FITFLOW: DECOUPLED FLOW MATCHING MODEL FOR FIT-AWARE VIRTUAL TRY-ON

Anonymous authors
Paper under double-blind review

ABSTRACT

Virtual try-on (VTON) has evolved from merely generating visually appealing results to focusing on the precise fit between the garment and the target body. Existing methods typically rely on simulated data and leverage numerical control or a structured prompt to achieve garment fit. However, such control signals lack spatial precision for continuous fit adjustment. While the reliance on synthetic data fails to capture fine-grained fit textures (e.g., elasticity and subtle draping), leading to visual realism degradation. Motivated by the nature of flow-matching trajectories, where high-noise timesteps establish global layout and low-noise timesteps restore fine details, we propose **FitFlow**, a decoupled flow matching model that leverages large-scale synthetic data to learn global body-aware fitting and real try-on pairs at late timesteps to recover fine-grained fit textures. We introduce a simple Garment Size Encoding, that warps the garment image onto a canonical grid based on its relative size to the target body, providing continuous fit control while maintaining photorealistic garment appearance. We also develop a measurement-annotated try-on dataset and evaluation protocol for authentic try-on and fit analysis. Experiments demonstrate that our method achieves improved fit consistency, spatial controllability, and visual fidelity over state-of-the-art baselines.

1 INTRODUCTION

Virtual try-on (VTON) aims to synthesize a realistic image of a target person wearing a specified garment. Earlier VTON methods improve garment alignment through geometric warping and appearance-flow-based transformations (Han et al., 2018; Wang et al., 2018; Ge et al., 2021; Xie et al., 2023), while recent diffusion-based approaches further advance visual realism and high-resolution generation (Zhu et al., 2023; Kim et al., 2024; Yang et al., 2024; 2025b). However, practical virtual try-on requires more than faithful garment appearance transfer; it also requires accurate and controllable modeling of how the garment fits the target body.

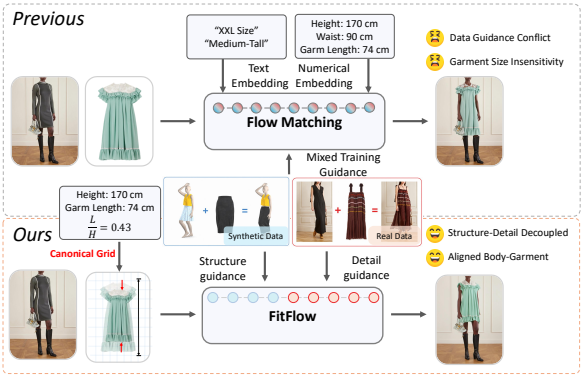

Figure 1: Comparison between previous studies and **FitFlow**. FitFlow uses structure-detail decoupled learning and garment size encoding.

Despite recent progress in fit-aware virtual try-on, accurately modeling relative garment–body appearance remains challenging. As illustrated in Fig. 1, existing methods try to introduce fitting information primarily through editable text prompts or numerical measurement controls (Kim et al., 2025; Yang et al., 2025a; Yao & Zheng, 2023; Chen et al., 2023; Yamashita et al., 2024). Although these approaches can generate quantitatively varied fitting results, they fail to realistically reflect how the same garment drapes and textures differently across body sizes. We attribute this limitation to the following two factors.

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.

Firstly, there is limited high-quality realistic data available, particularly paired contrastive observations of the same garment worn by individuals with different body measurements, termed "one-garment-to-multiple-bodies". Secondly, we experimentally find that garment size is essential for generated try-on quality but typical numerical encodings achieve sub-optimal performance caused by generative models inherently insensitive to numerical input.

To alleviate the data scarcity dilemma, some efforts like FIT (Karras et al., 2026) attempt to mix limited real-world try-on data mined from the Internet and a massive synthetic "one-garment-to-multiple-bodies" contrastive dataset generated by physics-based simulators (Korosteleva & Sorkine-Hornung, 2023) during model training. However, it suffers from guidance conflict: synthetic data provides abundant cross-size pairs but lacks realism, while real data offers high-fidelity textures but lacks cross-size observations.

In this paper, we propose a hybrid training paradigm **FitFlow**, which is inspired by a key observation that flow-matching models generate images from coarse structure to fine details. As shown in Fig. 2, early high-noise timesteps shape macro geometry and body–garment alignment, whereas late low-noise timesteps focus on localized textures and wrinkles.

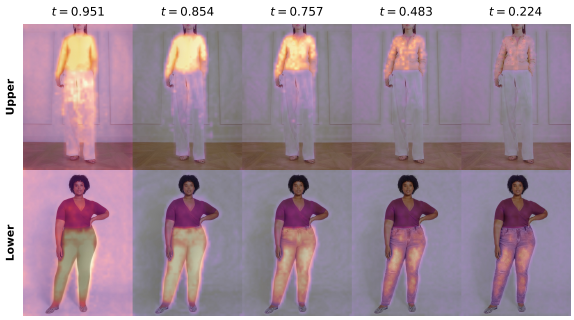

Figure 2: Attention maps over flow-matching timesteps. Garment-token attention transitions from global layout at high noise to local texture and boundary details at low noise.

Therefore, we build FitFlow as a decoupled flow matching model, as shown in Fig. 1, that leverages different data sources along a single generation trajectory to learn fitting dynamics. In early timesteps, we use synthetic cross-size pairs to teach the model how garments deform across body sizes, thereby controlling the overall fit. In late timesteps, we utilize real try-on data to render natural textures and localized fit details, such as wrinkles and stretching. To ensure a smooth transition between the two stages, we introduce a **structural bridge** at the switching timestep. Because all steps share a single generator, FitFlow completes inference in a single sampling trajectory without requiring secondary refinement models.

In addition, to make size information more directly usable by the generator, we introduce **Garment Size Encoding (GSE)**, a parameter-free scheme that maps the garment-to-body length ratio onto a canonical grid. Instead of requiring the model to infer the body–garment relationship from abstract numerical embeddings or coarse fit prompts, GSE represents the desired garment extent explicitly in image space. By resizing the garment to a body-relative spatial scale, this image-native condition alleviates numerical insensitivity and better leverages the pretrained model's visual tokens and positional priors for precise and continuous fit control.

Existing VTON benchmarks (Morelli et al., 2022; Choi et al., 2021) mainly evaluate appearance transfer and identity preservation, but lack paired body–garment measurements for assessing size-dependent fit. We therefore construct a measurement-annotated real-world dataset covering diverse body shapes, height groups, and garment sizes. To evaluate, we propose two fitting metrics Hem Fitting Error (HFE) and Sleeve Fitting Error (SFE), together with the standard garment-mask IoU and image-quality metrics such as SSIM, LPIPS, FID, and KID (Wang et al., 2004; Zhang et al., 2018; Heusel et al., 2017; Bińkowski et al., 2018). This protocol jointly evaluates fitting accuracy and photorealistic generation quality. Our contributions are summarized as follows:

- We propose **FitFlow**, a decoupled coarse-to-fine VTON framework that assigns synthetic fit supervision to high-noise timesteps and real appearance supervision to low-noise timesteps within a single flow-matching trajectory.
- We introduce **Garment Size Encoding (GSE)**, a parameter-free scheme that converts body–garment measurements into a body-relative visual scale for fine-grained fit control.
- We construct a **Fit-aware VTON benchmark** with diverse body shapes and garment sizes, together with fit-specific evaluation metrics.

2 Related Work

2.1 Image-Based Virtual Try-On

Image-based virtual try-on aims to synthesize a realistic image of a target person wearing a specified garment while preserving garment appearance and personal identity. Early methods typically separate garment alignment from image synthesis, using geometric warping, appearance flow, or semantic parsing to establish garment–body correspondence (Han et al., 2018; Wang et al., 2018; Ge et al., 2021; Xie et al., 2023; Wan et al., 2024). Diffusion-based approaches substantially improve visual quality and garment-detail preservation by leveraging pretrained generative priors (Zhu et al., 2023; Zhang et al., 2025; Kim et al., 2024; Xu et al., 2025; Yang et al., 2024; Morelli et al., 2023; Lin et al., 2025; Wang et al., 2025; Han et al., 2026). More recent methods further enhance flexibility and reference utilization: Any2AnyTryon (Guo et al., 2025) supports versatile clothing tasks through adaptive position embeddings, RefTON (Li et al., 2026) introduces additional reference-person guidance, and OmniTry (Feng et al., 2026) enables mask-free try-on across diverse garments and scenarios. Despite their strong synthesis quality, most existing methods focus on appearance transfer, while the measurement-dependent fit between the garment and the target body remains largely implicit. In contrast, FitFlow explicitly conditions generation on body–garment measurements to produce realistic try-on results consistent with the specified fit relationship.

2.2 Fit-Aware Virtual Try-On

Fit-aware VTON further considers how a garment spatially fits the target body. Existing methods introduce fit labels, text prompts, garment length, body shape, or garment-size conditions to improve controllability (Kim et al., 2025; Yang et al., 2025a; Yao & Zheng, 2023; Chen et al., 2023; Yamashita et al., 2024). Although these approaches enable semantic or attribute-level adjustment, their conditioning signals are often too coarse to capture fine-grained fitting variations governed by explicit body–garment measurements. More recently, FIT (Karras et al., 2026) incorporates body and garment measurements and leverages GarmentCode (Korosteleva & Sorkine-Hornung, 2023) for controllable fit supervision. However, the spatial effect of numerical measurements still needs to be inferred implicitly by the generator. In contrast, FitFlow represents garment measurements directly as a body-relative visual scale in the garment condition, while assigning synthetic fit supervision and real appearance supervision to different portions of the flow-matching trajectory.

3 Method

As illustrated in Fig. 3, we first present **FitFlow**, a decoupled flow-matching framework that assigns synthetic fit supervision and real appearance supervision to different timestep intervals of a single flow-matching trajectory. We then introduce **Structural Bridge**, which exposes the low-noise stage to synthetic-guided boundary states and learns a correction toward the real target. Finally, we present **Garment Size Encoding (GSE)** for constructing the measurement-aware garment condition.

3.1 Decoupled Flow Matching Framework

3.1.1 Preliminaries.

Given a target try-on image I_y, it can be encoded into the latent space as $z_0 = \mathcal{E}(I_y)$. Following flow matching (Liu et al., 2022), a noisy latent can be achieved via introducing Gaussian noise as follows, together with the corresponding target velocity:

$$z_t = (1-t)z_0 + t\epsilon, \tag{1}$$

where $t \in [0,1]$, and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ denote Gaussian noise. The target velocity is defined as $v = \epsilon - z_0$ and guides the optimization of the virtual try-on model v_θ using the standard flow-matching objective:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{z_0,t,\epsilon}\left[\left\|v_\theta\left(z_t, t, c, p\right) - v\right\|_2^2\right], \tag{2}$$

where p denotes an instruction for the try-on task and $c = [I_p, I_g]$ denotes the conditional inputs, which can simply be a concatenation of a person image I_p and a garment image.

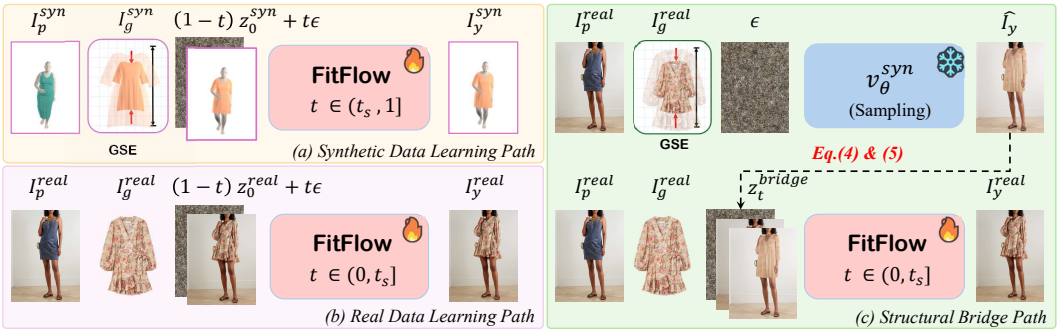

Figure 3: Overview of the three learning paths in FitFlow. (a) Synthetic data path ($t \in (t_s, 1]$): Trains the model in the high-noise regime to establish macro spatial alignment and body-relative fit priors. (b) Real data path ($t \in (0, t_s]$): Trains on real images to recover photorealistic appearance and fine-grained local details. (c) Structural bridge path ($t \in (0, t_s]$): Uses the auxiliary model v_θ^{syn} to generate a synthetic-guided anchor $z_{t_s}^{\text{anchor}}$, enabling the model to bridge coarse synthetic geometry with photorealistic target details. During training, paths (b) and (c) are stochastically activated with probabilities $1 - p_b = 0.3$ and $p_b = 0.7$, respectively.

3.1.2 Decoupled Synthetic-Real Supervision.

Teaching the generator how garments adapt across body sizes requires paired cross-size observations. Because real-world try-on datasets $\mathcal{D}_{\text{real}} = \{(I_p^{real}, I_g^{real}, I_y^{real})\}$ lack such fine-grained annotations, incorporating simulation-based synthetic data $\mathcal{D}_{\text{syn}} = \{(I_p^{\text{syn}}, I_g^{\text{syn}}, I_y^{\text{syn}})\}$ is essential. However, naively mixing these domains causes guidance conflict: synthetic try-on images provide precise structural fit control but lack photorealism, whereas images captured from the real world offer rich fabric textures but lack cross-size variations.

To resolve this conflict, we take full advantage of the coarse-to-fine nature of flow matching that high-noise stages establish macro structures, while low-noise stages refine local details (Fig. 2). **FitFlow** decouples the sampling trajectory by introducing a switching timestep $t_s \in (0, 1)$, assigning synthetic and real targets to separate noise intervals:

$$(I_p, I_g, I_y) \sim \begin{cases} \mathcal{D}_{\text{syn}}, & t \in (t_s, 1], \\ \mathcal{D}_{\text{real}}, & t \in (0, t_s]. \end{cases} \tag{3}$$

Specifically, synthetic samples are paired with high-noise timesteps $t \in (t_s, 1]$ to supervise macro spatial alignment and body-relative fit ($\mathcal{L}_{\text{syn}}$), whereas real samples are paired with low-noise timesteps $t \in (0, t_s]$ to learn photorealistic textures ($\mathcal{L}_{\text{real}}$).Both intervals update a single LoRA-adapted generator v_θ. We set $t_s = 0.95$ in practice.

3.1.3 Structural Bridge.

Abruptly switching supervision sources at t_s can induce structural inconsistency between the synthetic-guided macro layout and real texture refinement. We therefore introduce a *structural bridge* to connect the two intervals smoothly. Specifically, we first build an auxiliary velocity model v_θ^{syn}, which is trained only on $\mathcal{D}_{\text{syn}}$. Given real conditions $c^{\text{real}} = [I_p^{\text{real}}, I_g^{\text{real}}]$ sampled from $\mathcal{D}_{\text{real}}$, we use v_θ^{syn} to perform a complete sampling process from $t = 1$ to $t = 0$, obtaining a clean synthetic-guided try-on image $\hat{I}_y$. We encode this image as latent feature $\hat{z}_0 = \mathcal{E}(\hat{I}_y)$ and add noise to construct the structural anchor at the switching timestep:

$$z_{t_s}^{\text{anchor}} = (1 - t_s)\hat{z}_0 + t_s\epsilon. \tag{4}$$

We construct the structure bridge representation z_t^{bridge} by linearly interpolating between the real clean latent z_0^{real} and the structural anchor:

$$z_t^{\text{bridge}} = \left(1 - \frac{t}{t_s}\right) z_0^{\text{real}} + \frac{t}{t_s} z_{t_s}^{\text{anchor}}, \quad t \in (0, t_s]. \tag{5}$$

The corresponding target velocity is $v^{\text{bridge}} = \left(z_{t_s}^{\text{anchor}} - z_0^{\text{real}}\right)/t_s$. The structural bridge guides the model to smoothly transition between the learning path of synthetic and real data. We denote the flow-matching loss along this learning path by $\mathcal{L}_{\text{bridge}}$.

Although the bridge path enables the model to correct the synthetic structural anchor toward the real target, relying exclusively on this path may weaken the standard real-domain generation capability. Therefore, we activate the bridge path with probability p_b. The overall training objective is

$$\mathcal{L} = \begin{cases} \mathcal{L}_{\text{syn}}, & t \in (t_s, 1], \\ \mathcal{L}_{\text{bridge}}, & t \in (0, t_s], \quad p_{\text{b}}, \\ \mathcal{L}_{\text{real}}, & t \in (0, t_s], \quad 1 - p_{\text{b}}, \end{cases} \tag{6}$$

where $p_{\text{b}} = 0.7$. Note that v_θ^{syn} is only used for training. FitFlow simplifies to standard flow-matching sampling in a single sampling pass during the inference time.

3.2 Garment Size Encoding

Generative models are often insensitive to abstract numerical inputs, while numerical embeddings and coarse fit prompts leave the body–garment size relationship implicit. As illustrated in Fig. 1, such representations do not explicitly specify the desired garment extent on the target body. We therefore introduce **Garment Size Encoding (GSE)**, a parameter-free scheme that explicitly represents the desired garment extent in image space by mapping the garment-to-body length ratio onto a canonical grid.

Given the garment length l_g and the target body height h_b, we compute the body-relative garment ratio as $r_g = l_g/h_b$. Let $H_{\mathcal{C}}$ denote the canonical body height; the corresponding target garment height on the grid is then given by $H_g^{\mathcal{C}} = r_g H_{\mathcal{C}}$. The garment foreground is resized to the computed height while preserving its aspect ratio and then centered on the canonical grid, producing the encoded condition I_{gse}. Consequently, garments with different physical lengths occupy different body-relative extents, enabling the model to directly perceive size variations through its visual tokens and positional priors. We set $H_{\mathcal{C}} = 875$ pixels, corresponding to a canonical body height of 175 cm at a physical resolution of 0.2 cm/pixel in practice.

4 Dataset and Evaluation Protocol

4.1 Data Curation and Construction.

For real-data curation, we collect diverse fashion images from public websites and extract human pose skeletons using MMPose (MMPose, 2020) and garment landmarks via DeepFashion2 (Ge et al., 2019). Filtering by mask/pose quality and measurement validity yields 4,000 samples spanning diverse body types, which are split into mutually exclusive training (2,000) and test (2,000) sets. Because real-world data lacks controlled variations across garment sizes and body shapes, we perform synthetic-data construction using GarmentCode (Korosteleva & Sorkine-Hornung, 2023). Specifically, we simulate 102,400 try-on images across 16 human models and 19 garment templates (3 uppers, 8 lowers, and 8 dresses), with each garment generated in five randomly sampled sizes. This dual pipeline couples realistic image dynamics with controllable body–garment measurements. More details are provided in the supplementary material.

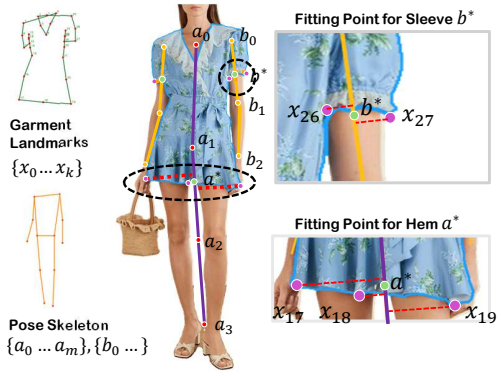

Figure 4: Illustration of Hem and Sleeve Fitting Point Calculation. Garment landmarks x_i are projected onto piecewise-linear body reference paths ($\mathcal{A}$ for hem, $\mathcal{B}$ for sleeves) constructed from pose keypoints. The aggregated projections yield fitting point a^* and b^*.

4.2 BODY-RELATIVE FIT EVALUATION PROTOCOL

In this work, we introduce two fitting metrics, Hem Fitting Error (HFE) and Sleeve Fitting Error (SFE). We first detect garment landmarks $\{x_i\}$ using DeepFashion2 (Ge et al., 2019). We then extract body keypoints using MMPose (MMPose, 2020). Based on these keypoints, we construct a category-specific body reference path $\mathcal{A} = \{a_0, \ldots, a_m\}$ for hem evaluation using the corresponding shoulder, hip, and ankle keypoints or their bilateral midpoints. For sleeve evaluation, we define $\mathcal{B} = \{b_0, b_1, b_2\}$ using the shoulder, elbow, and wrist keypoints of each arm. We project each garment landmark x_i onto its closest segment along the piecewise-linear body reference path, computing the fitting point for hem and sleeve as the mean of these projections:

$$a^* = \frac{1}{|\mathcal{H}|} \sum_{x_i \in \mathcal{H}} \pi_{\text{seg}(\mathcal{A})}(x_i), \quad b^* = \frac{1}{|\mathcal{S}|} \sum_{x_i \in \mathcal{S}} \pi_{\text{seg}(\mathcal{B})}(x_i), \tag{7}$$

where $\mathcal{H}$ and $\mathcal{S}$ denote the sets of hem and sleeve landmarks, respectively. $\pi_{\text{seg}(\cdot)}$ is a projection function that maps a garment landmark onto its corresponding segment of the reference path. The body-relative fitting positions are then computed as:

$$r_{\text{hem}} = \frac{\ell_{a_0 \to a^*}}{\ell_{a_0 \to a_m}}, \quad r_{\text{sleeve}} = \frac{\ell_{b_0 \to b^*}}{\ell_{b_0 \to b_2}}, \tag{8}$$

where $\ell_{u \to v}$ denotes the accumulated length along the reference path from u to v. For HFE, $\mathcal{A}$ is defined as the shoulder–hip path for upper garments, the hip–ankle path for lower garments, and the shoulder–ankle path for dresses. HFE is computed as $|r_{\text{hem}}^{\text{pred}} - r_{\text{hem}}^{\text{gt}}|$. For SFE, the same computation is performed along the left and right shoulder–elbow–wrist paths, and the errors of all valid sides are averaged. All errors are reported as percentages.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

Evaluation Metrics. We evaluate measurement-specific garment fit using Hem Fitting Error (HFE), Sleeve Fitting Error (SFE), and garment-mask Intersection over Union (IoU). HFE and SFE measure hem- and sleeve-length accuracy, respectively, while garment-mask IoU evaluates silhouette consistency. We further assess perceptual image quality using Structural Similarity (SSIM) (Wang et al., 2004) and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018), and distribution-level realism using Fréchet Inception Distance (FID) (Heusel et al., 2017) and Kernel Inception Distance (KID) (Bińkowski et al., 2018).

Implementation Details. Rather than sampling a timestep globally and using it to determine the data source, we first sample a training example and then draw t from a shared SNR-shifted (Esser et al., 2024) timestep distribution restricted to the interval corresponding to its data domain. Synthetic samples use $t \in (t_s, 1]$, while real samples use $t \in (0, t_s]$. We set $t_s = 0.95$. Since real-world data lacks paired try-on triplets, we follow (Chong et al., 2025) to generate pseudo input person images I_b using an editing model (Google, 2025). We adopt FLUX.2 Klein Base 4B (BlackForest, 2025) as the backbone and fine-tune the flow transformer using LoRA (Hu et al., 2022) with rank 64 and scaling factor 128, while keeping the text encoder and autoencoder frozen. All images are resized to 768×1024. We use AdamW (Loshchilov & Hutter, 2017) with a learning rate of 1×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$. The model is trained in bfloat16 precision with a global batch size of 8. The bridge path is activated with probability of 0.7. Inference is performed using a single 28-step trajectory. More details of the implementation are provided in the supplementary material.

5.2 COMPARISON WITH STATE-OF-THE-ART METHODS

We evaluate overall try-on performance on our collected dataset against representative VTON systems, including the general-purpose image editing model Gemini 2.5 Flash Image (Nano Banana) (Google, 2025),the original FLUX.2 Klein Base 4B (BlackForest, 2025) without fit-specific

Table 1: Quantitative comparison on our measurement-annotated VTON benchmark. HFE and SFE denote the body-relative hem and sleeve fitting errors, respectively. KID is reported in units of 10^{-3}. Best and second-best results are highlighted in bold and underlined, respectively.

	Visual Quality				Fit Consistency		
Method	FID ↓	KID ($\times 10^{-3}$) ↓	SSIM ↑	LPIPS ↓	HFE (%) ↓	SFE (%) ↓	IoU ↑
Nano Banana (Google, 2025)	15.8612	0.981	0.7919	0.1582	22.82	15.61	0.7936
Any2AnyTryon (Guo et al., 2025)	17.7583	2.306	0.8287	0.1536	11.53	7.79	0.7643
CatVTON (Chong et al., 2025)	21.8222	4.213	0.7904	0.1804	11.02	<u>6.80</u>	0.7866
OmniTry (Feng et al., 2026)	15.8087	**0.189**	0.8253	0.1478	13.04	7.99	0.7513
RefTON (Li et al., 2026)	<u>14.8936</u>	0.752	<u>0.8367</u>	0.1280	9.89	7.17	<u>0.8050</u>
FitVTON (Ning et al., 2026)	15.8709	1.033	0.8387	0.1393	<u>8.28</u>	8.05	0.7837
FLUX.2 Klein Base 4B (BlackForest, 2025)	17.4184	1.599	0.8174	0.1652	11.43	9.03	0.7658
FitFlow (Ours)	**13.9300**	<u>0.516</u>	**0.8526**	**0.1152**	**5.54**	**5.86**	**0.8423**

Table 2: Ablation study of the proposed decoupled training strategy, bridge path, GSE, and training data sources.KID is reported in units of 10^{-3}. Best and second-best results are highlighted in bold and underlined, respectively.

	Components					Visual Quality				Fit Consistency		
Method	Decoupled	Bridge	GSE	Syn.	Real	FID ↓	KID ($\times 10^{-3}$)↓	SSIM ↑	LPIPS ↓	HFE(%) ↓	SFE(%) ↓	IoU ↑
Synthetic-Only	✗	✗	✓	✓	✗	19.8070	3.481	0.8322	0.1535	11.13	8.32	0.7707
Real-Only	✗	✗	✓	✗	✓	15.3340	1.092	<u>0.8497</u>	0.1308	8.71	6.39	0.7922
Synthetic-Real Mixed	✗	✗	✓	✓	✓	14.8540	<u>0.665</u>	0.8438	0.1232	9.21	6.98	0.8027
FitFlow w/o Bridge	✓	✗	✓	✓	✓	<u>14.3940</u>	0.753	0.8487	<u>0.1224</u>	<u>7.61</u>	<u>6.24</u>	<u>0.8195</u>
FitFlow w/o GSE	✓	✓	✗	✓	✓	16.5270	1.597	0.8312	0.1297	9.90	8.17	0.7825
FitFlow (Ours)	✓	✓	✓	✓	✓	**13.9300**	**0.516**	**0.8526**	**0.1152**	**5.54**	**5.86**	**0.8423**



Figure 5: Qualitative comparisons on our curated real-world benchmark. The left column shows inputs, followed by baselines, our method, and GT. Compared to existing methods, FitFlow produces more accurate global fit (e.g., hem/sleeve length) and photorealistic local textures, especially under challenging size mismatches.

training, and open-source VTON methods such as Any2AnyTryon (Guo et al., 2025), CatVTON (Chong et al., 2025), OmniTry (Feng et al., 2026), RefTON (Li et al., 2026), and FitVTON (Ning et al., 2026). As shown in Tab. 1, FitFlow achieves the best performance on six of the seven evaluation metrics, including FID, SSIM, LPIPS, HFE, SFE, and garment-mask IoU. These results demonstrate that our method not only preserves high perceptual quality but also produces garment layouts that are substantially more consistent with the target measurements. In particular, FitFlow reduces HFE from 8.28% to 5.54% compared with another fit-aware method, FitVTON (Ning et al., 2026). It also improves garment-mask IoU from 0.7837 to 0.8423, indicating more accurate garment coverage and spatial alignment. Although OmniTry (Feng et al., 2026) obtains the lowest KID, its considerably higher HFE, SFE, and lower IoU suggest that image-distribution similarity alone does not guarantee accurate fitting effects. Fig. 5 provides qualitative comparisons of fitting



Figure 6: Continuous garment-length control. Compared to the real-only baseline, FitFlow achieves consistent length and fit control across diverse garments, enabled by multi-size synthetic data that explicitly exposes the model to size-aware variations.

accuracy and continuous length control. Existing methods often generate incorrect hem positions, particularly for dresses and pants, resulting in garment lengths and body coverage that deviate from the target measurements.

5.3 Ablation Studies

Effect of Real and Synthetic Data. We first investigate the respective contributions of real and synthetic training data by comparing models trained with synthetic-only, real-only, and mixed data. We find that the model can learn a certain degree of fitting control from real data. However, garment measurements in real-world images are entangled with diverse poses, wearing styles, and garment constructions, making it difficult to provide stable measurement-dependent supervision. As shown in Fig. 6, using real-only model changes the shorts into a skirt-like garment when the target length is increased by 20 cm, rather than consistently extending the original garment structure. In contrast, by decoupling the two domains across different timestep ranges, more consistent garment-length changes are demonstrated. Besides, the involvement of real data in the low-noise interval substantially improves visual quality, resulting in an FID reduction from 19.807 to 14.854, as shown in Tab. 2, which demonstrates the effectiveness of this sim-to-real framework.

Effect of Timestep Decoupling. To isolate the effect of timestep decoupling, we compare full-timestep mixing (*Synthetic-Real Mixed*) with decoupled training (*FitFlow w/o Bridge*). As shown in Table 2, decoupling the data domains improves fitting accuracy (HFE drops from 9.21% to 7.61%, SFE from 6.98% to 6.24%) and garment-mask IoU (0.8027 to 0.8195), confirming that time-partitioned supervision better aligns synthetic structural priors with real texture refinement.

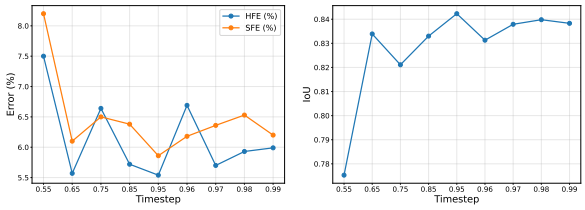

Figure 7: Ablation on boundary timestep t_s. Optimal fit accuracy (HFE/SFE/IoU) are achieved around $t_s = 0.95$.

We examine the choice of the boundary timestep t_s in Fig. 7. Under the shared SNR-shifted schedule (Esser et al., 2024), the model achieves robust performance across a wide high-noise interval ($t_s \in [0.65, 0.99]$), reaching optimal fit accuracy and mask IoU at $t_s = 0.95$.

Notably, even when synthetic supervision covers only the earliest noise steps, the model maintains fit control. This reveals a key insight: macro body–garment layout is largely determined during initial trajectory steps, where concise synthetic guidance suffices to inject a strong measurement-aware structural prior. Conversely, setting t_s too low (e.g., $t_s = 0.55$) curtails the real-domain trajectory interval, leaving insufficient steps to refine textures and correct local boundary errors.

Table 3: Comparison of fit-control strategies on the FIT benchmark. All conditioning strategies are reproduced under the same FLUX.2 Klein Base 4B (BlackForest, 2025) backbone.

Method	HFE (%) ↓	SFE (%) ↓	IoU ↑
FIT (Karras et al., 2026)	8.57	8.34	0.864
FitVTON (Ning et al., 2026)	8.90	8.07	0.863
FitFlow-MLP	8.76	8.98	0.858
FitFlow (Ours)	**5.13**	**5.87**	**0.904**

Effect of the Structural Bridge. We further evaluate the structural bridge by comparing FitFlow against a variant using timestep decoupling alone (*FitFlow w/o Bridge*). As shown in Table 2, incorporating the bridge boosts both fitting accuracy and visual quality, reducing HFE from 7.61% to 5.54% and FID from 14.394 to 13.930, while improving garment-mask IoU to 0.8423.



Figure 8: Comparison of alternative training strategies.

Although timestep decoupling separates structural learning from appearance refinement, it creates an interface mismatch between the two stages. Without the bridge, the late-stage vector field is trained solely on noisy states from real targets, but during inference, it receives boundary states generated by the synthetic stage, causing a severe train–inference trajectory shift. As illustrated in Fig. 8, this shift hampers the model's ability to correct residual layout errors and render sharp garment boundaries. The structural bridge bridges this gap by conditioning the low-noise trajectory on synthetic-predicted boundary states, forcing the network to learn paired trajectory corrections toward real targets.

Effect of Garment Size Encoding. We first evaluate the contribution of GSE within the FitFlow framework by removing it while keeping all other components unchanged. Without GSE, HFE increases from 5.54% to 9.90%, SFE increases from 5.86% to 8.17%, and garment-mask IoU decreases from 0.8423 to 0.7825. These substantial degradations demonstrate that GSE is essential for accurate measurement adherence and garment-layout control.

We further evaluate different fit control strategies on the FIT benchmark (Karras et al., 2026). We compare our method with an MLP-based measurement embedding following FIT (Karras et al., 2026) and prompt-based conditioning following FitVTON (Ning et al., 2026). For a controlled comparison, all methods use FLUX.2 Klein Base 4B (BlackForest, 2025) as their generation backbone. The MLP-based strategy requires the model to infer garment extent from abstract numerical features, while prompt-based control provides only coarse semantic guidance. In contrast, GSE directly converts the garment-to-body measurement ratio into an explicit image-space garment scale, reducing the gap between physical measurements and spatial layout. As shown in Tab. 3, this explicit visual conditioning yields more accurate measurement adherence and garment layout.

6 Conclusion

We present **FitFlow**, a decoupled flow-matching framework for measurement-guided virtual try-on. FitFlow assigns synthetic fit supervision to high-noise timesteps and real appearance supervision to low-noise timesteps within a single generation trajectory. A structural bridge connects the two learning regimes, while Garment Size Encoding enables continuous and spatially precise fit control. We further construct a measurement-annotated dataset and a body-relative fit evaluation protocol. Extensive experiments demonstrate that FitFlow consistently improves fitting accuracy, controllability, and visual quality through unified structural and appearance learning along the flow trajectory. This design provides a practical foundation for precise, measurement-responsive, and visually realistic virtual try-on across diverse body and garment conditions in real-world deployment scenarios.

AI USE STATEMENT

Generative AI tools were used to assist with the writing and language editing of this paper, including improving clarity, grammar, and readability. The authors reviewed and revised all AI-assisted text and take full responsibility for the final content of the paper.

ETHICS STATEMENT

Our real-world data are collected from publicly accessible fashion and e-commerce webpages and are used solely for research purposes. As the collected images may contain identifiable individuals, we crop facial regions before training and evaluation to reduce identity cues and mitigate privacy concerns. The proposed dataset and evaluation protocol are intended for research on virtual try-on and garment-fit modeling rather than for identifying or profiling individuals. Any release of data or annotations will respect applicable licenses, terms of use, and redistribution restrictions of the original data sources.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed FitFlow framework, training objectives, and Garment Size Encoding in the main paper. The experimental settings, including the backbone, optimization configuration, timestep sampling strategy, and inference procedure, are reported in Sec. 5.1. Additional details on real- and synthetic-data construction are provided in Appendix A, the HFE and SFE evaluation protocol in Appendix B, and auxiliary-model training, structural-bridge training, and controlled baseline reproductions in Appendix C. These details are provided to facilitate reproduction of the proposed method and evaluation protocol.

REFERENCES

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

BlackForest. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

D. Bojanić. Smpl-anthropometry. https://github.com/DavidBoja/SMPL-Anthropometry, 2024.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7513–7522, 2023.

Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14131–14140, 2021.

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.

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 *Forty-first international conference on machine learning*, 2024.

Yutong Feng, Linlin Zhang, Hengyuan Cao, Yiming Chen, Xiaoduan Feng, Jian Cao, Yuxiong Wu, and Bin Wang. Omnitry: Virtual try-on anything without masks. *Advances in Neural Information Processing Systems*, 38:132640–132667, 2026.

Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5337–5345, 2019.

Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8485–8493, 2021.

Google. Introducing gemini 2.5 flash image, our state-of-the-art image model. https://developers.googleblog.com/introducing-gemini-2-5-flash-image/, August 2025. Accessed: 2026-07-26.

Hailong Guo, Bohan Zeng, Yiren Song, Wentao Zhang, Jiaming Liu, and Chuang Zhang. Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19085–19096, 2025.

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.

Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7543–7552, 2018.

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. *Advances in neural information processing systems*, 30, 2017.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Johanna Karras, Yuanhao Wang, Yingwei Li, and Ira Kemelmacher-Shlizerman. Fit: A large-scale dataset for fit-aware virtual try-on. *arXiv preprint arXiv:2604.08526*, 2026.

Jeongho Kim, Guojung 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.

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.

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

Liuzhuozheng Li, Yue Gong, Shanyuan Liu, Zanyi Wang, Dengyang Jiang, Leibucha Wu, Bo Cheng, Yuhang Ma, Dawei Leng, and Yuhui Yin. Refton: Reference person shot assist virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14915–14925, 2026.

Ente Lin, Xujie Zhang, Fuwei Zhao, Yuxuan Luo, Xin Dong, Long Zeng, and Xiaodan Liang. Dreamfit: Garment-centric human generation via a lightweight anything-dressing encoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5218–5226, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Miles Macklin. Warp: A high-performance python framework for gpu simulation and graphics. `https://github.com/nvidia/warp`, 2022. NVIDIA GPU Technology Conference (GTC).

Miles Macklin, Matthias Müller, and Nuttapong Chentanez. Xpbd: position-based simulation of compliant constrained dynamics. In *Proceedings of the 9th International Conference on Motion in Games*, pp. 49–54, 2016.

MMPose. 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2231–2235, 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.

Yiqun Ning, Ao Shen, Chenhang He, and Lei Zhang. Fitvton: Fit-aware virtual try-on via body-garment size control. *arXiv preprint arXiv:2606.12012*, 2026.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10975–10985, 2019.

Qwen. Qwen3 technical report, 2025. URL `https://arxiv.org/abs/2505.09388`.

Siqi Wan, Yehao Li, Jingwen Chen, Yingwei Pan, Ting Yao, Yang Cao, and Tao Mei. Improving virtual try-on with garment-focused diffusion models. In *European Conference on Computer Vision*, pp. 184–199. Springer, 2024.

Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 589–604, 2018.

Haoyu Wang, Zhilu Zhang, Donglin Di, Shiliang Zhang, and Wangmeng Zuo. Mv-vton: Multi-view virtual try-on with diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7682–7690, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23550–23559, 2023.

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*, 2025a.

Zhaotong Yang, Zicheng Jiang, Xinzhe Li, Huiyu Zhou, Junyu Dong, Huaidong Zhang, and Yong Du. D 4-vton: Dynamic semantics disentangling for differential diffusion based virtual try-on. In *European Conference on Computer Vision*, pp. 36–52. Springer, 2024.

Zhaotong Yang, Yuhui Li, Shengfeng He, Xinzhe Li, Yangyang Xu, Junyu Dong, and Yong Du. Omnivton: Training-free universal virtual try-on. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16702–16711, 2025b.

Jinliang Yao and Haonan Zheng. Lc-vton: Length controllable virtual try-on network. *IEEE Access*, 11:88451–88461, 2023.

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.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26399–26408, 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4606–4615, 2023.

Appendix

In this appendix, we provide supplementary materials for **FitFlow**. Specifically, we present:

A. Details of the data construction pipeline, including real-data curation, measurement annotation, and synthetic-data generation through body-specific garment construction and cloth simulation.

B. Details of the body-relative fit evaluation protocol, including the qualification criteria for HFE and SFE,and a discussion of potential extensions to 3D fit evaluation.

C. Additional implementation details, including auxiliary model training, training with the structural bridge, and controlled reproduction of fit-conditioning baselines.

D. Additional qualitative results on our curated real-world benchmark, continuous garment-length control, and the FIT benchmark.

A Details of the Data pipeline

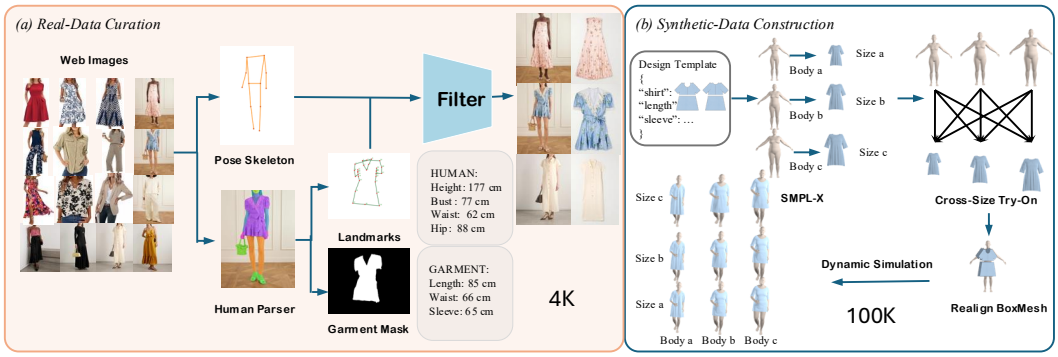

Figure 9: Overview of our data construction pipeline. (a) Real-data curation through pose, landmark, and garment-mask extraction, followed by quality filtering and measurement annotation. (b) Synthetic-data construction through size-specific garment generation, cross-size outfit combination, BoxMesh realignment, and dynamic cloth simulation.

Real-Data Curation. As illustrated in Fig 9(a), we collect fashion images from open e-commerce platforms, including Mytheresa[1], SSENSE[2], Net-a-Porter[3], and Universal Standard[4]. We first extract human pose skeletons using MMPose (MMPose, 2020), and garment landmarks and masks using DeepFashion2 (Ge et al., 2019). Samples with failed or low-quality pose estimation, garment segmentation, or landmark detection are removed.

Each retained sample is required to provide four body measurements: height, bust, waist, and hip circumference. For garment measurements, we manually verify and annotate the most relevant attributes for each category, including garment length and, when available, bust, waist, hip, and sleeve length. These measurements jointly describe the body–garment size relationship and support measurement-guided fit supervision.

After filtering and annotation, the final real-world dataset contains 4,000 high-quality samples with body and garment measurements. To reduce identity cues and protect personal privacy, we crop the facial regions before training and evaluation. The dataset is divided into mutually exclusive training and test sets, each containing 2,000 samples: 600 upper-garment, 700 lower-garment, and 700 dress samples.

[1] https://www.mytheresa.com/
[2] https://www.ssense.com/
[3] https://www.net-a-porter.com/
[4] https://www.universalstandard.com/

Synthetic-Data Construction. As shown in Fig 9(b), we use GarmentCode (Korosteleva & Sorkine-Hornung, 2023) to generate sewing patterns and garment specifications, and NVIDIA Warp (Macklin, 2022) to perform XPBD-based cloth simulation (Macklin et al., 2016). We follow the dynamic simulation pipeline of FitVTON (Ning et al., 2026), with additional modifications to support stable and size-consistent garment generation.

Specifically, we use SMPL-Anthropometry (Bojanić, 2024) to extract body measurements from SMPL-X (Pavlakos et al., 2019) meshes and optimize the SMPL-X shape parameters to match the target height, bust, waist, and hip measurements. Based on the resulting measurements, we generate a body-specific GarmentCode specification for each target size, rather than uniformly scaling a fixed body mesh or garment pattern. We further adopt category-specific anatomical anchors to improve garment initialization and reduce simulation artifacts.

This pipeline produces synthetic data showing the same garment worn across different body shapes with multiple sizes. The synthetic dataset contains 19 garment templates (3 uppers, 4 pants, 4 skirts, and 8 dresses), each instantiated in five randomly sampled sizes. Combining upper and lower garments and rendering dresses independently produces 640 outfits for each body–pose pair, yielding 102,400 images across 16 body models and 10 poses. During training, triplets are randomly sampled, with each size-specific garment used once as the target for every body–pose pair, resulting in 15,200 triplets per epoch.

B Details of Body-Relative Fit Evaluation

This section elaborates on the body-relative fit evaluation protocol described in the main paper. Both HFE and SFE rely on human body keypoints and garment landmarks, using category-specific structural cues for hem and sleeve evaluation, respectively. All ground-truth samples are pre-screened to ensure that the required structural cues can be reliably detected. For generated images, if the required garment landmarks or body keypoints cannot be reliably detected, we treat the case as an evaluation failure and assign the maximum fitting error of 100%, rather than excluding the sample.

HFE Qualification. For each hem sample, the required body keypoints and at least one hem landmark are used to estimate the garment fitting position. When multiple hem landmarks are available, their projected fitting positions are averaged to reduce localization noise. If the required structural cues cannot be reliably detected in the generated image, the sample is assigned the maximum HFE of 100%.

SFE Qualification. For each sleeve sample, the cuff landmark together with the corresponding shoulder, elbow, and wrist keypoints are used to estimate the sleeve fitting position. The left and right sleeves are evaluated independently. If the required structural cues cannot be reliably detected in the generated image, or if the generated sleeve structure does not provide a valid cuff corresponding to the ground truth, the sample is treated as a fitting failure and assigned the maximum SFE of 100%.

Extension to 3D Fit Evaluation. Our current HFE and SFE are defined in a 2D body-relative coordinate system to provide a robust and directly measurable evaluation of longitudinal garment fit. While 3D body reconstruction methods such as SMPL-X (Pavlakos et al., 2019) could provide a more complete geometric reference, reliable 3D garment reconstruction from a single generated image remains challenging, particularly for loose garments, self-occlusion, and non-rigid draping. Using such estimates directly for evaluation may therefore introduce additional reconstruction bias that is unrelated to the quality of the try-on model itself. A promising direction is to extend the current framework to 3D by jointly recovering the body and garment surfaces and measuring region-wise garment–body clearance, enabling evaluation of tightness, ease, and draping in addition to longitudinal fit.

C Additional Implementation Details

Auxiliary Model Training. The auxiliary velocity model v_θ^{syn} is trained exclusively on the synthetic dataset described above, using 15,200 dynamically sampled triplets per epoch. We initialize it from FLUX.2 Klein Base 4B (BlackForest, 2025) and fine-tune it with LoRA (Hu et al., 2022) using

rank 64 and scaling factor 128. The model is trained for 5 epochs over the full timestep range using the flow-matching MSE objective in bfloat16 precision. We use AdamW (Loshchilov & Hutter, 2017) with a learning rate of 1×10^{-4}, a constant learning-rate schedule, and a batch size of 1 per GPU on 8 GPUs. Garment conditions are resized using GSE with a canonical body height of 175 cm and a physical resolution of 0.20 cm/pixel, together with random scaling and translation augmentations. After training, v_θ^{syn} is frozen and used to generate the synthetic-guided predictions required by the structural bridge. Following the same 28-step sampling schedule used during inference, we perform complete sampling from $t = 1$ to $t = 0$ for each real condition $c^{\text{real}} = [I_p^{\text{real}}, I_g^{\text{real}}]$, obtain the clean prediction $\hat{I}_y$, encode it into $\hat{z}_0$, and cache the result for subsequent FitFlow training.

Training with the Structural Bridge. We set the switching timestep to $t_s = 0.95$. Each epoch uses 15,200 dynamically sampled synthetic triplets and 2,000 real training samples. After sampling a training example, we draw t from a shared SNR-shifted (Esser et al., 2024) timestep distribution conditioned on the interval corresponding to its data domain. Specifically, synthetic samples use $t \in (t_s, 1]$, while real samples use $t \in (0, t_s]$.

For synthetic samples, GSE is used to encode the body-relative garment scale, together with the same random scaling and translation augmentations adopted during auxiliary-model pretraining. For real samples in the low-noise interval, we stochastically activate the structural-bridge and standard real-data paths with probabilities 0.7 and 0.3, respectively. In this low-noise interval, the original garment image is used without GSE, since excessive resizing may reduce fine-grained garment details, while structural guidance has already been introduced through the synthetic-guided bridge anchor.

The model is trained for 5 epochs using the same backbone, LoRA configuration, optimizer, learning rate, precision, and distributed training settings as the auxiliary model. Unless otherwise specified, all ablation variants follow the same training configuration as FitFlow, with only the corresponding component or training strategy modified.

Controlled Comparison and Baseline Reproductions. To evaluate GSE against alternative fit-conditioning strategies, we conduct controlled comparisons on the FIT benchmark (Karras et al., 2026) using a shared 20,000-sample training subset and a clean 2,000-sample test split, evaluated via HFE, SFE, and garment-mask IoU. For a fair comparison, all baselines, including numerical-conditioning FIT (Karras et al., 2026), prompt-based FitVTON (Ning et al., 2026), and our GSE-only variant—are reproduced on the same FLUX.2 Klein Base 4B backbone with identical optimization protocols: 5 epochs of AdamW training at 1×10^{-4} learning rate, batch size of 8, resolution 768×1024, and rank-64/scale-128 LoRA adapters applied to the frozen backbone.

Specifically, **FIT** encodes seven normalized numerical measurements via multi-frequency Fourier features and an MLP into a 3,072-dimensional token, whereas **FitVTON** converts body and garment metrics into discrete textual attributes (*short/tall*, *slim/fat*, *XS–3XL*) formatted by a natural-language template and encoded via a frozen Qwen3-4B-FP8 text encoder (Qwen, 2025). Our GSE-only variant trains over the full timestep range without timestep decoupling or the structural bridge, ensuring that performance gains stem strictly from the conditioning representation.

D Additional Qualitative Results

We provide qualitative comparisons on our curated real-world benchmark, continuous garment-length control results, and comparisons on the FIT benchmark. As shown in Fig. 10, FitFlow better preserves garment fit, including hem and sleeve length, while producing more realistic textures, particularly under challenging body–garment size mismatches. Fig. 11 further compares FitFlow with the real-only baseline under continuous garment-length control from -20 cm to $+20$ cm. FitFlow produces more gradual and consistent changes in hem position while better preserving garment appearance and person identity. Finally, Fig. 12 compares FIT, FitVTON, and FitFlow on the FIT benchmark, where FitFlow more accurately follows the target body-relative garment scale and produces more consistent garment coverage across diverse body shapes and garment sizes.

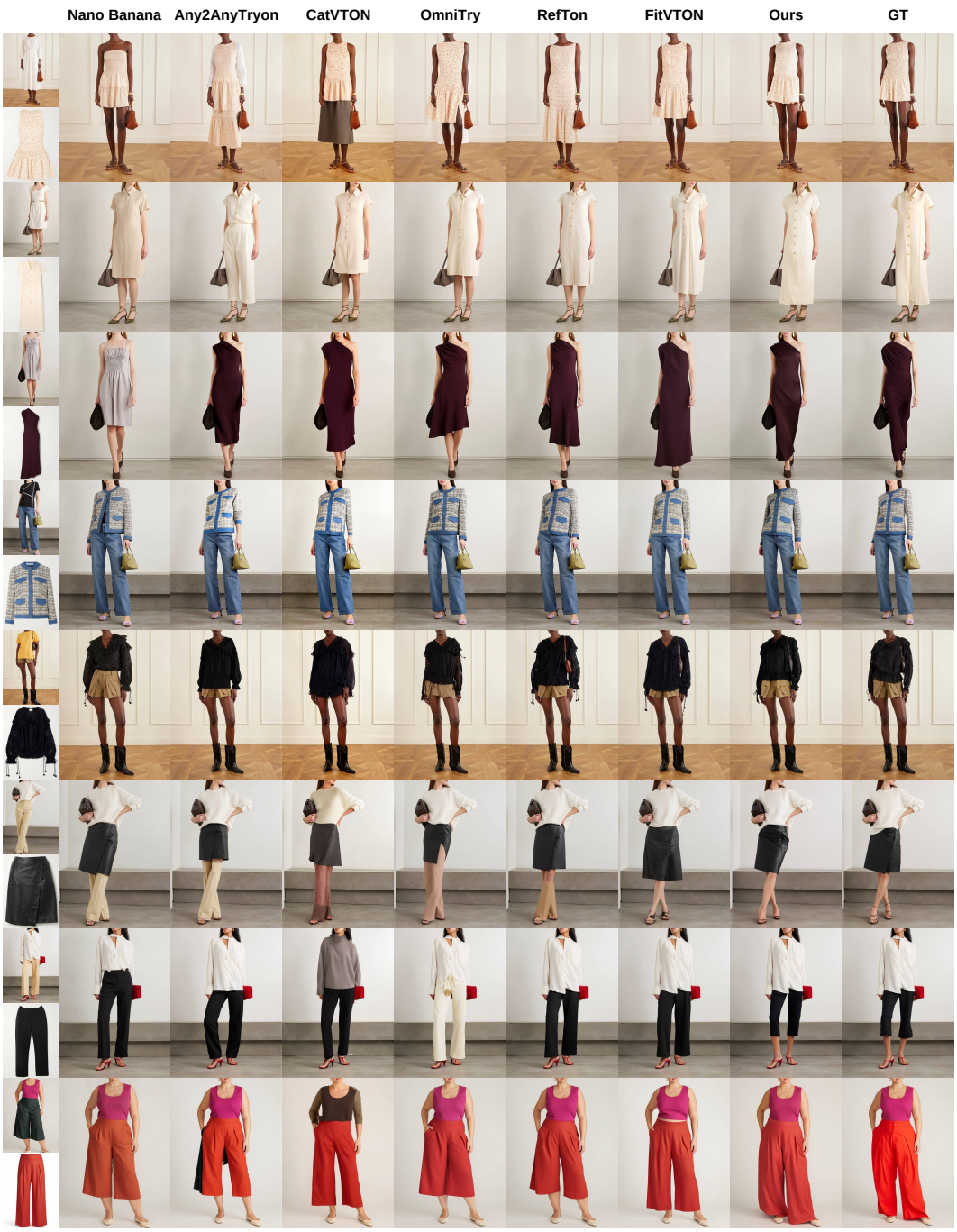

Figure 10: Qualitative comparisons on our curated real-world benchmark. From left to right: inputs, baseline methods, FitFlow, and ground truth. FitFlow achieves more accurate global garment fit, including hem and sleeve length, while preserving realistic local textures, especially under challenging size mismatches.

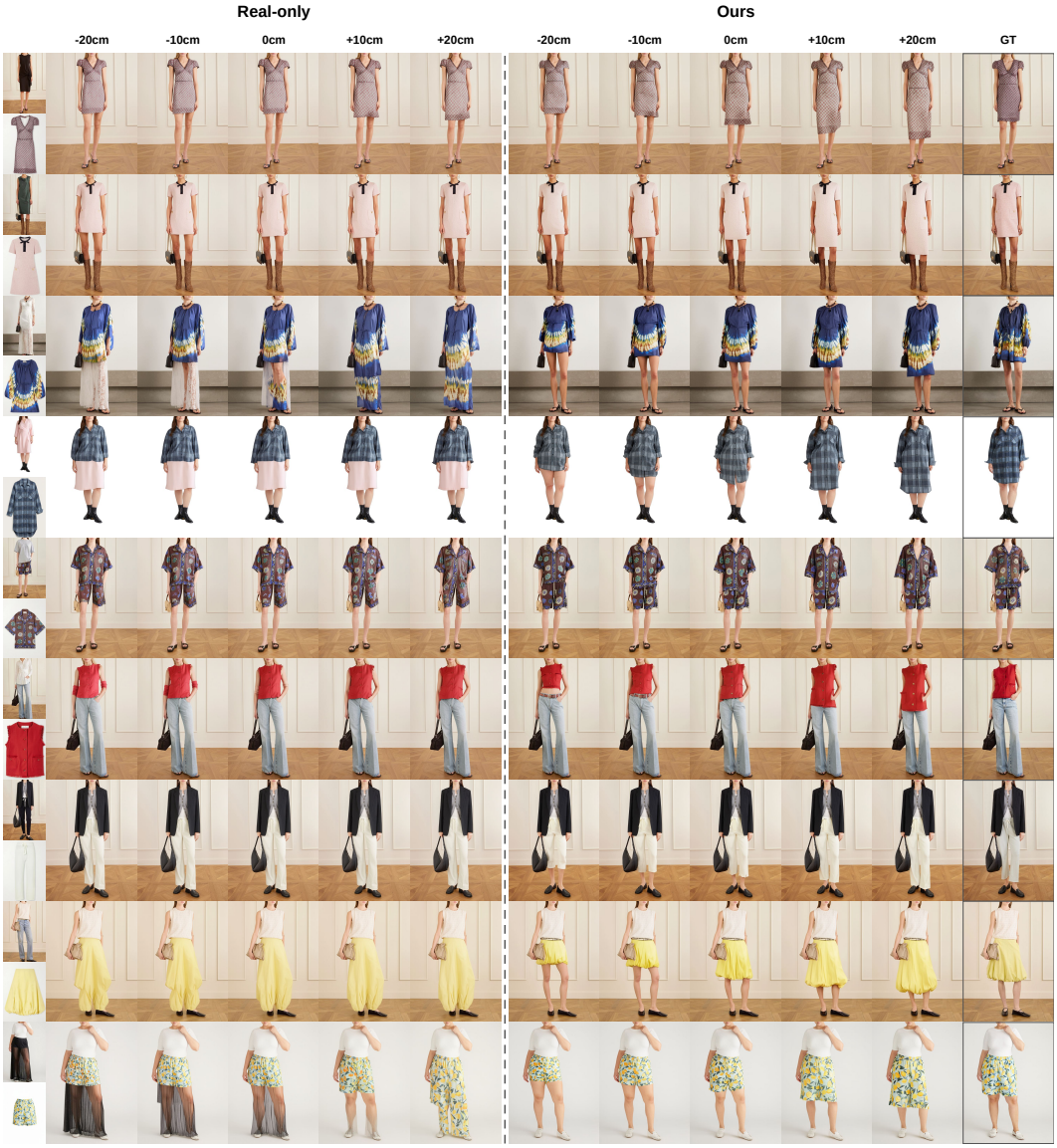

Figure 11: Qualitative comparison of continuous garment-length control from -20 cm to $+20$ cm. Compared with the real-only baseline, FitFlow produces more gradual and measurement-responsive changes in garment length while better preserving garment appearance, person identity, and overall image quality.



Figure 12: Qualitative comparison on the FIT benchmark among FIT, FitVTON, and FitFlow. Compared with existing fit-control methods, FitFlow produces garment layouts that more closely follow the target body–garment measurements while preserving garment appearance and realistic details.