sample stay in one split; validation/test source samples are excluded from both training stages and auxiliary singles. The control test set contains 308 front-opening, 673 sleeve, 673 hem, and 112 fit changes; these subsets overlap because a pair may change multiple attributes. Each target is evaluated independently, yielding 2,000 test images. For standard VTON, we use the 2,500-pair test split of the *proprietary e-commerce dataset*. The dataset contains mainly upper-body person images and upper-garment references, with complex patterns, backgrounds, poses, and layered outfits. Its original split comprises 50,602 training pairs and 2,500 testing pairs; our quality comparison uses the latter, separately from the control test set.

Implementation details. ProxyFit uses an SDXL-based dual-UNet backbone (Podell et al., 2024; Hu et al., 2023; Li et al., 2024). The proxy representation branch is inserted into 16 spatial Transformer blocks, with 3×3 garment neighborhoods and at most $K = 128$ mappings. Person and garment inputs have resolutions of 1024×768 and 1024×1024, respectively. We pretrain the backbone for standard VTON for 100 epochs using only real (non-Qwen-generated) samples from the second-stage training split, then add the proxy branch and train on proxy-point data for another 100 epochs. Training uses 8 NVIDIA H20 GPUs, AdamW (Loshchilov & Hutter, 2019), a learning rate of 5×10^{-5}, batch size 8, and DeepSpeed ZeRO-2 (Rajbhandari et al., 2020). Inference uses 30 DDIM steps (Song et al., 2021) and classifier-free guidance of 1.2 (Ho & Salimans, 2022).

Baselines. For control, we compare with the released Qwen-Image-Edit-2509 checkpoint (Wu et al., 2025), using attribute-matched text instructions. Official code for Wear-Any-Way (Chen et al., 2024) and MOFA-VTON (Han et al., 2026) was not publicly available at the time of our experiments. We therefore reimplement their respective point-attention and dual-region curve-control mechanisms within our backbone, yielding the *WAW-style* and *MOFA-style* baselines. These variants share ProxyFit's training data and optimization budget; MOFA-style is evaluated only for its supported hem control. Standard VTON comparisons include LaDI-VTON (Morelli et al., 2023), DCI-VTON (Gou et al., 2023), StableVITON (Kim et al., 2024), IDM-VTON (Choi et al., 2024), and CatVTON (Chong et al., 2025) flux version.

Metrics. We use Gemini 3.5 Flash (Google, 2026) to measure spatial control accuracy. Given the flat garment, target reference person image, and generated image, it judges whether the corresponding regions match spatially in four aspects: *front opening*, *sleeve placement*, *hem–waist relationship*, and *garment silhouette*. We report single-image accuracy on each relevant control subset and their unweighted mean, mAcc. This protocol achieves 87% agreement with human judgments. We also

Table 1: Cumulative filtering yield and spatial control accuracy of retained pairs, assessed by Gemini 3.5 Flash.

Filtering stage	Retained pairs	Retention	Spatial control accuracy
Raw DINO matches	60,695	100.00%	63.73%
+ Human-structure realism	55,183	90.92%	61.28%
+ Garment identity preservation	46,287	76.26%	66.46%
+ Proxy-point accuracy (final)	17,704	29.17%	91.42%

Table 2: Spatial control on 1,000 test pairs. The four accuracies are evaluated per image (0–1); mAcc is their unweighted mean. d_{Euc} is point-placement error in pixels.

Method	Open ↑	Sleeve ↑	Hem ↑	Fit ↑	mAcc ↑	d_{Euc} ↓
Qwen-Image-Edit	0.84	**0.83**	0.88	0.71	0.815	59.08
WAW-style	0.75	0.69	0.83	0.66	0.733	51.40
MOFA-style	–	–	0.74	–	–	–
ProxyFit	**0.85**	0.79	**0.90**	**0.77**	**0.828**	**40.13**

report d_{Euc}, the Euclidean distance in pixels between requested targets and their DINO-matched generated locations on a 1024×768 canvas, averaged first over the matches per image and then over evaluable images.

Following LaDI-VTON (Morelli et al., 2023), we evaluate standard VTON with four metrics. LPIPS (Zhang et al., 2018) and SSIM (Wang et al., 2004) measure perceptual and structural similarity to reference images in the paired setting. FID (Heusel et al., 2017) and KID (Bińkowski et al., 2018) measure distributional similarity between generated and real images in the unpaired setting. KID values are multiplied by 10^3. Lower LPIPS/FID/KID and higher SSIM indicate better quality.

4.2 Data Filtering and Yield

Table 1 summarizes the yield of the automated pipeline. Starting from 60,695 candidate pairs with DINO matches, the human-structure and garment-identity checks retain 55,183 and 46,287 pairs, respectively. The final proxy-point check leaves 17,704 pairs, or 29.17% of the initial pool, with 91.42% spatial control accuracy assessed by Gemini 3.5 Flash. These retained pairs provide supervision in which the wearing-state change is accompanied by reliable garment–person point mappings.

4.3 Main Results

Spatial control accuracy. ProxyFit achieves the highest mAcc of 0.828 in Table 2, improving over WAW-style by 9.5 percentage points and Qwen-Image-Edit by 1.3 points. It leads on front opening, hem placement, and fit, while Qwen performs best on sleeves. Despite its competitive mAcc, Qwen-Image-Edit has the largest reported d_{Euc} (59.08 pixels), indicating that its text-based control provides limited precision in placing garment parts at the requested target points. Compared with WAW-style under the same backbone and training budget, ProxyFit improves all four accuracies and reduces d_{Euc} from 51.40 to 40.13 pixels (21.9%), supporting the benefit of explicit feature routing for executing proxy point mappings.

Following an instruction versus reaching a target. Figure 4 shows that Qwen-Image-Edit often follows the requested semantic change, but can also alter the person's pose or unedited clothing. In the bottom-right black-hoodie example, Qwen's hem height deviates from the requested target, and the arm pose also changes. WAW-style better preserves the original content but under-executes several point edits. ProxyFit follows the requested sleeve, front, and hem displacements more closely, including joint edits. Its 32.1% lower d_{Euc} than Qwen complements their relatively close mAcc scores: successful style changes do not by themselves imply precise point placement.

Standard VTON quality. On the proprietary test set, ProxyFit achieves the lowest LPIPS (0.201), FID (44.35), and scaled KID (0.607) among the methods in Table 3. Its SSIM of 0.721 remains



Figure 4: Wearing-style control with ProxyFit, WAW-style, and Qwen-Image-Edit. Source images show requested point movements; Qwen receives corresponding text instructions. ProxyFit more closely follows the intended local displacements, while Qwen can change pose or unrelated content and WAW-style can leave requested edits incomplete.

Table 3: Standard VTON quality on 2,500 real-world e-commerce test pairs.

Method	LPIPS ↓	SSIM ↑	FID ↓	KID ↓ ($\times 10^3$)
LaDI-VTON	0.252	**0.734**	52.24	6.51
DCI-VTON	0.264	**0.734**	57.96	12.35
StableVITON	0.242	0.720	53.80	8.13
IDM-VTON	0.247	0.701	48.76	4.35
CatVTON	0.291	0.678	47.46	2.22
ProxyFit	**0.201**	0.721	**44.35**	**0.607**

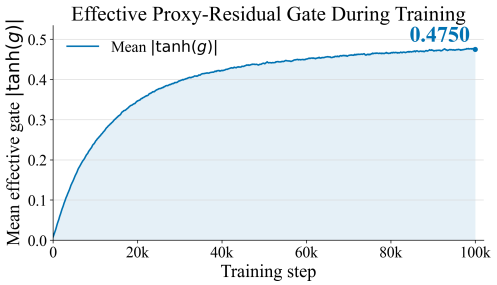


Figure 5: Mean effective proxy gate magnitude $|\tanh(g)|$ across layers during training.

below the joint best of 0.734. These results show that precise control is compatible with strong try-on texture quality. Figure 6 further illustrates preservation of the jacket's color-block layout, the sweater's repeated motifs, and the sweatshirt's printed design across varied people and backgrounds.

4.4 Ablation Study

Proxy-point injection. We fix the backbone, training data, and optimization budget and vary only the injection mechanism. RGB ControlNet (Zhang et al., 2023) rasterizes mappings into images; Embedding ControlNet encodes point pairs through a side branch. As Table 4 shows, both improve over no proxy, but ProxyFit's explicit Fold–Attention–ReFold routing raises mAcc further to 0.828, versus 0.722 and 0.740. Relative to embedding injection, it gains 8.8 percentage points and reduces point error by 23.5%. Thus, supplying point conditions alone does not provide the same spatial precision as reading and writing features at the specified locations.

Paired training. With the architecture fixed, paired sampling with independent noise and timesteps improves mAcc from 0.741 to 0.784 and reduces d_{Euc} from 50.98 to 43.87 pixels (Table 5). Sharing noise and timesteps adds another 4.4 percentage points of mAcc and lowers the error to 40.13 pixels. Pairing and matched corruption therefore make complementary contributions to learning the response to edited proxy point mappings, without introducing a contrastive loss.



Figure 6: Comparison on the real-world VTON data. Each input pairs a garment with a target person. ProxyFit retains garment structures and complex patterns across varied poses and backgrounds.

Table 4: Proxy-point injection ablation.

Variant	**mAcc** ↑	$d_{\mathbf{Euc}}$ ↓
No proxy	0.645	63.08
RGB ControlNet	0.722	54.11
Embed. ControlNet	0.740	52.43
ProxyFit	**0.828**	**40.13**

Table 5: Paired-training ablation.

Sampling	**mAcc** ↑	$d_{\mathbf{Euc}}$ ↓
Random	0.741	50.98
Paired, indep. ϵ, t	0.784	43.87
Paired, shared ϵ, t	**0.828**	**40.13**

Gate dynamics. Figure 5 shows the mean effective gate magnitude, $|\tanh(g)|$, rising rapidly early in training and gradually approaching approximately 0.475 at 100k steps. Together with zero-initialized output projections, this gradual activation lets the proxy branch enter the pretrained feature stream progressively. The increased gate magnitude gives greater weight to proxy-guided residuals, enabling the branch to participate more actively in steering garment placement during try-on generation. The control ablations, rather than gate magnitude alone, establish its contribution to spatial accuracy.

5 Conclusion

We presented **ProxyFit** for precise, interactive wearing-style control through proxy point mappings. Its automated pipeline constructs quality-filtered correspondences and contrastive try-on pairs, and its proxy branch explicitly reads garment features and places them at requested body locations through Fold–Attention–ReFold. Experiments show that explicit routing and matched-noise paired training improve spatial control, while preserving strong standard VTON quality. By connecting user intent with explicit spatial control, ProxyFit paves the way for personalized virtual try-on systems that let users decide not only what to wear, but precisely how it is worn.

AI USE STATEMENT

We used generative AI tools to assist with language polishing, literature and reference checking, and code implementation for ablation studies and the WAW-style and MOFA-style baselines. The AI-assisted code was reviewed and tested for correctness by one author. We also used Qwen3.5 and Qwen-Image-Edit to generate editing instructions and synthetic contrastive try-on pairs, and Gemini 3.5 Flash to assist with data filtering and spatial control evaluation. We assessed the quality of the generated data through human annotation and automated evaluation, and checked the automated evaluator against human judgments. We reviewed and revised the AI-assisted manuscript content and take responsibility for the final text, claims, code, data, and other artifacts in this work.

REFERENCES

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep ViT features as dense visual descriptors. In *European Conference on Computer Vision Workshops*, 2022. URL https://arxiv.org/abs/2112.05814.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *Proc. ICLR*, 2018. URL https://openreview.net/forum?id=r1lUOzWCW.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *arXiv preprint arXiv:2304.08465*, 2023. URL https://arxiv.org/abs/2304.08465.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://arxiv.org/abs/2104.14294.

Mengting Chen, Xi Chen, Zhonghua Zhai, Chen Ju, Xuewen Hong, Jinsong Lan, and Shuai Xiao. Wear-any-way: Manipulable virtual try-on via sparse correspondence alignment. In *European Conference on Computer Vision*, pp. 124–142. Springer, 2024. doi: 10.1007/978-3-031-72630-9_8.

Ye Chen, Yupeng Zhu, Xiongzhen Zhang, Zhewen Wan, Yingzhe Li, Wenjun Zhang, and Bingbing Ni. Proxyimg: Towards highly-controllable image representation via hierarchical disentangled proxy embedding. *arXiv preprint arXiv:2602.01881*, 2026. URL https://arxiv.org/abs/2602.01881.

Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. VITON-HD: High-resolution virtual try-on via misalignment-aware normalization. In *Proc. IEEE/CVF CVPR*, 2021. URL https://arxiv.org/abs/2103.16874.

Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *European Conference on Computer Vision*, pp. 206–235. Springer, 2024. doi: 10.1007/978-3-031-73016-0_13.

Zheng Chong, Xiao Dong, Haoxiang Li, Shiyue Zhang, Wenqing Zhang, Xujie Zhang, Hanqing Zhao, Dongmei Jiang, and Xiaodan Liang. CatVTON: Concatenation is all you need for virtual try-on with diffusion models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a6a1e4c756d700d9aedcc1896a7e6fb0-Abstract-Conference.html.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proc. ICML*, 2024. URL https://arxiv.org/abs/2403.03206.

Google. Gemini 3.5 Flash. Gemini API documentation, 2026. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.5-flash.

Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. 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*, pp. 7599–7607, 2023. doi: 10.1145/3581783.3612255.

Xiaoyu Han, Chenyang Wang, Jing Wang, Shunyuan Zheng, Quanling Meng, and Shengping Zhang. MOFA-VTON: More fashion possibilities with fine-grained adaptations in virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1895–1905, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Han_MOFA-VTON_More_Fashion_Possibilities_with_Fine-Grained_Adaptations_in_Virtual_Try-On_CVPR_2026_paper.html.

Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. VITON: An image-based virtual try-on network. In *Proc. IEEE CVPR*, 2018. URL https://arxiv.org/abs/1711.08447.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. URL https://arxiv.org/abs/2208.01626.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.08500.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. URL https://arxiv.org/abs/2006.11239.

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. URL https://arxiv.org/abs/2311.17117.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://arxiv.org/abs/1506.02025.

Boyuan Jiang, Xiaobin Hu, Donghao Luo, Qingdong He, Chengming Xu, Jinlong Peng, Jiangning Zhang, Chengjie Wang, Yunsheng Wu, and Yanwei Fu. FitDiT: Advancing the authentic garment details for high-fidelity virtual try-on. *arXiv preprint arXiv:2411.10499*, 2024. URL https://arxiv.org/abs/2411.10499.

Zhenchao Jin. Sssegmenation: An open source supervised semantic segmentation toolbox based on pytorch. *arXiv preprint arXiv:2305.17091*, 2023. URL https://arxiv.org/abs/2305.17091.

Jeongho Kim, Gyojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. 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*, pp. 8176–8185, 2024. URL https://arxiv.org/abs/2312.01725.

Jeongho Kim, Hoiyeong Jin, Sunghyun Park, and Jaegul Choo. PromptDresser: Improving the quality and controllability of virtual try-on via generative textual prompt and prompt-aware mask. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16026–16036, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Kim_PromptDresser_Improving_the_Quality_and_Controllability_of_Virtual_Try-On_via_ICCV_2025_paper.html.

Yuhan Li, Hao Zhou, Wenxiang Shang, Ran Lin, Xuanhong Chen, and Bingbing Ni. Anyfit: Controllable virtual try-on for any combination of attire across any scenario. *arXiv preprint arXiv:2405.18172*, 2024. URL https://arxiv.org/abs/2405.18172.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: Open-set grounded text-to-image generation. In *Proc. IEEE/CVF CVPR*, 2023. URL https://arxiv.org/abs/2301.07093.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022W/CVFAD/html/Morelli_Dress_Code_High-Resolution_Multi-Category_Virtual_Try-On_CVPRW_2022_paper.html.

Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On. In *Proc. ACM MM*, 2023. URL https://arxiv.org/abs/2305.13501.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proc. AAAI*, 2024. URL https://ojs.aaai.org/index.php/AAAI/article/view/28226.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2304.07193.

Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your GAN: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH Conference Proceedings*, 2023. URL https://vcai.mpi-inf.mpg.de/projects/DragGAN/.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proc. IEEE/CVF ICCV*, 2023. URL https://arxiv.org/abs/2212.09748.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/081b08068e4733ae3e7ad019fe8d172f-Abstract-Conference.html.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2020. doi: 10.1109/SC41405.2020.00024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF CVPR*, 2022. URL https://arxiv.org/abs/2112.10752.

Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent Y. F. Tan, and Song Bai. DragDiffusion: Harnessing diffusion models for interactive point-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2306.14435.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Ke Sun, Jian Cao, Qi Wang, Linrui Tian, Xindi Zhang, Lian Zhuo, Bang Zhang, Liefeng Bo, Wenbo Zhou, Weiming Zhang, et al. OutfitAnyone: Ultra-high quality virtual try-on for any clothing and any person. *arXiv preprint arXiv:2407.16224*, 2024. URL https://arxiv.org/abs/2407.16224.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0503f5dce343a1d06d16ba103dd52db1-Abstract-Conference.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.03762.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004. doi: 10.1109/TIP.2003.819861.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025. URL https://arxiv.org/abs/2508.02324.

Yuhao Xu, Tao Gu, Weifeng Chen, and Arlene Chen. OOTDiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8996–9004, 2025. doi: 10.1609/aaai.v39i9.32973.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. URL https://arxiv.org/abs/2308.06721.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proc. IEEE/CVF ICCV*, 2023. URL https://arxiv.org/abs/2302.05543.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 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. URL https://arxiv.org/abs/1801.03924.

Xujie Zhang, Ente Lin, Michael Kampffmeyer, Zhenyu Xie, Jiang Li, Ting Liu, Xiaochao Qu, Luoqi Liu, and Xiaodan Liang. MMTryon: Multi-modal multi-reference control for high-quality fashion generation. *arXiv preprint arXiv:2405.00448*, 2024. URL https://arxiv.org/abs/2405.00448.

Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *Proc. IEEE/CVF CVPR*, 2023. URL https://arxiv.org/abs/2306.08276.