REALFIT: MEASUREMENT-CONDITIONED VIRTUAL TRY-ON

Anonymous authors
Paper under double-blind review



Figure 1: **Measurement-conditioned virtual try-on with RealFIT.** Given an input person of height **H** and a layflat garment (*left*), RealFIT synthesizes photorealistic try-on results conditioned on physical garment measurements in centimeters (*right*; **L**: length, **B**: bust, **W**: waist, **S**: sleeve length), smoothly transitioning from tight to oversized fit with authentic fabric drape.

ABSTRACT

The core value of real-world try-on lies in allowing consumers to both a) visualize how a garment looks on themselves, as well as b) see how a garment will fit on them: whether any given size will stretch tightly over body contours, or drape loosely with excess slack. Prevailing Virtual Try-On (VTO) models address the first use-case, but ignore the second: by conditioning solely on 2D visual inputs, they default to synthesizing deceptively well-fitted try-on results regardless of garment dimensions. Data scarcity is a central bottleneck for developing fit-accurate VTO systems. Standard public benchmarks consist of catalog imagery lacking both garment dimensions and body measurements. While recent synthetic approaches leverage 3D physics simulation to generate exact measurement labels, they suffer from sim-to-real domain gaps and remain constrained to simplistic garment topologies.

In this work, we address the issue of fit through **RealFIT**, the first framework for scalable, real-world measurement-conditioned virtual try-on across diverse and structurally complex apparel categories. To bridge the data gap, we introduce a multi-tier data hierarchy that bootstraps geometric priors from large-scale web retail metadata and VLM-assisted measurement annotations, anchored by a newly curated real-world dataset with tape-measured metric ground truth. Using this data, we propose a fit-aware virtual try-on model based on flow matching that integrates a dedicated measurement encoder to condition generation on human and garment dimensions. To enhance the model's sensitivity to metric conditions, we also introduce a **Localized Metric Contrastive (LMC)** loss, which explicitly boosts the conditioning signal by contrasting velocity fields across paired sizing perturbations within targeted regions. Furthermore, we establish a systematic evaluation protocol to benchmark fit accuracy. Extensive experiments demonstrate that RealFIT achieves state-of-the-art photorealism and fit accuracy and controllability across varied body geometries and complex garment styles. We plan to release our benchmark dataset, evaluation suite, and code upon acceptance.

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.

1 INTRODUCTION

The field of Image-based Virtual Try-On aims to bring the classic consumer fitting-room experience online. While recent diffusion-based methods (Zhu et al., 2023; Chong et al., 2025; Choi et al., 2024) achieve remarkable photorealism in garment synthesis, they remain agnostic to physical dimensions and show deceptively well-fitted garments. This hides key sizing mismatches, such as tight cloth or excess fabric drape, that consumers rely on try-on to reveal.

The core bottleneck to developing fit-aware systems is the training data. Standard real-world VTO benchmarks (Liu et al., 2016; Choi et al., 2021; Morelli et al., 2022; Cui et al., 2025) are sourced from retail catalogs where garments are pinned to fit each model, thus lacking size annotations and "ill-fitting" examples (too loose or tight). To fill this gap, FIT (Wang et al., 2026) uses 3D simulation to drape garments with exact measurements over parametric bodies. However, synthetic simulation introduces its own downsides: synthetic artifacts caused by the domain gap, and the inability to generate complex apparel with seams, pockets, or structural outerwear.

We therefore present **RealFIT**, a fit-aware try-on framework that is grounded in real photography in order to preserve authentic fabric dynamics and allow use on complex garment patterns. To overcome the lack of measurements in real-world images, we assemble **RealFIT-Web**, a large-scale dataset with measurement labels collected from: annotations by vision-language models, as well as retailers' published size charts, to learn a broad prior relating dimensions to fit. We then finetune and evaluate on a tape-measured benchmark of real subjects (**RealFIT-Bench**). Because real photos rarely contain ill-fit examples, we augment our training data with a frontier image generation model, re-rendering garments at altered sizes under automated QA to teach the model how appearance changes with size.

Using this data, we propose a fit-aware virtual try-on architecture built upon a state-of-the-art flow matching backbone (Black Forest Labs, 2025). Since traditional clothing-agnostic masks leak the target garment's size, we adopt a mask-free formulation that conditions directly on a synthetic reference person image, and utilize a dedicated measurement encoder to handle body and garment measurements. However, because sizing shifts alter only localized regions (e.g., hemlines or sleeves), global flow-matching loss dominates and dilutes the measurement signal. We resolve this by introducing a **Localized Metric Contrastive** (LMC) loss that amplifies the sizing signal at active garment boundaries. Given paired size-perturbed training examples, LMC evaluates the network under target and perturbed measurements on a shared noisy latent, enforcing directional alignment between the predicted velocity difference and ground-truth deformation within the image tokens governing size change. This concentrates gradient supervision directly on shifting boundaries without disrupting static regions like facial identity or background.

Finally, because image similarity metrics commonly used in VTO evaluations provide little insight into garment size accuracy, we establish a systematic protocol for VTO fit evaluation. We project segmented garments onto keypoint-defined reference lines along the torso and limbs, and thereby introduce two complementary metrics: **Landmark Fit Score (LFS)** to measure garment coverage against a ground truth, and **Size Monotonicity (SM)** to quantify continuous, monotonic size increase across a set of results as input dimensions scale. Our main contributions are summarized as follows:

- We introduce **RealFIT-Web** and **RealFIT-Bench**, pairing a large-scale VLM-annotated pretraining corpus (345k samples) with a tape-measured benchmark of real subjects (12k samples).
- We present **RealFIT**, a measurement-conditioned virtual try-on model with Localized Metric Contrastive loss that achieves state-of-the-art visual realism and faithful physical sizing.
- We establish a landmark-anchored evaluation protocol, Landmark Fit Score (LFS) and Size Monotonicity (SM), to systematically quantify garment fit accuracy and consistency.

2 RELATED WORK

2.1 VIRTUAL TRY-ON

Image-based virtual try-on has evolved significantly with the adoption of deep generative models. Early approaches relied on explicit 2D warping pipelines and Thin-Plate Spline (TPS) transformations to align flat-lay garments with target body representations (Han et al., 2018; Choi et al., 2021; Ge et al., 2021). More recently, diffusion models have achieved tremendous progress in transferring

high-fidelity textures and garment appearances onto target subjects (Zhu et al., 2023; 2024; Chong et al., 2025; Kim et al., 2024; Zhang et al., 2024; Huang et al., 2026). In the absence of paired data, mainstream VTO pipelines train via mask-based self-reconstruction (Gou et al., 2023; Morelli et al., 2023; Baldrati et al., 2023; Choi et al., 2024; Xu et al., 2025), masking out the worn garment and regenerating it from the layflat reference. However, because the mask traces the worn garment's outline, it leaks that garment's size and forces synthesis into a fixed silhouette regardless of the target size. Recent mask-free architectures (Sun et al., 2024; Niu et al., 2024; Zhang et al., 2025; Shen et al., 2025) instead bypass explicit masks via cross-attention or spatial concatenation, avoiding this size leakage while preserving identity, pose, and background directly from unmasked context.

Fit-aware virtual try-on aims to depict how a garment of specific dimensions fits a given body. Early methods rely on geometric heuristics, warping clothing landmarks (Chen et al., 2023) or rescaling segmentation maps by body-to-clothing ratios (Kuribayashi et al., 2023); operating on intermediate representations, they degrade under complex poses. FitControler (Yang et al., 2025) redraws the body–garment layout from categorical fit labels (e.g., slim or loose), which cannot specify metric dimensions. Yamashita et al. (2024) condition on physical sizes but require the original garment's dimensions and can only resize that same garment. Progress is bottlenecked by data: standard benchmarks (Choi et al., 2021; Morelli et al., 2022; Cui et al., 2025) lack measurements and depict only well-fitted catalog sizes, and size-annotated corpora (Tiwari et al., 2020; Yamashita et al., 2024) remain small. FIT (Wang et al., 2026) sidesteps this scarcity by procedurally generating garments (Korosteleva & Sorkine-Hornung, 2023) and draping them in simulation with exact labels, but inherits a synthetic domain gap and cannot represent seams, pockets, or structured outerwear. RealFIT instead grounds supervision in real photography, combining measurements of varying precision (VLM estimates, retailer size charts, and tape measurements) with ill-fit augmentation.

2.2 Contrastive Learning in Diffusion and Flow Matching

A body of work integrates contrastive learning into diffusion and flow matching models to strengthen conditional discriminability (Stoica et al., 2025; Wang & He, 2025). While Classifier-Free Guidance (CFG) (Ho & Salimans, 2022; Armandpour et al., 2023) steers trajectories only at inference time, recent methods embed contrastive objectives directly into training: ΔFM (Stoica et al., 2025) and VeCoR (Hong et al., 2026) penalize velocity alignment with competing batch flows or off-manifold perturbations, Concept Sliders (Gandikota et al., 2024) and Diffusion-DPO (Wallace et al., 2024) contrast counterfactual or preference pairs for attribute and alignment control, and localized editing methods (Hertz et al., 2023; Nam et al., 2024) contrast spatial tokens to preserve unedited regions. Whereas prior objectives separate discrete classes, global preference pairs, or full-canvas 1D sliders, fit-aware try-on must translate continuous multi-dimensional measurements into spatially sparse boundary shifts. Our Localized Metric Contrastive (LMC) loss bridges this gap by projecting the velocity difference between paired sizing perturbations onto the ground-truth deformation strictly within high-energy boundary tokens, amplifying metric sensitivity without disturbing static regions.

3 The RealFIT Corpora

Existing real-world try-on datasets lack the body and garment measurement annotations required for measurement-conditioned synthesis. Measurement information is available at varying levels of precision, and we exploit both ends of this spectrum (Figure 2): **RealFIT-Web** ($\mathcal{D}_{\text{web}}$), a large-scale corpus with automatically inferred measurements for pretraining, and **RealFIT-Bench** ($\mathcal{D}_{\text{bench}}$), a curated dataset of real subjects with tape-measured ground truth for finetuning and evaluation. Each sample is a tuple $(I_p, I_g, \mathbf{m}, I_{\text{gt}})$ that is composed of: a synthesized *triplet* image I_p of the person in the target pose wearing a different garment, a real layflat garment image I_g, physical measurements $\mathbf{m} \in \mathbb{R}^5$, and the real target photograph I_{gt}.

3.1 RealFIT-Web: Pretraining Corpus

RealFIT-Web (Figure 2a) draws on over 170K samples from existing virtual try-on datasets, each supplying I_{gt} and I_g; the remaining components are produced by the steps below. We first obtain measurement annotations with a frontier vision language model (VLM), Gemini 3.7 Flash (Google DeepMind, 2026a). Given I_{gt}, the model begins by estimating the height of the person, then uses



(a) RealFit-Web Samples (b) RealFit-Bench Samples (c) RealFit-Bench Statistics

Figure 2: **Overview of the RealFIT corpora: (a)** RealFIT-Web (large-scale dataset) showing synthesized ill-fit augmentations and estimated measurements, **(b)** RealFIT-Bench (small-scale tape-measured benchmark) preview **(c)** RealFIT-Bench selected statistics.

that height as an anchor to estimate four garment dimensions, namely total length, bust, waist, and sleeve length, from visual proportions. One schema spans all categories, with dimensions undefined for a category or missing from a source set to -1. For a subset of garments, the retailer publishes a measurement chart alongside the product listing; we crawl these charts directly and read off the dimensions of the size worn, yielding a smaller pool of annotations that require no visual estimation.

Existing VTO datasets do not supply a triplet image I_p of the same person in the exact same pose wearing a different garment. To enable mask-free training, we therefore synthesize I_p by prompting Nano Banana 2 (Google DeepMind, 2026b) to replace the garment worn in I_{gt} with a different one while holding subject, camera, pose, and background fixed, so that I_p and I_{gt} differ only in the garment to be transferred. We explicitly instruct the editor to select a replacement of different length, so that I_p carries no cue to the size of the garment being transferred. Furthermore, models in catalog images are typically photographed in the size chosen for them, so the data naturally lacks ill-fit cases in which a garment is too tight or too loose. We therefore add an ill-fit augmentation step, again using Nano Banana 2 to re-render the garment of I_{gt} at an altered size while holding subject, pose, background, and garment design and style fixed, and re-estimate $\mathbf{m}'$ from the resulting image I'_{gt} to form paired size-augmented samples $\mathcal{D}_{\text{aug}}$. We apply strict quality assurance (QA) to every step above (see the appendix for more details). In total, we curated 345k training samples including $\mathcal{D}_{\text{aug}}$. Crucially, because Nano Banana 2 performs localized editing directly on real photographs, the resulting images remain fully photorealistic and free of synthetic-to-real domain gaps.

3.2 RealFIT-Bench: Tape-Measured Benchmark

Automatic annotation is sufficient to learn a coarse prior, but it lacks real-world grounding. Annotations inferred from web data are inherently noisy: visual estimation is approximate, and published size charts follow measurement conventions that differ from retailer to retailer. We therefore curate a benchmark of real subjects through crowdsourcing (Figure 2b–c). Each participant reports a set of self-measured body dimensions and selects a set of personal garments. For every garment, the participant tape-measures the garment, photographs it layflat, and then wears it for a series of full-body photographs in a standing pose. Every entry is manually reviewed for quality and reliability.

During processing, we replace each subject's face for identity protection, substitute a studio background, and enhance image quality. These edits are confined to appearance: body and garment geometry is preserved exactly, as verified by automatic and manual QA. We then produce the remaining components of each sample as in RealFIT-Web, deriving the layflat garment image I_g from the participant's garment photograph, and synthesizing the triplet image I_p. Despite our effort to collect natural ill-fit samples, each garment is observed in only a single size and mostly within 1–2 sizes of standard fit; we therefore apply ill-fit augmentation to supply paired same-garment size perturbations and expand fit extremes for training. Notably, we do not synthesize ill-fit images in our test set to ensure full real-world grounding in evaluation. The final benchmark comprises 168 subjects, 987 garments, and 11,646 on-body images, spanning most common categories, including T-shirts, jackets, hoodies, pants, dresses, and skirts. Because participants also wear garments across a wide range of natural fit types (tight, loose, and well-fitting), ill fit is directly observed on real bodies in addition to our size-augmented pairs. We partition the benchmark by subject and by garment into a finetuning split $\mathcal{D}^{\text{ft}}_{\text{bench}}$ and a held-out test split $\mathcal{D}^{\text{test}}_{\text{bench}}$, so that no test garment is seen during training.

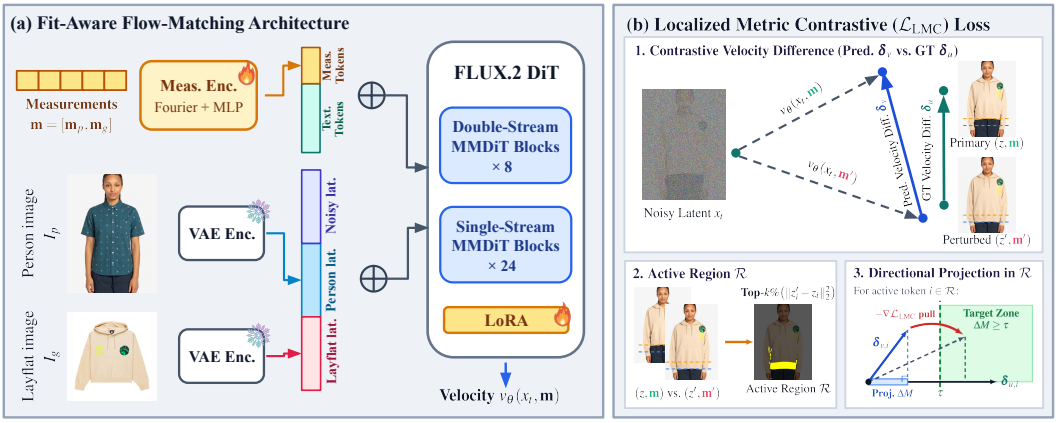

Figure 3: **RealFIT overview.** **(a)** The reference I_p and layflat garment I_g are VAE-encoded and sequence-concatenated with the noisy latent x_t in the image stream, while measurements $\mathbf{m} = [\mathbf{m}_p, \mathbf{m}_g]$ are Fourier–MLP embedded and prepended to the text tokens. **(b)** LMC evaluates v_θ twice on a shared x_t, under $\mathbf{m}$ and stop-gradient $\mathbf{m}'$, and aligns the predicted velocity difference δ_v with the ground truth $\delta_u = z' - z$ inside the top-$k\%$ active region $\mathcal{R}$.

4 Fit-Aware Virtual Try-On

4.1 Task Formulation

Conventional try-on renders a garment plausibly but leaves its size unspecified; fit-aware try-on instead conditions on physical measurements $\mathbf{m}$ (in cm) so the synthesized drape reflects the true dimensions of a garment on a specific body. We argue that the traditional mask-based reconstruction approach is ill-suited to this task since the clothing-agnostic mask leaks the garment size during training. We therefore adopt a mask-free formulation and pose try-on as measurement-conditioned image editing, $\hat{I} = \mathcal{G}(I_p, I_g, \mathbf{m})$, trained so $\hat{I}$ matches the ground-truth photograph I_{gt}. While $\mathbf{m} \in \mathbb{R}^5$ includes four garment dimensions (length, bust, waist, and sleeve length), it represents the body using only the subject's height h. This is a deliberate choice: body shape and proportions are already visible in I_p, but their absolute size is ambiguous from a 2D image; height resolves this scale ambiguity. It is also the most practical and frictionless body input for potential real-world applications, as most users know their height but rarely their body circumferences.

4.2 Model Architecture

As shown in Fig. 3, we build on FLUX.2 (Black Forest Labs, 2025). A frozen VAE $\mathcal{E}$ encodes the target image into $z = \mathcal{E}(I_{\text{gt}})$ and the reference inputs into conditioning latents $c = \big(\mathcal{E}(I_p), \mathcal{E}(I_g)\big)$. Given noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and timestep $t \in [0, 1]$, the transformer v_θ takes the noisy latent $x_t = (1-t)z + t\epsilon$ and regresses the target velocity $u = \epsilon - z$ via $\mathcal{L}_{\text{FM}} = \mathbb{E}_{z,\epsilon,t}\big[\|v_\theta(x_t, t, c, \mathbf{m}) - u\|_2^2\big]$.

To inject the visual conditioning c into v_θ, we concatenate person and garment latents with the noisy latent x_t along the sequence dimension for joint self-attention, avoiding a separate reference branch. Each token receives a 4D rotary positional embedding (T, Y, X, L) over frame, 2D spatial grid, and text index: separated by frame index ($T \in \{0, 10, 20\}$), I_p is spatially aligned to the target image to preserve pose and identity, whereas the layflat I_g occupies an unaligned frame for global conditioning.

Measurements enter through a lightweight measurement encoder. The five scalars are lifted to Fourier features, then passed through an MLP and reshaped into a set of tokens. These tokens are concatenated to the sequence of text tokens, so $\mathbf{m}$ acts as a global condition and the generative prior of the backbone is preserved. Training updates the measurement encoder and low-rank adapters inserted into the backbone; the autoencoder and the pretrained transformer weights stay frozen.

4.3 Localized Metric Contrastive (LMC) Loss

The standard flow-matching objective $\mathcal{L}_{\text{FM}}$ averages velocity errors uniformly across all spatial tokens and timesteps. Because the majority of the image regions are unaffected by garment size change,

global reconstruction dominates the loss landscape and dilutes the gradient signal driving subtle size changes. However, two observed structural properties of flow-based try-on allow us to isolate and amplify this signal: (1) *macro-structure is resolved in the high-noise regime*, where garment silhouettes are determined prior to low-noise texture refinement; and (2) *size changes are spatially localized*, altering latent representations mostly along garment boundaries (e.g., hemlines or cuffs).

Motivated by these observations, we introduce the **Localized Metric Contrastive (LMC) Loss** to explicitly boost the model's sensitivity to metric conditioning. Let $(z, \mathbf{m})$ denote a primary training latent and its ground-truth measurement vector, and let $(z', \mathbf{m}')$ denote the size-perturbed counterpart supplied by the ill-fit augmentation of Section 3, with $z' = \mathcal{E}(I'_{\text{gt}})$. Sharing ϵ across $u = \epsilon - z$ and $u' = \epsilon - z'$ cancels out the noise in the target velocity shift, $\delta_u = u - u' = z' - z$.

At large timestep t, we evaluate v_θ twice on the same noisy latent x_t, once conditioned on $\mathbf{m}$ and once on $\mathbf{m}'$, so that the predicted shift isolates the effect of the measurement change:

$$\delta_v = v_\theta(x_t, t, c, \mathbf{m}) - \text{sg}(v_\theta(x_t, t, c, \mathbf{m}')) \,, \tag{1}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. To focus gradients on active garment boundaries without disturbing static regions, we compute the per-token deformation energy $e_i = \|\delta_{u,i}\|_2^2$ and restrict supervision to the localized active region $\mathcal{R} \subset \{1, \dots, N\}$ containing the top-$k\%$ highest-energy tokens. Over $\mathcal{R}$, we project δ_v onto δ_u and normalize by the ground-truth deformation energy so the resulting directional alignment score ΔM is invariant to perturbation magnitude:

$$\Delta M = \frac{\sum_{i \in \mathcal{R}} \langle \delta_{v,i},\, \delta_{u,i} \rangle}{\sum_{i \in \mathcal{R}} \|\delta_{u,i}\|_2^2 + \epsilon_{\text{floor}}}. \tag{2}$$

We then penalize insufficient directional response using a smooth softplus margin loss ($\tau, \beta > 0$) that saturates once $\Delta M \geq \tau$ to prevent over-repulsion, warmed up linearly via $\lambda(s)$ over training steps s:

$$\mathcal{L}_{\text{LMC}} = \frac{1}{\beta} \log\left(1 + \exp\left(\beta \left(1 - \frac{\Delta M}{\tau}\right)\right)\right), \qquad \mathcal{L}_{\text{total}}(s) = \mathcal{L}_{\text{FM}} + \lambda(s)\, \mathcal{L}_{\text{LMC}}. \tag{3}$$

4.4 Fit Evaluation Protocol

Standard image similarity metrics (e.g., FID, SSIM, LPIPS) assess global appearance but do not distinguish garment sizes. Since fit perception is anchored to skeletal landmarks (e.g., hemlines relative to knees), we evaluate sizing on an anatomical coordinate frame. Given a try-on output I, we segment the garment mask $M(I)$ with SAM 3 (Carion et al., 2025) and extract 2D body keypoints with MMPose (MMPose Contributors, 2020) to define reference lines $\{L_k\}$ along the torso, arms, and legs (Fig. 4). Along each line L_k of length ℓ_k, we measure the normalized garment coverage $c_k(I) = \|L_k \cap M(I)\|/\ell_k$ and weight them by category weights w_k ($\sum_k w_k = 1$).



Figure 4: LFS computes garment mask coverage along reference lines defined by skeletal landmarks, allowing us to measure localized fit changes.

To evaluate pointwise fit accuracy against the ground-truth image I_{gt}, we define the **Landmark Fit Score (LFS)** from the weighted coverage discrepancy, $\text{LFS}(\hat{I}, I_{\text{gt}}) = 1 - \sum_k w_k |c_k(\hat{I}) - c_k(I_{\text{gt}})|$. However, pointwise accuracy on a single target size is insufficient on its own: a model can score well on typical fits simply by defaulting to an average fit while ignoring $\mathbf{m}$. A genuine fit-aware model must also respond consistently as garment dimensions are varied from smaller to larger without mode collapse or size reversals. We therefore pair LFS with **Size Monotonicity (SM)**, which scales the garment measurements across S ordered factors $s_1 < \cdots < s_S$ and evaluates whether the aggregate coverage $C(s_j) = \sum_k w_k c_k(\hat{I}(s_j))$ increases monotonically across all $\binom{S}{2}$ ordered pairs ($s_j < s_l$) as $\text{SM} = (N_{\text{conc}} + 0.5\, N_{\text{tied}})/\binom{S}{2}$, where N_{conc} and N_{tied} denote the number of concordant pairs ($C(s_l) > C(s_j)$) and ties.

5 Experiments

5.1 Experimental Setup

Implementation Details. We build RealFIT upon FLUX.2 [klein] 9B (Black Forest Labs, 2025), training rank-128 LoRA adapters and our Fourier–MLP measurement encoder with batch size 128

Table 1: Quantitative comparison on RealFIT-Web and RealFIT-Bench. Best results are in **bold**; second-best are underlined. †Evaluated on upper-body tops only. *Grayed out: inpainting with a mask cut from I_{gt}, which leaks the target garment's silhouette and leads to artificially boosted scores.

	RealFIT-Web					RealFIT-Bench				
Method	**LFS ↑**	**SM ↑**	**SSIM ↑**	**LPIPS ↓**	**FID ↓**	**LFS ↑**	**SM ↑**	**SSIM ↑**	**LPIPS ↓**	**FID ↓**
IDM-VTON	0.8058	50.18	0.8306	0.1773	20.05	0.7984	52.15	0.8449	0.1520	24.09
Qwen-Image-Edit	0.8648	51.19	0.8344	0.1606	17.00	0.8241	53.05	0.8255	0.1767	22.45
FIT†	0.9072	85.81	0.8218	0.1935	24.95	0.8402	80.50	0.8384	0.1764	29.74
Nano Banana 2	0.9211	70.14	0.8285	0.1672	14.94	0.8876	70.33	0.8255	0.1697	19.56
Ours w/ Agn. Mask*	0.9846	50.34	0.8759	0.1277	11.30	0.9393	49.74	0.8678	0.1183	14.03
Ours w/o Ill-Fit Aug.	0.9569	56.94	0.8329	0.1539	13.52	0.9472	54.72	0.8327	0.1435	14.24
Ours w/ Text Meas.	0.9590	89.53	0.8358	0.1499	13.31	0.9299	85.74	0.8313	0.1425	16.35
Ours w/o LMC Loss	0.9646	90.15	**0.8549**	0.1207	12.57	0.9406	84.86	0.8519	0.1136	13.74
Ours	**0.9691**	**93.14**	0.8540	**0.1202**	**12.47**	**0.9476**	**90.54**	**0.8535**	**0.1088**	**13.69**

and learning rate 10^{-4}. Target images are generated at 1024×768 resolution while conditioning images I_p and I_g are encoded at 512×384. All RealFIT variants are pretrained on RealFIT-Web for 100k steps; during finetuning the model is co-trained on a 50/50 mixture of RealFIT-Web and RealFIT-Bench ($\mathcal{D}^{\text{ft}}_{\text{bench}}$) for 10k steps; full implementation details are provided in Appendix B.

Benchmarks and Baselines. We evaluate fit accuracy (LFS) and controllability (SM) alongside visual fidelity (SSIM, LPIPS, and FID) on two held-out benchmarks: RealFIT-Web Test (5,000 pairs) and RealFIT-Bench Test (1,395 pairs). Both test sets contain no ill-fit augmentation for real-world grounded evaluation. To quantify measurement controllability (SM) while being computationally manageable, we randomly sample a subset of 500 pairs from each benchmark and synthetically sweep all garment measurements across 9 uniform scaling increments ranging from -20% to $+20\%$ relative to their ground-truth sizes. We compare our approach against state-of-the-art mask-based VTO (IDM-VTON (Choi et al., 2024)), simulation-trained fit-aware VTO (FIT (Wang et al., 2026)), and powerful multimodal image editors (Qwen-Image-Edit (Wu et al., 2025) and Nano Banana 2 (Google DeepMind, 2026b)). Since FIT is only trained on top garments for measurement condition, we evaluated it on the subsets of top garments. For Qwen-Image-Edit and Nano Banana 2, we passed the measurement input into the editing prompts.

5.2 Quantitative Analysis

Quantitative results are shown in Table 1. IDM-VTON and Qwen-Image-Edit generate default well-fitted try-on results and are effectively blind to changing dimensions: their Size Monotonicity remains near random chance. Frontier multi-modal generation model (Nano Banana 2) captures coarse sizing intent, but text-based reasoning lacks continuous calibration, leading to frequent size reversals across fine-grained metric perturbations. Simulation-trained FIT captures better directional sizing on upper-body tops where its 3D parametric templates apply, but cannot generalize to diverse garment categories and exhibits poorer image quality as shown by the high FID scores. By grounding continuous metric conditioning in real photography, **RealFIT** achieves the strongest point-wise fit accuracy, size controllability, and visual fidelity (SSIM, LPIPS, and FID) across both benchmarks.

5.3 Qualitative Evaluation

As shown in Fig. 5, visual comparisons across scale perturbations (-10%, Input Size, and $+10\%$) mirror these quantitative findings. IDM-VTON and Qwen-Image-Edit render static, size-agnostic silhouettes across all three scales. While FIT successfully adjusts garment length on basic tops, it introduces synthetic artifacts and struggles on structurally complex categories such as layered dresses (Fig. 5a) or trousers (Fig. 5b). Nano Banana 2 responds to text-based size prompts but frequently exhibits inconsistent length jumps. In contrast, **RealFIT** accurately matches the ground-truth physical dimensions at the *Input Size*, and smoothly transitions the garment drape from a cropped look at -10% to a relaxed fit at $+10\%$, all while preserving the wearer's identity and fine fabric textures. In Fig. 6, we further demonstrate our model's ability to perform continuous, fine-grained size control from -20% to $+20\%$ in 5% increments across different garment categories. The garment coverage and tightness smoothly interpolate across sizes while maintaining high visual fidelity.

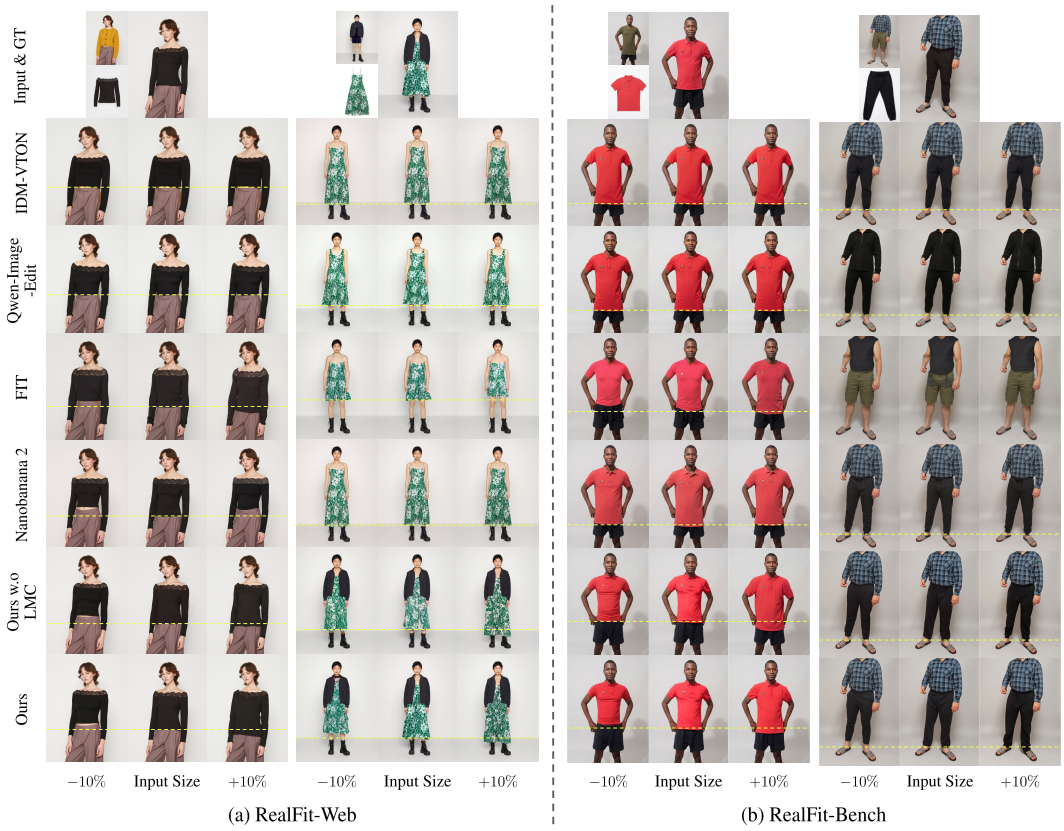

Figure 5: Qualitative comparison on RealFIT-Web and RealFIT-Bench. For each example, the top row shows the input person and garment (*left*) alongside the ground-truth image I_{gt} (*center*) corresponding to the *Input Size*. Each method generates three try-on results with garment dimensions varied by $\pm 10\%$ (-10%, *Input Size*, and $+10\%$). **RealFIT** (*Ours*) matches I_{gt} at the input size and achieves accurate, consistent size changes across scales with high visual quality. Please zoom in for a detailed view.



Figure 6: Qualitative results of RealFIT across increasing garment measurements, scaled from -20% to $+20\%$.

5.4 Ablation Studies

To validate RealFIT's key design choices, we evaluate four variants isolating individual components under identical training configurations (Table 1). First, adopting a traditional agnostic mask (**Ours w/ Agnostic Mask***) exposes a catastrophic shortcut: leaking the garment's spatial outline allows the model to reconstruct the silhouette while ignoring $\mathbf{m}$, inflating static image similarity while

collapsing Size Monotonicity to random chance (49.74%). Second, removing the size-perturbed pairs $\mathcal{D}_{\text{aug}}$ (**Ours w/o Ill-Fit Aug.**) achieves high LFS on test sets without ill-fit augmentation, as the network can infer a plausible default coverage from the layflat garment's visual proportions and moderate natural fit variations, but collapses SM to near random chance (54.72%): without paired counterfactuals of the same garment across sizes, the model tends to ignore $\mathbf{m}$. Third, serializing $\mathbf{m}$ as numerical prompt text (**Ours w/ Text Meas.**) degrades fit accuracy, controllability, and image fidelity, showing that even strong text encoders lack the continuous geometric inductive bias provided by Fourier features. Finally, removing the contrastive loss (**Ours w/o LMC Loss**) degrades fit accuracy (LFS) and monotonicity (SM); without $\mathcal{L}_{\text{LMC}}$, the model struggles to synthesize subtle sizing shifts at boundaries (Fig. 5), confirming that contrastive loss sharpens metric sensitivity.

5.5 Additional Analysis and Results

Subject Height Conditioning. RealFIT also adapts garment fit to wearer height under fixed garment dimensions: as shown in Fig. 7, decreasing height H elongates hemline drape, while increasing H yields a shorter fit. Sweeping H by ± 10 cm at fixed garment size, RealFIT achieves 88.89% monotonic concordance on RealFIT-Web and 78.54% on RealFIT-Bench.

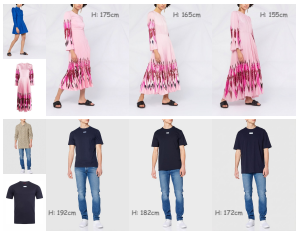

Figure 7: Varying height H.

Independent Measurement Control. In retail apparel and in our dataset, garment dimensions correlate strongly across standard grading sizes. Despite this co-variation, RealFIT exhibits emergent independent control over individual coordinates of $\mathbf{m}$: selectively modifying individual measurements alters only the corresponding dimension while leaving the others unchanged (Appendix A.3), confirming that our model learns to decouple localized dimensions rather than memorizing global scale factors.

LLM Annotation Accuracy. To evaluate our automated data curation, we benchmark the VLM annotation pipeline against tape-measured garments on RealFIT-Bench. Anchored to wearer height, population mean differences (and median absolute errors) between VLM estimates and manual measurements are: length -0.89 cm (3.0 cm), bust $+2.10$ cm (4.7 cm), waist $+0.56$ cm (2.5 cm), and sleeve -0.11 cm (2.25 cm). Negligible mean differences and median errors within manual measuring variance confirm that our automated annotations provide reliable supervision.

Standard Virtual Try-On Quality. Beyond fit-accurate synthesis, setting $\mathbf{m} = -\mathbf{1}$ allows RealFIT to operate as a conventional virtual try-on model, where it achieves state-of-the-art performance. On the VITON-HD paired test set with triplet inputs (the subject wearing a different garment), RealFIT outperforms both specialized VTO baselines and frontier image editors across all metrics (Appendix A.1), showing that our framework advances general try-on quality and metric fit control.

6 Conclusion

Limitations and Future Work. First, physical measurements carry inherent noise: manual tape measurements exhibit discrepancies bounding precision in supervision and evaluation. Second, garments with identical dimensions may fit differently depending on fabric elasticity and drape, which RealFIT does not explicitly condition on. Third, RealFIT cannot synthesize physically implausible wearing states absent from training data (e.g., oversized trousers slipping off). Furthermore, we observe minor color drifts in our results, caused by occasional color inconsistency between garment and person images in the dataset due to lighting variations. Finally, our Landmark Fit Score evaluates spatial extent rather than fabric tightness. We leave to future work to address these limitations.

We presented **RealFIT**, a virtual try-on framework conditioned on physical body and garment measurements for fit-accurate synthesis. We pair a large-scale, VLM-annotated pretraining corpus (RealFIT-Web) with a tape-measured benchmark of real subjects (RealFIT-Bench). To condition precisely on measurements, we embed them via a measurement encoder and introduce a Localized Metric Contrastive loss during training. To benchmark fit beyond visual similarity, we introduce Landmark Fit Score (LFS) and Size Monotonicity (SM). RealFIT outperforms existing VTO and multimodal editing baselines in both fit accuracy and controllability, while remaining a state-of-the-art standard try-on model when measurements are omitted. We hope our work encourages virtual try-on systems to show not only how a garment looks, but also how it fits.

7 Additional Information

AI use statement

In this work, we used generative AI tools to generate synthetic data sets, implement methods, and clean and reformat datasets. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, support qualitative and thematic data analysis, or interpret results. The use case of assisting with translation is not applicable to this work.

Additionally, we used generative AI tools to create or modify scientific figures or images, create or edit software code, polish writing, and format references.

We have reviewed all AI-assisted work by manually verifying and testing the edited code, as well as verifying our generated datasets through rounds of both synthetic and human QA. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

We have taken efforts to ensure reproducibility by publishing detailed descriptions of our model and dataset as well as the methods taken to create the dataset. The complete process and prompts used for all synthetic dataset generations and QA, as well as the data curation pipeline used to augment real data, are listed in the Appendix. The implementation details of the model and our measurement encoder are also listed in the Appendix.

References

Mohammadreza Armandpour, Ali Sadeghian, Huangjie Zheng, Amir Sadeghian, and Mingyuan Zhou. Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond. *arXiv preprint arXiv:2304.04968*, 2023.

Alberto Baldrati, Davide Morelli, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Multimodal garment designer: Human-centric latent diffusion models for fashion image editing. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23336–23345. IEEE, 2023.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2`, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL `https://arxiv.org/abs/2511.16719`.

Chieh-Yun Chen, Yi-Chung Chen, Hong-Han Shuai, and Wen-Huang Cheng. Size does matter: Size-aware virtual try-on via clothing-oriented transformation try-on network. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7479–7488. IEEE, 2023.

Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. *arXiv preprint arXiv:2103.16874*, 2021.

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.

Zheng Chong, Xiao Dong, Haoxiang Li, Wenqing Zhang, Hanqing Zhao, Dongmei Jiang, Xiaodan Liang, et al. Catvton: Concatenation is all you need for virtual try-on with diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 66586–66601, 2025.

Aiyu Cui, Jay Mahajan, Viraj Shah, Preeti Gomathinayagam, Chang Liu, and Svetlana Lazebnik. Street tryon: Learning in-the-wild virtual try-on from unpaired person images. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1414–1423. IEEE, 2025.

Yutong Feng, Linlin Zhang, Hengyuan Cao, Yiming Chen, Xiaoduan Feng, Jian Cao, Yuxiong Wu, and Bin Wang. Omnitry: Virtual try-on anything without masks. *arXiv preprint arXiv:2508.13632*, 2025.

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pp. 172–188. Springer, 2024.

Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. *arXiv preprint arXiv:2103.04559*, 2021.

Google DeepMind. Gemini 3.7 Flash: Model Card. `https://deepmind.google/models/model-cards/gemini-3-7-flash/`, August 2026a. Accessed: 2026-09-26.

Google DeepMind. Gemini 3.1 flash image. `https://gemini.google/overview/image-generation/`, 2026b.

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.

Hailong Guo, Bohan Zeng, Yiren Song, Wentao Zhang, Chuang Zhang, and Jiaming Liu. Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. *arXiv preprint arXiv:2501.15891*, 2025.

Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7543–7552. IEEE, 2018.

Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2328–2337. IEEE, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Zong-Wei Hong, Jing-lun Li, Lin-Ze Li, Shen Zhang, and Yao Tang. Vecor–velocity contrastive regularization for flow matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4739–4748, 2026.

Ding-Jiun Huang, Yuanhao Wang, Cheng Zhang, Hugo Bertiche, Alexandru-Eugen Ichim, Thabo Beeler, and Fernando De la Torre. Omnifabric: Coherent uv space texture synthesis for 3d garment reconstruction. *arXiv preprint arXiv:2609.30234*, 2026.

Jeongho Kim, Gyojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stable viton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8176–8185. IEEE, 2024.

Maria Korosteleva and Olga Sorkine-Hornung. Garmentcode: Programming parametric sewing patterns. *ACM Transactions on Graphics (TOG)*, 42(6):1–15, 2023.

Minoru Kuribayashi, Koki Nakai, and Nobuo Funabiki. Image-based virtual try-on system with clothing-size adjustment. *arXiv preprint arXiv:2302.14197*, 2023.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1096–1104, 2016.

MMPose Contributors. OpenMMLab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020.

Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *European Conference on Computer Vision*, pp. 345–362. Springer, 2022.

Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 8580–8589, 2023.

Hyelin Nam, Gihyun Kwon, Geon Yeong Park, and Jong Chul Ye. Contrastive denoising score for text-guided latent diffusion image editing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9192–9201. IEEE, 2024.

Yunfang Niu, Dong Yi, Lingxiang Wu, Zhiwei Liu, Pengxiang Cai, and Jinqiao Wang. Pfdm: Parser-free virtual try-on via diffusion model. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3780–3784. IEEE, 2024.

Le Shen, Yanting Kang, Rong Huang, and Zhijie Wang. Mfp-vton: Enhancing mask-free person-to-person virtual try-on via diffusion transformer. *arXiv preprint arXiv:2502.01626*, 2025.

George Stoica, Vivek Ramanujan, Xiang Fan, Ali Farhadi, Ranjay Krishna, and Judy Hoffman. Contrastive flow matching. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1185–1194. IEEE, 2025.

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.

Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2020.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 2024.

Runqian Wang and Kaiming He. Diffuse and disperse: Image generation with representation regularization. *arXiv preprint arXiv:2506.09027*, 2025.

Yuanhao Wang, Johanna Karras, Yingwei Li, and Ira Kemelmacher-Shlizerman. Fit: A large-scale dataset for fit-aware virtual try-on. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2026.

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.

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.

Yohei Yamashita, Chihiro Nakatani, and Norimichi Ukita. Size-variable virtual try-on with physical clothes size. *arXiv preprint arXiv:2412.06201*, 2024.

Lu Yang, Yicheng Liu, Yanan Li, Xiang Bai, and Hao Lu. Fitcontroler: Toward fit-aware virtual try-on. *arXiv preprint arXiv:2512.24016*, 2025.

Cheng Zhang, Yuanhao Wang, Francisco Vicente, Chenglei Wu, Jinlong Yang, Thabo Beeler, and Fernando De la Torre. Fabricdiffusion: High-fidelity texture transfer for 3d garments generation from in-the-wild images. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–12, 2024.

Xuanpu Zhang, Dan Song, Pengxin Zhan, Tianyu Chang, Jianhao Zeng, Qingguo Chen, Weihua Luo, and An-An Liu. Boow-vton: Boosting in-the-wild virtual try-on via mask-free pseudo data training. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26399–26408. IEEE, 2025.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4606–4615. IEEE, 2023.

Luyang Zhu, Yingwei Li, Nan Liu, Hao Peng, Dawei Yang, and Ira Kemelmacher-Shlizerman. M&m vto: Multi-garment virtual try-on and editing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1346–1356. IEEE, 2024.

APPENDIX

A Additional Experiment Results **15**

A.1 Standard Virtual Try-On on VITON-HD 15

A.2 Additional Qualitative Results 15

A.3 Emergent Independent Measurement Control 15

A.4 Failure Cases: Out-of-Distribution Measurements 17

B Implementation Details **17**

C Data Curation Pipeline **18**

C.1 VLM Size Annotation 18

C.2 Triplet Synthesis 19

C.3 Ill-Fit Augmentation 20

C.4 Dataset Statistics 21

This appendix provides supplementary material omitted from the main text due to space constraints. Appendix A.1–A.4 present additional experimental results, including standard measurement-free virtual try-on on VITON-HD, continuous size sweeps, emergent independent size control, and failure cases under out-of-distribution measurements. Appendix B details the network architecture, measurement encoder configuration, and training hyperparameters, including the setup for the Localized Metric Contrastive (LMC) loss. Appendix C describes the data curation pipeline for RealFIT-Web and RealFIT-Bench, including VLM size annotation, triplet synthesis, ill-fit augmentation, and dataset statistics.

A Additional Experiment Results

A.1 Standard Virtual Try-On on VITON-HD

Table 2: **Standard virtual try-on on the VITON-HD paired test set.** RealFIT is evaluated with all measurement conditions dropped. Each method receives a triplet input (the subject wearing a different garment) and is compared against the ground-truth photograph. All methods are mask-free except IDM-VTON†, which receives the person image with the garment region masked out. KID is reported $\times 10^3$. **Bold**: best; underline: second best.

Method	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
CatVTON (Chong et al., 2025)	0.7819	0.2244	19.91	3.74
Any2AnyTryon (Guo et al., 2025)	0.8191	0.1908	12.05	5.01
OOTDiffusion (Xu et al., 2025)	0.8020	0.1866	9.92	2.21
OmniTry (Feng et al., 2025)	<u>0.8247</u>	0.1734	11.18	3.66
IDM-VTON† (Choi et al., 2024)	0.8239	0.1673	9.49	2.38
Qwen-Image-Edit (Wu et al., 2025)	0.8004	0.1795	9.19	<u>1.99</u>
Nano Banana 2 (Google DeepMind, 2026b)	0.8246	<u>0.1644</u>	<u>8.73</u>	2.24
RealFIT (measurement-free)	**0.8321**	**0.1438**	**7.87**	**0.82**

Although RealFIT is designed for measurement-conditioned try-on, the measurement dropout used for classifier-free guidance (Appendix B) allows the same model to operate as a conventional garment-transfer model when all entries of $\mathbf{m}$ are replaced by null tokens. We evaluate this measurement-free mode on the paired test set of VITON-HD (Choi et al., 2021) without any dataset-specific finetuning, using triplet inputs in which the subject wears a different garment, so that each output can be compared against the ground-truth photograph. We compare against try-on models evaluated in the mask-free setting (CatVTON (Chong et al., 2025), Any2AnyTryon (Guo et al., 2025), OOTDiffusion (Xu et al., 2025), and OmniTry (Feng et al., 2025)), multimodal image editors (Qwen-Image-Edit (Wu et al., 2025) and Nano Banana 2 (Google DeepMind, 2026b)), and the mask-based IDM-VTON (Choi et al., 2024). All methods receive only the person and garment images, except IDM-VTON, which additionally takes the paired person image with the garment region masked out. As shown in Table 2, RealFIT achieves the best SSIM, LPIPS, FID, and KID, outperforming even IDM-VTON despite its access to an agnostic mask. This shows that measurement conditioning does not come at the cost of general try-on quality: training on real paired photography with our mask-free formulation also yields a strong standard VTO model.

A.2 Additional Qualitative Results

Fig. 8 shows additional results as all garment measurements are scaled jointly from -20% to $+20\%$, covering diverse garment categories, body types, and poses on both RealFIT-Web and RealFIT-Bench test examples.

A.3 Emergent Independent Measurement Control

In real-world apparel catalogs and benchmark photography, individual garment dimensions (length, bust, waist, and sleeve length) co-vary strongly across standard grading sizes (e.g., an XL garment is simultaneously longer, wider in the chest, and longer in the sleeves than an S). Even in our ill-fit augmentation pipeline ($\mathcal{D}_{\text{aug}}$), sizes are perturbed holistically rather than one dimension at a time, and we apply no explicit disentanglement objective during training. Nevertheless, as shown in Fig. 9, RealFIT exhibits emergent independent control over individual coordinates of $\mathbf{m}$: varying only the garment length shifts the lower torso hemline vertically while keeping chest fit and sleeve cuffs stationary; varying only the bust adjusts torso fullness without lengthening the hem or sleeves; and varying only the sleeve length moves the cuff along the forearm while keeping torso length and width fixed. This emergence indicates that our continuous Fourier–MLP measurement encoder and localized contrastive supervision teach the model to associate each entry of $\mathbf{m}$ with its corresponding physical boundary on the body.

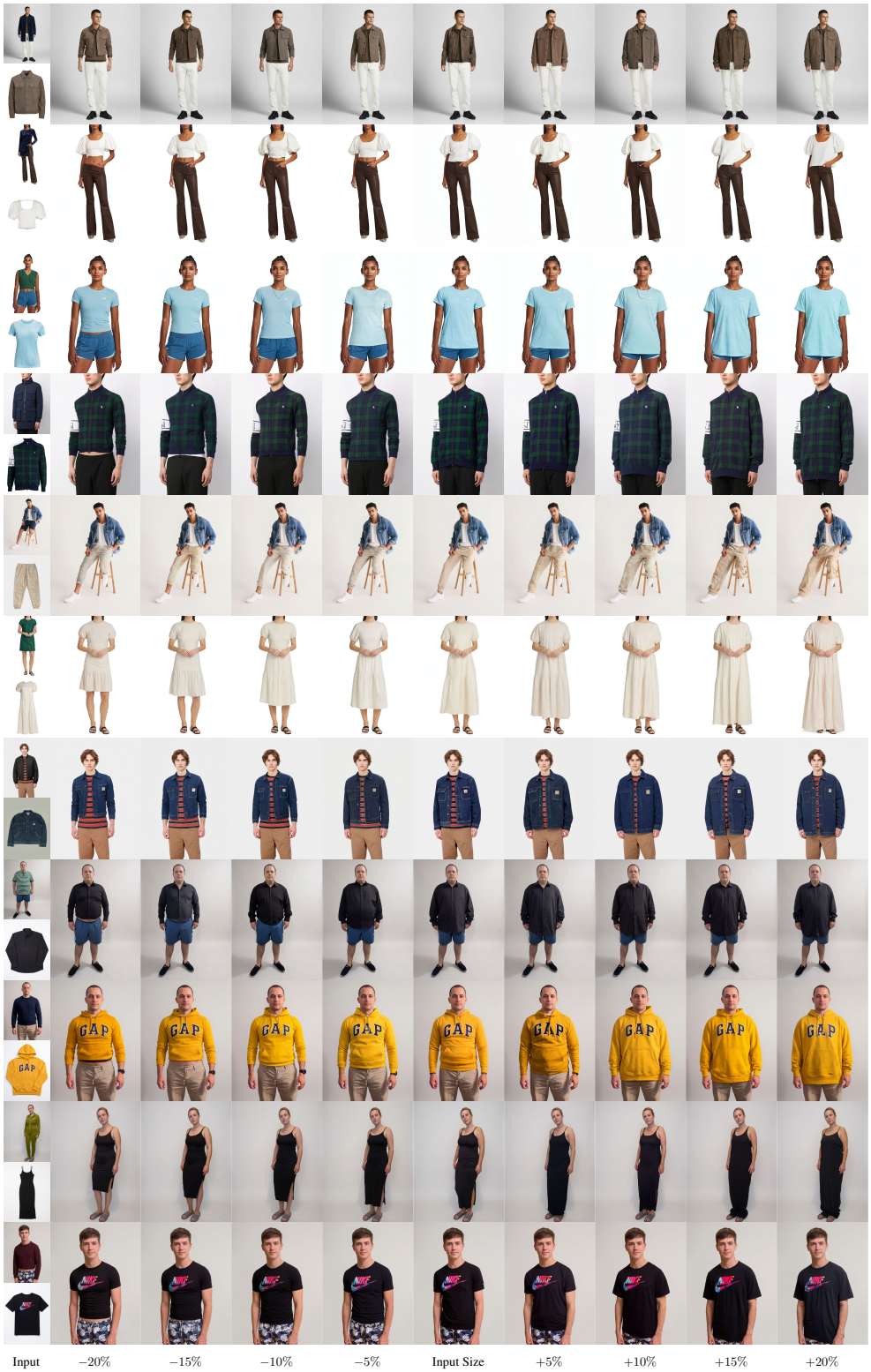

Figure 8: **Additional qualitative results.** RealFIT results as all garment measurements are scaled from -20% to $+20\%$. The bottom four rows are test examples from RealFIT-Bench; all other rows are from RealFIT-Web.



Figure 9: **Emergent independent control over individual garment dimensions.** Starting from the input person and layflat garment (*left*), we vary a single measurement by $\pm 10\%$, garment length (*Length Change*), bust (*Bust Change*), or sleeve length (*Sleeve Length Change*), while holding all other dimensions in $\mathbf{m}$ fixed at $+0\%$. Although garment dimensions are strongly correlated in real-world training data and are never explicitly disentangled during training, RealFIT learns to adjust each dimension independently without altering the others.

A.4 Failure Cases: Out-of-Distribution Measurements



Figure 10: **Failure case under out-of-distribution measurement scaling.** Starting from the input person and dress (*Input*), RealFIT accurately renders the target fit at *GT Size* and elongates the dress to ankle length at $+20\%$. However, when prompted with an extreme $+40\%$ scale increase that lies outside the training measurement distribution, the model fails to extrapolate and instead collapses to an implausibly short hemline.

Because RealFIT learns the mapping between physical dimensions and garment drape from data, its controllability is bounded by the support of the training measurement distribution. As illustrated in Fig. 10, when all measurements of the input dress are scaled within the in-distribution range (up to $+20\%$), RealFIT smoothly elongates the hemline from mid-calf at *GT Size* to ankle length at $+20\%$. However, when queried with an extreme $+40\%$ increase that exceeds both plausible grading ranges for this garment category and the perturbations seen during training, the model struggles to extrapolate and experiences mode collapse, reverting to an erroneously short dress.

B Implementation Details

Backbone and Asymmetric Multi-Resolution Packing. We instantiate RealFIT on the FLUX.2 (9B) rectified flow transformer (Black Forest Labs, 2025), which comprises 8 double-stream MMDiT blocks followed by 24 single-stream DiT blocks with hidden dimension 4,096 and joint conditioning width 12,288. To make high-resolution joint self-attention tractable, we adopt an asymmetric multi-resolution token budget: the target latent x_t is encoded at 1024×768 resolution, whereas the reference person I_p and layflat garment I_g are encoded at 512×384 resolution. In the 4D rotary positional embedding (T, Y, X, L), the spatial coordinates of I_p are scaled by $2\times$ so each conditioning patch (y, x) aligns with its corresponding 2×2 neighborhood $(2y, 2x)$ in x_t, while significantly reducing the total visual sequence length. We insert trainable LoRA adapters of rank $r = 128$ across all linear layers of the transformer while keeping the pretrained backbone and autoencoder frozen.

Measurement Encoder. Each 5-dimensional measurement vector $\mathbf{m} \in \mathbb{R}^5$ (subject height h and garment length, bust, waist, and sleeve length, in centimeters, with -1 indicating unused category

dimensions) is normalized and lifted using $K = 4$ Fourier frequencies $\{2^0, 2^1, 2^2, 2^3\}$, yielding a 40-dimensional $\sin / \cos$ feature vector. This vector is processed by a two-layer MLP with hidden dimension 512 and SiLU activations, followed by a linear projection reshaped into 4 conditioning tokens of dimension $12{,}288$ that are prepended to the 128-token text sequence. During training, we drop $\mathbf{m}$ (replacing it with null sentinel tokens) with probability 0.15 to support classifier-free guidance at inference.

LMC Hyperparameters and Optimization. In our final model, the LMC loss weight $\lambda(s)$ ramps linearly from 0 to 0.05 over the first $2{,}000$ training steps with no cold-start delay and remains at 0.05 thereafter. We apply $\mathcal{L}_{\text{LMC}}$ on high-noise timesteps $t \in [0.85, 1.0)$ (allocating 20% of training steps to this structural regime) for paired samples whose primary and perturbed garment dimensions differ by at least 5%. The active boundary region $\mathcal{R}$ selects the top $k = 20\%$ spatial tokens ranked by ground-truth latent deformation energy $e_i = \|z'_i - z_i\|_2^2$, and the softplus margin loss uses sharpness $\beta = 5.0$ and target margin $\tau = 0.55$. All models and ablation variants are trained with AdamW at a constant learning rate of 10^{-4} and a global batch size of 128 across 128 TPU v6e chips (1 sample per chip) for 100k pretraining steps on RealFIT-Web, followed by 10k finetuning steps on a 50/50 mixture of RealFIT-Web and RealFIT-Bench ($\mathcal{D}^{\text{ft}}_{\text{bench}}$). At inference, we generate samples using 28 Euler integration steps.

C Data Curation Pipeline

This section details the three generative and annotation steps used to construct RealFIT-Web and RealFIT-Bench (Section 3): VLM-based size annotation, triplet synthesis, and ill-fit augmentation. Each step is followed by automated quality assurance (QA).

C.1 VLM Size Annotation

We estimated garment dimensions with Gemini 3.7 Flash (Google DeepMind, 2026a) from a side-by-side image of the same person wearing the original garment (left) and the ill-fit augmented garment in a different size (right). We treat height estimation and garment measurement as a two-call self-verification procedure designed to improve label stability. Call 1 jointly estimates the person's height and all category-specific garment measurements. Call 2 fixes the height estimated by Call 1 as the scale anchor and independently re-estimates the same measurements. We label length, bust circumference, and sleeve length for upper garments; length and waist circumference for lower garments; and all four measurements for full-body garments. Bust and waist denote full circumferences, while length is measured shoulder-to-hem for upper and full-body garments and waistband-to-hem for lower garments. For each field, both calls estimate the original and resized garments, and the resizing delta is the right-image value minus the left-image value.

We flag a field when either its original- or resized-garment estimate differs between calls by more than the value tolerance. We also flag it when the two calls disagree on the magnitude of the resizing-induced change by more than the delta tolerance; this criterion does not limit the size change itself. The value/delta tolerances are 8/10 cm for length, 6/8 cm for sleeve length, 10/12 cm for bust circumference, and 10/12 cm for waist circumference. In the final aggregation, flagged measurements are discarded and represented by the sentinel value -1; for each unflagged measurement, the final label is the arithmetic mean of the two call-level predictions. The second call therefore both verifies the first prediction and reduces sampling noise in the retained labels.

Complete example prompt. The following is the complete Call 2 prompt for an upper-body sample depicting a blue long-sleeve shirt, resized smaller, using the 170 cm height estimated in Call 1.

```
You are an expert in fashion tailoring, human anatomy, and computer vision.

This side-by-side image shows the SAME person in two photos:
- LEFT: wearing blue long-sleeve shirt in its original fit
- RIGHT: wearing the same blue long-sleeve shirt in an altered size

The person's total standing height is 170 cm. Use this exact value as your
scale for every measurement below.
```

```
All values are PRODUCT measurements -- how the garment itself measures, not
the wearer's body. An oversized garment measures much larger than its wearer.

For EACH of the two garments, estimate:
  - blue long-sleeve shirt length (shoulder seam to hem)
  - bust circumference (full circumference around the chest)
  - sleeve length (shoulder seam to cuff)

If the garment is sleeveless, output 0 for sleeve length.

Provide concise reasoning (3-4 sentences), then output exactly these tags:
<left_length>[integer cm]</left_length>
<right_length>[integer cm]</right_length>
<left_bust>[integer cm]</left_bust>
<right_bust>[integer cm]</right_bust>
<left_sleeve>[integer cm]</left_sleeve>
<right_sleeve>[integer cm]</right_sleeve>
```

C.2 TRIPLET SYNTHESIS

To obtain the person image I_p for mask-free training, we edit I_{gt} with Nano Banana 2 (Google DeepMind, 2026b) to replace the target garment with a different one while preserving subject identity, pose, camera, and background. The replacement was required to differ visibly in color, material, texture, style, and length, while person identity, body shape, pose, skin, other garments, accessories, background, camera, crop, and framing remained fixed. A lightweight vision-language model generated three descriptions of the edited garment to improve SAM3 segmentation. We then accepted a candidate only when SAM3 localized the pixel changes to the garment region and semantic QA confirmed that the target garment changed while pose, body shape, and all non-target content were preserved. Failed candidates were retried with QA correction feedback; the fourth attempt used a fallback prompt emphasizing a different design and cut. The accepted garment-swap image, together with the original person and garment conditioning inputs, formed the training triplet.

Garment-swap edit prompt example. The following is the complete primary prompt for replacing a blue long-sleeve shirt.

```
Target garment is 'blue long-sleeve shirt'. Replace only this target garment
with a completely different top or shirt. The replacement must look obviously
different: choose a new color, fabric, texture, and style that clearly contrasts
with the original. It must also have a noticeably different length from the
original (e.g. cropped vs. longline). CRITICAL: this is a pure garment swap --
everything except the target garment must stay pixel-for-pixel identical. Do not
change the person's face, head angle, hand position, body pose, body shape, skin
tone, hairstyle, any other garment, accessories, or the background. Do not
change the camera angle, zoom level, crop, or framing -- the same amount of body
and face must remain visible as in the original.
```

Semantic QA prompt.

```
You are a quality-control reviewer for garment-swap image edits.

You are given three images:
1. ORIGINAL: the original model photo
2. EDITED: the edited photo where the target garment was replaced with a
   different one
3. DIFF: a grayscale pixel difference map; brighter pixels indicate larger
   change

The target garment is: blue long-sleeve shirt

Evaluate these criteria:
1. garment_changed: Is the target clearly replaced with a visually distinct
```

```
   garment? It must differ in color, style, or design. Flag false if identical
   or nearly identical.
2. only_garment_changed: Are all other elements unchanged? Expected skin
   exposure changes from neckline, sleeve, hem, or leg-length changes are
   allowed. Flag only independently changed clothing, accessories, face, hair
   structure, background, or skin tone.
3. pose_unchanged: Are body pose and image framing identical?
4. body_shape_preserved: Are visible skin contours and bust/waist proportions
   preserved, allowing garment-volume changes?

Respond with exactly:
<garment_changed>true/false</garment_changed>
<only_garment_changed>true/false</only_garment_changed>
<pose_unchanged>true/false</pose_unchanged>
<body_shape_preserved>true/false</body_shape_preserved>
<reason>one sentence describing failures, or "pass"</reason>
<correction_note>if any criterion failed, a 1-2 sentence imperative correction;
otherwise "pass"</correction_note>
```

C.3 Ill-Fit Augmentation

To supply training examples of garments that are too tight or too loose, we re-render the garment in I_{gt} at an altered size with Nano Banana 2, holding subject, pose, background, and garment design fixed, yielding a paired image I'_{gt}. A deterministic hash assigned each sample to a "smaller" or "larger" edit, and the model was instructed to alter only the target garment by one to two sizes while preserving its design, fabric, pattern, and open/closed state. The requested change included a visible fit adjustment and an approximately 5–10cm hem displacement. Person identity, body shape, pose, all non-target garments and accessories, background, camera, crop, and framing were explicitly locked. Each candidate passed two quality-control stages: SAM3 segmented the target garment and person to ensure that changes outside the garment remained below an outside-difference threshold, and a semantic vision-language check verified that the requested fit change was achieved without unintended edits. Failed candidates were regenerated with correction feedback for up to four attempts; from the third attempt onward, the target direction was reversed to improve recovery. Only candidates passing both checks were retained.

Ill-fit edit prompt example. The following is the complete instantiated prompt for making a blue long-sleeve shirt smaller.

```
Target garment is 'blue long-sleeve shirt'. Resize only this target garment so
it fits 1-2 sizes too small: snug and tight, with the hem sitting noticeably
higher than a properly fitted version. A 1-2 size difference should shift the
hem by about 5-10 cm. Keep the garment's exact original design, silhouette,
fabric, and pattern; change only fit and length. Preserve the garment's exact
state: if it is open, unzipped, or unbuttoned, keep it that way. Edit only the
target garment. Every other element is locked: person identity, body shape,
skin, face, hair, pose, shoes, bags, accessories, other garments, and background
must remain pixel-identical. Do not regenerate or recompose any area outside the
target garment. No zoom, crop, camera, or framing changes.
```

Gemini semantic QA prompt. The exact production prompt is shown below, instantiated for a blue long-sleeve shirt and line-wrapped for typesetting.

```
You are a quality-control reviewer for garment ill-fit (resize) image edits.

You are given three images:
1. GARMENT: a clean product photo of the target garment
2. ORIGINAL: the original model photo
3. ALTERED: the edited model photo, where only the target garment's fit/length
was resized (made intentionally too small or too large)

The target garment is: blue long-sleeve shirt
```

```
Evaluate these two criteria:

1. design_preserved: In ALTERED, does blue long-sleeve shirt retain the same
design elements, fabric, texture, color, and construction as shown in GARMENT?
IMPORTANT: blue long-sleeve shirt in ALTERED is intentionally a different size
and fit -- it will look longer or shorter, tighter or looser, than in GARMENT.
These length and fit changes are expected and must NOT be flagged. Only flag if
non-size design elements changed: buttons added or removed, fabric pattern or
texture changed, pocket placement altered, neckline/collar style changed,
silhouette type changed, color changed, logo/graphic changed or moved, or the
garment replaced entirely with a different design.

2. body_shape_preserved: Comparing ORIGINAL to ALTERED, is the person's
underlying body unchanged?

Pay particular attention to BUST SIZE, but flag only an obvious change -- a
clearly different cup size or projection. Compression, lift, and coverage
changes from a tighter or looser garment are expected; ignore subtle or
borderline differences.

Then check these anchors, which do not depend on what the garment is doing:
- shoulder width measured relative to head width
- exposed skin: upper-arm and forearm thickness, wrist width, and where visible,
  thigh, calf, and ankle width
- waist and hip width where skin or a non-target garment is visible
- head-to-body height ratio and overall stature

Do NOT flag the garment's own volume, drape, tightness, length, or how much skin
it covers. A tighter garment revealing a narrower silhouette, or a looser one
concealing it, is the intended edit and must pass. Judge the body itself, not
the garment's outline: if the only evidence of change is the outline of the
target garment, answer true.

Respond with exactly:
<design_preserved>true/false</design_preserved>
<body_shape_preserved>true/false</body_shape_preserved>
<reason>one sentence describing any failures, or "pass" if both pass</reason>
<correction_note>if either criterion failed, a 1-2 sentence imperative
correction for the next edit attempt; otherwise "pass"</correction_note>
```

C.4 Dataset Statistics

Table 3 summarizes the composition and physical measurement distributions of both **RealFIT-Web** and **RealFIT-Bench** (complementing Fig. 2c). Starting from 172,386 real catalog pairs, our ill-fit augmentation pipeline generates 172,386 paired counterfactual samples in $\mathcal{D}_{\text{aug}}$ (balanced between 51.5% smaller/tighter and 48.5% larger/looser perturbations), yielding 344,772 total pretraining samples in $\mathcal{D}_{\text{web}}$ alongside a held-out test split of 5,000 pairs. Across paired perturbations in RealFIT-Web, garment dimensions shift by an average of 9.6%–12.1% (9.1 cm in length, 11.4 cm in bust, 8.4 cm in waist, and 3.5 cm in sleeve length), expanding the 5th–95th percentile coverage while preserving the underlying mean dimensions across 12 fine-grained apparel categories (59.7% upper-body tops, 24.4% full-body dresses, and 15.9% lower-body bottoms).

For **RealFIT-Bench** (168 subjects and 987 unique garments in total), partitioning by subject and garment yields a held-out evaluation split $\mathcal{D}^{\text{test}}_{\text{bench}}$ of 1,395 real on-body test pairs (19 subjects, 199 garments, with all synthesized ill-fit images excluded) and a finetuning split $\mathcal{D}^{\text{ft}}_{\text{bench}}$ of 11,646 samples across 149 subjects and 788 garments (5,829 original on-body photographs and 5,817 ill-fit augmented counterparts, split 45.0% smaller/tighter and 54.8% larger/looser with 10.0%–13.6% mean dimensional shifts). Compared to studio catalog models in RealFIT-Web (179.9 ± 5.7 cm, P5–P95: 175.0–188.0 cm), RealFIT-Bench spans a much broader real-world subject height distribution (168.2 ± 9.7 cm, P5–P95: 152.0–183.0 cm). Furthermore, across the 5,829 unaugmented on-body photographs in $\mathcal{D}^{\text{ft}}_{\text{bench}}$, participants' self-reported natural garment fit spans 35.9% well-fitting (2,095), 28.8% somewhat loose (1,679, 1–2 sizes larger), 16.7% somewhat tight (976, 1–2 sizes smaller),

10.4% very loose (604, >2 sizes larger), and 8.1% very tight (475, >2 sizes smaller)—providing over 64% naturally occurring tight or loose real-world fit examples prior to synthetic augmentation.

Table 3: **Summary statistics of RealFIT-Web ($\mathcal{D}_{\text{web}}$, $N = 344{,}772$) and RealFIT-Bench ($\mathcal{D}^{\text{ft}}_{\text{bench}}$, $N = 11{,}646$). Left:** Physical measurement distributions (in cm, over valid positive entries) and mean paired size perturbations $|\Delta|$. **Right:** Garment category breakdown across both corpora.

Dimension	Mean ± Std	Median	P5–P95	Paired $\lvert\Delta\rvert$
(a) RealFIT-Web *($\mathcal{D}_{web}$,* 344,772 *samples /* 172,386 *pairs)*				
Height H (cm)	179.9 ± 5.7	177.0	175.0–188.0	—
Length L (cm)	81.0 ± 23.3	76.0	47.0–125.5	9.1 cm (12.1%)
Bust B (cm)	102.3 ± 14.3	102.0	81.5–126.0	11.4 cm (11.2%)
Waist W (cm)	78.8 ± 12.1	76.0	65.5–104.0	8.4 cm (10.6%)
Sleeve S (cm)	47.9 ± 19.3	59.0	17.0–67.5	3.5 cm (9.6%)
(b) RealFIT-Bench *($\mathcal{D}^{ft}_{bench}$,* 11,646 *samples across* 149 *subjects,* 788 *garments)*				
Height H (cm)	168.2 ± 9.7	167.0	152.0–183.0	—
Length L (cm)	75.5 ± 22.1	72.0	42.0–113.0	9.4 cm (13.6%)
Bust B (cm)	98.5 ± 18.0	98.0	70.0–128.0	12.4 cm (12.7%)
Waist W (cm)	76.6 ± 13.8	76.0	58.0–100.0	7.5 cm (10.0%)
Sleeve S (cm)	38.4 ± 20.3	28.0	14.0–67.0	4.1 cm (12.9%)

Garment Type	RealFIT-Web	RealFIT-Bench
Top (Upper Body)	**205,948 (59.7%)**	**7,063 (60.7%)**
T-Shirt / Polo	60,210 (17.5%)	3,020 (25.9%)
Jacket / Coat	52,360 (15.2%)	1,288 (11.1%)
Shirt / Blouse	34,838 (10.1%)	1,210 (10.4%)
Hoodie / Sweatshirt	26,354 (7.6%)	741 (6.4%)
Sweater / Tank / Other	32,186 (9.3%)	804 (6.9%)
Full Body (Dresses)	**84,028 (24.4%)**	**1,004 (8.6%)**
Bottom (Lower Body)	**54,796 (15.9%)**	**3,579 (30.7%)**
Pants / Jeans / Joggers	47,642 (13.8%)	2,633 (22.6%)
Skirt / Shorts	7,154 (2.1%)	946 (8.1%)