FITVTON: FIT-AWARE VIRTUAL TRY-ON VIA BODY-GARMENT SIZE CONTROL

Anonymous authors
Paper under double-blind review



Figure 1: With Garment-Body Size prompts, Nano Banana (Google, 2025) produces "neutral fit" results across different body shapes. In contrast, FitVTON produces more faithful fit results. (Zoom in to see how the hem position, tightness and the elasticity of the cuffs are varied across different bodies.)

ABSTRACT

While diffusion-based virtual try-on has achieved impressive visual realism, most methods treat the task as 2D inpainting, prioritizing texture preservation over physical plausibility. Consequently, they often produce plausible-looking images that fail to reflect authentic garment fit across diverse body shapes. We present FitVTON, a Fit-aware virtual try-on model on different bodies in the wild. FitVTON encodes Garment-Body Size, a pair of measured body and garment-size levels, through structured text prompts, and learns from simulated try-on triplets that drape five garment sizes on diverse parametric bodies. To improve the fitting effects over garment silhouettes, we introduce two auxiliary heads to predict the masks for both the garment and the exposed body. We further introduce a texture rectification stage to improve realistic appearance from simulated data. To evaluate the fitting fidelity, we curate a real-world dataset, *FittingEffect3K*, combining VLM-based scoring protocol. Both subjective and quantitative experiments show that FitVTON demonstrates authentic fitting fidelity, with significant sizing accuracy and shape preservation over state-of-the-art methods while maintaining competitive image quality. Code and data will be released upon acceptance.

1 INTRODUCTION

Virtual try-on (VTON) visualizes how a garment would look on a person, serving as a critical technology for e-commerce and content creation (Han et al., 2018; Wang et al., 2018; Yang et al., 2020; Lewis et al., 2021). Powered by diffusion models, modern VTON systems can synthesize compelling results with sharp textures and plausible lighting (Zhu et al., 2023; Choi et al., 2024; Yang et al., 2024a; Li et al., 2023). Yet, visual realism does not guarantee physical authenticity: outputs may look convincing while failing to reflect how garments drape and fit across body shapes.

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

Figure 1 illustrates this gap. Even when Nano Banana (Google, 2025) is given explicit garment and body size prompts, it produces "neutral fit" try-on results across different bodies. This suggests that fitting fidelity cannot be obtained from language prompts alone; models need supervision that captures how garments physically deform with body shape.

We identify two barriers to authentic fit. First, public datasets are skewed toward professional models with tall, slim figures (Choi et al., 2021; Morelli et al., 2022; Ge et al., 2019), so models often generalize poorly to diverse bodies and default to an averaged fit that ignores individual characteristics. Second, current VTON datasets such as VITON-HD (Choi et al., 2021) and DressCode (Morelli et al., 2022) lack explicit *(human, garment, target try-on)* triplets with controllable garment-body fit variation. As a result, most VTON approaches formulate the task as clothing-region inpainting (Choi et al., 2024; Yang et al., 2024a; Kim et al., 2024; Wan et al., 2025; Li et al., 2024). Although masks are effective for texture transfer and local realism, they can remove body-shape cues around fit-critical regions and let the prescribed mask boundary act as an implicit shape prior. The final result is therefore sensitive to mask quality and may follow the inpainting region rather than learning body-specific garment deformation.

We introduce **FitVTON**, a fit-aware virtual try-on model that generates more authentic fitting effects based on Garment-Body Size control. The model fine-tunes a *FLUX.1 Kontext* (Labs et al., 2025) Image Editing model with simulated garment-fitting data, which encode cloth deformations across different body shapes and poses. As shown in Figure 1, by encoding the garment size and body levels as structured text prompts (e.g., an *"M-size skirt"* on a *"tall, plus-size body"*), FitVTON produces try-on results that adapt more faithful garment cut, drape, and tightness under different body shapes and poses.

To generate diverse garment fitting data, we use GarmentCode (Korosteleva & Sorkine-Hornung, 2023) as our data source and SMPL-X (Pavlakos et al., 2019; Loper et al., 2023) serves as our fitting mannequin. Both of them are parametric models, which allow us to control both the garment size and the body shape of the simulated data. We develop a multi-step simulation tool based on NVIDIA Warp (Macklin, 2022) that can generate garment fitting data with different garment sizes, body shapes, poses, and wearing styles, including one-piece garments, tucked-in/untucked upper outfits, and lower outfits.

To help the model learn fitting dynamics from simulated data, we introduce two innovations. First, we propose Fit-Grounded Dual-Branch Mask Supervision, which uses training-only auxiliary heads to predict both garment and exposed-body masks. With segmentation losses, these heads spatially ground the generated outputs to fit-sensitive silhouettes (e.g., expanded garment masks for an "oversized" cloth). Second, to bridge the sim-to-real domain gap, we introduce a texture rectification stage by fine-tuning the model on an unpaired real-world dataset. This enhances texture transferability while preserving the model's fitting capability.

Finally, assessing the semantic and geometric accuracy of garment fit remains a significant challenge, as standard perceptual metrics (e.g., FID (Heusel et al., 2017), KID (Bińkowski et al., 2018), LPIPS (Zhang et al., 2018)) prioritize global image fidelity over sizing correctness. To address this, we curate a specialized real-world benchmark, *FittingEffect3K*, captured with 10 human participants with diverse body shapes, poses and cloth types, which produce ground-truth for direct evaluation of fitting accuracy. We design a multi-dimensional evaluation protocol that integrates both VLM-based scoring and subjective human preference. Empirically, we observe strong concordance between the VLM assessments and human judgments across various baselines. Both metrics consistently demonstrate that FitVTON significantly outperforms state-of-the-art methods, particularly in achieving accurate sizing and faithful shape preservation. In summary, our contributions are as follows:

- **Controllable Simulation Data Pipeline.** We generate large-scale, aligned try-on triplets with diverse garment sizes, body shapes, poses, and wearing styles.
- **Fit-aware Flow Matching Model.** We fine-tune a FLUX-based model with dual-branch mask supervision and texture rectification for controllable and realistic fitting across various Garment-Body Sizes.
- **Fit Evaluation Dataset and Protocol.** We introduce the real-world *FittingEffect3K* benchmark and evaluate fitting accuracy with VLM scoring and human assessment.

2 Related Work

2.1 Image-based Virtual Try-On

Image-based virtual try-on (VTON) aims to synthesize a person wearing a target garment while preserving both garment identity and human appearance. Early methods (Han et al., 2018; Wang et al., 2018; Ge et al., 2021; Xie et al., 2023) combine geometric warping with texture synthesis; later pipelines add appearance-flow and local-global constraints. More recently, diffusion-based VTON methods (Zhu et al., 2023; Kim et al., 2024; Xu et al., 2025; Kim et al., 2025; Jiang et al., 2024; Morelli et al., 2023; Zhang et al., 2024; Wan et al., 2025; Yang et al., 2024b) formulate try-on as mask-guided inpainting on person, garment, and pose cues. However, cloth-agnostic masks may suppress body geometry near fit-critical boundaries (e.g., waistline, hips, sleeves, and hems). Recent mask-free VTON approaches increasingly build on large text-to-image backbones such as *FLUX.1* (Black Forest Labs, 2024) and remove explicit mask requirements at inference time (Zhang et al., 2025; Guo et al., 2025; Feng et al., 2025; Wang et al., 2025). These methods remain largely appearance-driven, and language is not explicitly aligned with fit semantics or body-garment interactions. FitVTON fine-tunes a FLUX-based model with simulation data that explicitly captures garment deformation under different body shapes and poses, enabling text prompts to control fit-related changes such as tightness, drape, and silhouette.

2.2 Physics-Based Garment Simulation

Physics-based cloth simulation provides a principled way to model garment deformation and body-garment interactions, and has started to be used in virtual fitting. Commercial systems (CLO3D, 2022; Style3D, 2022) achieve high fidelity but require expert-intensive garment authoring, limiting scalability. Recent work on procedural pattern representations, such as GarmentCode (Korosteleva & Sorkine-Hornung, 2023; Korosteleva et al., 2024), significantly reduces garment creation cost and enables large-scale synthesis, with further extensions enriching garment diversity and semantic control (Bian et al., 2025; Zhou et al., 2025). Nonetheless, simulation pipelines are typically not optimized for photorealistic rendering and struggle to cover the appearance diversity needed for image-based VTON. Building on these efforts, we extend GarmentCode-based dynamic simulation to generate controllable, fitting-aware supervision across body shapes, and use it to train an image-based try-on model.

3 Method

3.1 Garment-Body Size and Simulation

The size tag of a garment is an unreliable description of its fit: the same tag, such as XL, corresponds to different measurements across brands, garment categories, and regional standards. We therefore describe fit with *Garment-Body Size*, a pair of levels, one for the body and one for the garment, both computed from physical measurements. On the body side, we construct 16 SMPL-X (Pavlakos et al., 2019; Loper et al., 2023) body prototypes per gender by combining four height levels (*short*, *medium-short*, *medium-tall*, *tall*) with four body-size levels (*slim*, *medium-slim*, *medium plus-size*, *plus-size*). On the garment side, we generate every GarmentCode (Korosteleva & Sorkine-Hornung, 2023) design in five sizes, XS, S, M, L, and XL: size M is the pattern fitted to the mean SMPL-X body, and the other sizes are graded from M with standard increments in length and girth. Simulating the five sizes on the 16 prototypes gives 80 garment-body combinations per design, from which the model learns how the same design becomes tighter or looser and ends higher or lower as the garment size and the body change. A real person or garment is placed on this grid by measurement. Each body measurement (height, chest, waist, hips) and each garment measurement (length, chest, waist, hips) is located between the two adjacent prototype or graded-size values and assigned to the nearer one, and the per-measurement levels are aggregated into a single level (Appendix A.5). Since the level is computed from measured dimensions, a garment tagged XL by one brand and L by another receives the same level whenever their measurements agree.

We use NVIDIA Warp (Macklin, 2022) with XPBD-based cloth simulation to simulate the garment on the body. The body first morphs from the mean shape to the target shape and then articulates to the target pose; we subdivide these transitions when needed to reduce cloth-body penetration.

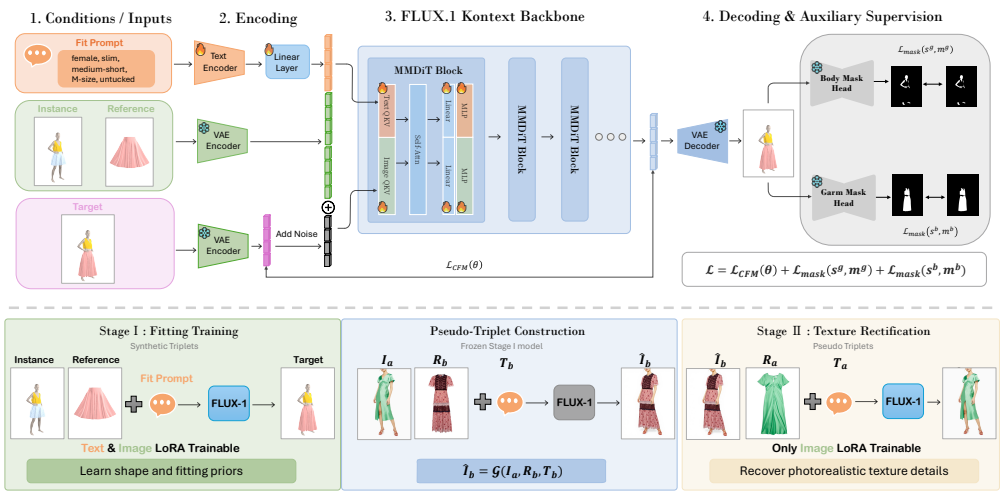

Figure 2: Overview of FitVTON. (**Top**) Given a person image, a reference garment, and a Garment-Body Size prompt, FitVTON performs fit-aware virtual try-on with FLUX.1 Kontext backbone, modality-specific LoRA adapters, and dual-branch garment/body mask supervision. (**Bottom**) Two-stage training strategy: Stage I learns prompt-driven fit priors from synthetic simulations, while Stage II rectifies real-image textures by updating only image LoRA layers.

The resulting simulations provide controlled fit variations across garment sizes and body prototypes, allowing the model to learn how prompt terms such as *slim*, *medium-tall*, *plus-size*, and *XL-size* correspond to garment-body fit. We additionally model the outfit styles within three cases: one-piece, tucked-in, and untucked. For tucked-in/untucked outfits, the simulation uses separate upper and lower garment meshes and controls whether the upper hem lies inside/outside the lower waistband. Besides, we add wrist/ankle stop constraints to prevent sleeve cuffs or pant hems from sliding beyond the hand or shoe region to align with the real-world try-on scenario. The body prototypes, garment size grading, measurement-based mapping, and simulation details are elaborated in Appendix A.

3.2 FitVTON Architecture

As shown in Fig. 2, FitVTON is formulated as fit-aware image editing. The input conditions consist of an instance image I_a, a reference garment image R_b, and a fit prompt $\mathbf{T}_b$ that describes Garment-Body Size and wearing style, e.g., *"The person is a medium-short and slim female. The cloth is an M-size upper garment. The wearing style is untucked."* The model generates a target try-on image I_b that preserves the visual appearance of R_b while adapting the garment silhouette to the body and pose in I_a according to the fit prompt.

We build FitVTON on *FLUX.1 Kontext* (Labs et al., 2025), a flow-matching Multimodal Diffusion Transformer. The instance image, reference garment, and target image are encoded into the latent space by the VAE encoder, while the prompt is encoded by the text encoder. These visual and textual tokens are then processed by the MMDiT backbone, where image and text LoRA adapters are inserted into the modality-specific projections. This design lets the image branch learn garment transfer and appearance preservation, while the text branch learns how Garment-Body Size controls fit-related geometry. Let $\mathcal{E}$ and $\mathcal{D}$ denote the VAE encoder and decoder. For a training triplet (I_a, R_b, I_b), we encode the target try-on image as $\mathbf{x}_1 = \mathcal{E}(I_b)$ and sample a noise latent $\mathbf{x}_0 \sim p_0$. The conditioning signal is denoted by $\mathbf{c}$ and contains the encoded instance image, reference garment, and fit prompt:

$$\mathbf{c} = \{\mathcal{E}(I_a), \mathcal{E}(R_b), \tau(\mathbf{T}_b)\}, \tag{1}$$

where $\tau(\cdot)$ is the text encoder. The flow-matching backbone learns a conditional vector field governed by the ODE:

$$\frac{d\mathbf{x}_t}{dt} = \mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}), \quad t \in [0, 1], \tag{2}$$

where $\mathbf{x}_t$ is the intermediate latent state and $\mathbf{v}_\theta$ predicts its instantaneous velocity. We use the linear probability path between the noise and data endpoints:

$$\mathbf{x}_t = (1 - t)\mathbf{x}_0 + t\mathbf{x}_1, \tag{3}$$

whose target velocity is $\mathbf{u}_t = \mathbf{x}_1 - \mathbf{x}_0$. The conditional flow-matching objective is:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1),\, \mathbf{x}_0\sim p_0,\, \mathbf{x}_1\sim p_{\text{data}}} \left[\|\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}) - \mathbf{u}_t\|_2^2 \right]. \tag{4}$$

At inference, the learned ODE is integrated from a sampled noise latent to obtain $\hat{\mathbf{x}}_1$, which is decoded as $\hat{I}_b = \mathcal{D}(\hat{\mathbf{x}}_1)$. While this objective provides the main image-generation supervision, it does not explicitly constrain whether the fit prompt controls local garment and body boundaries. We address this limitation with the auxiliary supervision below.

3.3 Dual-Branch Mask Supervision

Fit variations are expressed most clearly through local geometry, including waistlines, hems, sleeves, cuffs, garment boundaries, and exposed-body regions. A pure flow-matching objective can reproduce the target appearance, but it provides only indirect supervision for whether each intermediate state follows the intended fit semantics. We therefore introduce Dual-Branch Mask Supervision as an auxiliary geometric regularizer. From the simulated training data, we export garment masks $\mathbf{M}^g$ and exposed-body masks $\mathbf{M}^b$, and use them to train two lightweight U-Net mask heads, $\mathcal{H}^g_\phi$ and $\mathcal{H}^b_\psi$. During flow-matching training, the predicted velocity at timestep t gives a one-step estimate of the endpoint latent:

$$\tilde{\mathbf{x}}_1^{(t)} = \mathbf{x}_t + (1-t)\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}). \tag{5}$$

We decode $\tilde{\mathbf{x}}_1^{(t)}$ to image space and apply the two mask heads to predict timestep-conditioned geometry:

$$\hat{\mathbf{M}}^g_t = \mathcal{H}^g_\phi\left(\mathcal{D}(\tilde{\mathbf{x}}_1^{(t)})\right), \quad \hat{\mathbf{M}}^b_t = \mathcal{H}^b_\psi\left(\mathcal{D}(\tilde{\mathbf{x}}_1^{(t)})\right), \tag{6}$$

where $\mathcal{D}$ denotes the image decoder. Each branch is supervised by a BCE-Dice objective:

$$\mathcal{L}_{\text{seg}}(\hat{\mathbf{M}}, \mathbf{M}) = \mathcal{L}_{\text{BCE}}(\hat{\mathbf{M}}, \mathbf{M}) + \eta\mathcal{L}_{\text{Dice}}(\hat{\mathbf{M}}, \mathbf{M}), \tag{7}$$

where η balances binary cross-entropy and Dice loss (Sudre et al., 2017). The final training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CFM}} + \lambda_g\mathcal{L}_{\text{seg}}(\hat{\mathbf{M}}^g_t, \mathbf{M}^g) + \lambda_b\mathcal{L}_{\text{seg}}(\hat{\mathbf{M}}^b_t, \mathbf{M}^b). \tag{8}$$

The garment branch encourages prompt-consistent clothing silhouettes, while the exposed-body branch preserves body-shape cues that should remain visible under different fit prompts. Since the U-Net heads are used only as training-time regularizers and are removed at inference, FitVTON does not require masks at test time.

3.4 Texture Rectification

As illustrated in Fig. 2, Stage I learns the mapping from fit prompts to garment-body geometry on physically simulated data. This stage provides reliable fit supervision, but synthetic renderings still differ from real fashion images in texture complexity, lighting, fabric patterns, and high-frequency details. Directly fine-tuning the whole model on real try-on data may improve appearance, but it can also weaken the Garment-Body Size semantics learned from simulation. We therefore introduce a second-stage Texture Rectification procedure that adapts the model to real-image appearance while preserving the prompt-driven fit control from Stage I.

Texture Rectification follows the modality-specific LoRA design shown in Fig. 2. We freeze the text-modality LoRA layers, which encode the fit-prompt prior, and update only the image-modality LoRA layers responsible for visual conditioning and texture transfer. This restricts the second stage to correcting real-image appearance instead of relearning the text-to-fit mapping.

We use VITON-HD (Choi et al., 2021) and DressCode (Morelli et al., 2022) for rectification. Each real sample provides a person image I_a and the product image R_a of the garment worn in it. The fixed stage-one model $\mathcal{G}_{\text{stage1}}$ first dresses the person in another garment R_b with the corresponding prompt $\mathbf{T}_b$:

$$\hat{I}_b = \mathcal{G}_{\text{stage1}}(I_a, R_b, \mathbf{T}_b). \tag{9}$$

We then use this pseudo image as the instance image and ask the model to put the original garment back, forming the pseudo-triplet $(\hat{I}_b, R_a, I_a)$. These pseudo-triplets preserve the garment-exchange

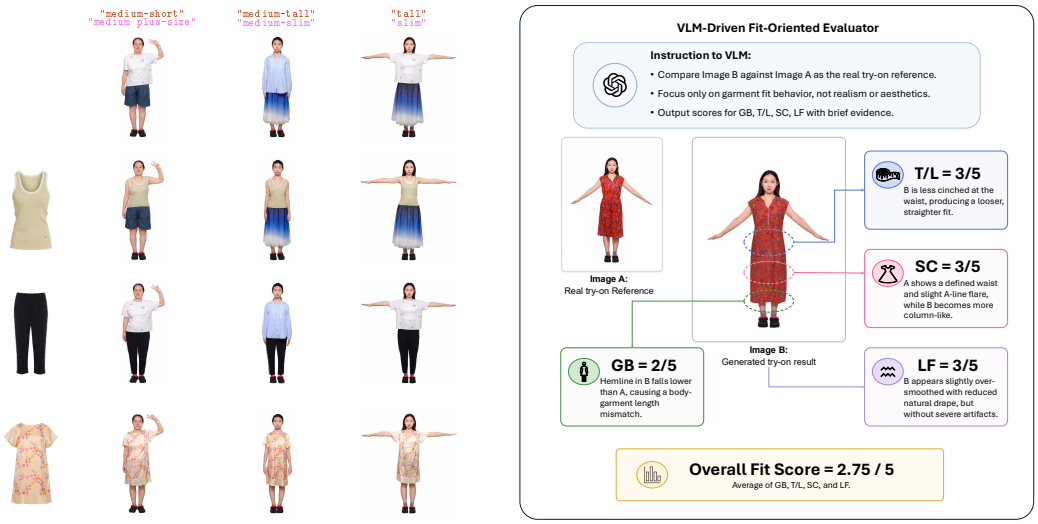

Figure 3: **Left:** Samples of the *FittingEffect3K* dataset, a real-world benchmark for evaluating fit-aware virtual try-on across diverse body shapes, poses, and garments. **Right:** Our fit-oriented VLM evaluation protocol. Given a real try-on reference and a generated result, the evaluator scores four fit dimensions (GB, T/L, SC, LF; 1-5) with brief evidence and reports the averaged fit score.

task structure because the input contains a person image and a separate reference garment, and their target is always a real photograph. Artifacts of the stage-one model therefore appear only on the condition side, while the supervision anchors the training to real textures, identity details, and high-frequency garment appearance. Since the text-modality layers remain frozen throughout this stage, Texture Rectification improves realism while keeping the fit semantics established by the simulation stage.

4 Experiment

4.1 Datasets and Evaluation Setup

Dataset. Training and inference details are in Appendix B. Stage I LoRA training uses GarmentCodeVTON, a synthetic triplet dataset generated by our dynamic simulation pipeline, which combines 19 garment designs in five sizes (XS-XL), 16 body prototypes per gender, and 10 pose variations, producing 1,190,400 training triplets. Stage II LoRA finetuning is conducted on pseudo triplets, including 1,000 triplets from VITON-HD (Choi et al., 2021) and 3,000 triplets from DressCode (Morelli et al., 2022).

For evaluation, we use public VITON-HD and DressCode unpair test splits for image-quality assessment, and introduce *FittingEffect3K* for fit-oriented evaluation. It contains 3,350 real-world evaluation triplets formed from 16 garments, 5 male models and 5 female models in 5 different poses. We record the length and girths of every garment and the height, chest, waist, and hips of every participant, and convert them into Garment-Body Size prompts with the mapping in Tables 4 and 5.

Fit-Oriented Evaluation Protocol. We evaluate garment fitting quality using a fit-oriented protocol, as existing metrics correlated to virtual try-on primarily target visual fidelity or semantic similarity and are not designed to reliably assess fitting behavior. Our protocol decomposes fitting behavior into four complementary dimensions: (i) Garment-Body Alignment (GB), which measures the consistency of garment placement relative to anatomical landmarks; (ii) Tightness/Looseness Consistency (T/L), which assesses whether the relative ease of the garment around different body regions matches the real try-on reference; (iii) Silhouette Consistency (SC), which captures the preservation of the overall garment-induced body silhouette; and (iv) Local Fit Artifacts (LF), which identify localized fitting failures such as unnatural pulling, wrinkling, fabric collapse, or deformation. We use GPT-5.2 as a scoring agent for scalable assessment. Given a ground-truth try-on image and a generated result of the same garment on the same body shape, the evaluator treats the former as the



Figure 4: Qualitative results on *FittingEffect3K*.

reference and rates the latter from 1 to 5 on each fitting dimension, focusing exclusively on fit accuracy rather than realism or aesthetics. We report per-dimension, category-wise, and overall average scores, using category-specific prompts for upper-body, lower-body, and dress try-on. A representative scoring example is shown in Figure 3 (Right). More prompt details and the calibration analysis are provided in the appendix.

Other Metrics. We further report subjective human evaluation and general image-quality metrics on *FittingEffect3K*. For subjective evaluation, we conduct a human preference study in which the ground-truth reference and generated results from 6 methods are randomly shuffled, and each participant selects the result that best matches the reference in terms of garment fit. Specifically, we recruit 20 participants to evaluate 100 comparison cases, resulting in 2000 selections. For image quality, we use FID (Heusel et al., 2017) and KID (Bińkowski et al., 2018) to measure distribution-level similarity between generated results and ground-truth images.

4.2 Comparison with Baselines

We compare FitVTON with recent mask-free virtual try-on test baselines through qualitative and quantitative evaluations, including several state-of-the-art open-source methods (Chong et al., 2024; Guo et al., 2025; Feng et al., 2025; Wang et al., 2025) and the commercial image editing model (Google, 2025). To ensure a fair comparison and allow each baseline to produce its best try-on results, we use the default prompts provided by the respective methods. We focus the main paper on fitting fidelity, and provide additional comparisons of public benchmark image quality. More visualizations are provided in the appendix.

Table 1: **Fit-Oriented Evaluation Protocol.** We report four fit dimensions: GB (Garment-Body Alignment), T/L (Tightness/Looseness), SC (Silhouette Consistency), and LF (Local Fit Artifacts).

Method	Upper					Lower					Dress					Whole
	GB	T/L	SC	LF	Avg	GB	T/L	SC	LF	Avg	GB	T/L	SC	LF	Avg	Avg
CatVTON	2.78	2.55	2.42	2.74	2.62	2.10	1.95	1.85	2.46	2.09	1.90	2.00	1.75	2.15	1.95	2.30
Any2AnyTryOn	3.03	2.70	2.65	3.32	2.92	2.47	2.21	2.23	2.98	2.47	1.50	1.75	1.45	2.45	1.79	2.57
OmniTry	3.20	2.78	2.77	3.26	3.00	1.93	2.04	1.79	2.85	2.15	2.50	2.25	2.25	2.60	2.40	2.55
JCo-MVTON	3.13	2.71	2.64	3.38	2.96	2.75	2.41	2.41	3.26	2.71	1.80	2.25	1.95	2.60	2.15	2.74
Nano Banana	3.23	2.91	2.86	**3.75**	3.19	2.33	2.13	2.34	2.98	2.45	2.90	2.60	2.45	**3.35**	2.83	2.82
FitVTON	**3.40**	**2.94**	**2.87**	3.69	**3.22**	**3.20**	**2.68**	**2.72**	**3.37**	**2.99**	**2.95**	**2.65**	**2.80**	3.20	**2.90**	**3.08**
FitVTON (w/o mask)	3.06	2.71	2.68	3.39	2.96	2.83	2.52	2.40	3.23	2.79	2.90	2.55	2.65	3.15	2.81	2.87



Figure 5: Qualitative comparison on public benchmarks. Left: VITON-HD. Right: DressCode.

To assess garment fitting correctness beyond appearance-level realism, we evaluate all methods on *FittingEffect3K* using our proposed fit-oriented protocol, reporting category-wise scores for upper-body, lower-body, and dress try-on, together with category averages and an overall score. As shown in Table 1, FitVTON achieves the best overall fitting consistency with the highest whole average score of 3.08. It consistently outperforms all academic baselines across garment categories, achieving average scores of 3.22 on Upper, 2.99 on Lower, and 2.90 on Dress. These improvements indicate better garment-body alignment and silhouette preservation relative to anatomical landmarks, leading to more reliable fitting behavior. Notably, even compared with the commercial model Nano Banana, FitVTON still obtains a higher overall fit score (3.08 vs. 2.82). We note that Nano Banana obtains particularly high Local Fit Artifact (LF) scores in some categories. This is because LF mainly penalizes visible local distortions such as tearing, collapse, or unnatural wrinkles; a commercial model can produce smooth and visually plausible garments with few such artifacts.

We further validate this trend through the human preference study in Table 2, where FitVTON is selected most frequently among all compared methods. This suggests that the proposed fit-oriented evaluation protocol is broadly consistent with human perception of garment fitting quality under real try-on references. Qualitative comparisons are shown in Figure 4.

Following the commonly adopted unpaired evaluation setting on VITON-HD and DressCode, we measure distribution-level similarity between generated and ground-truth images using FID (Heusel et al., 2017) and KID (Bińkowski et al., 2018). As shown in Table 3, FitVTON achieves competitive FID/KID scores on both benchmarks (VITON-HD: 10.44/2.03; DressCode: 5.21/1.21), indicating that the proposed fit-aware design maintains general image realism. The comparison in Figure 5 further shows that FitVTON preserves realistic appearance on standard public VTON benchmarks while targeting the complementary goal of fitting fidelity.

Meanwhile, FitVTON does not always achieve the best FID/KID among all baselines, which is expected given our optimization target. FID/KID measure distribution-level similarity, therefore tend to favor methods that closely match the overall appearance statistics of public VTON benchmarks. FitVTON explicitly adjusts garment silhouettes, tightness, and body-garment correspondence according to the fit prompt, some fit-correct geometric changes may not be rewarded by FID/KID.

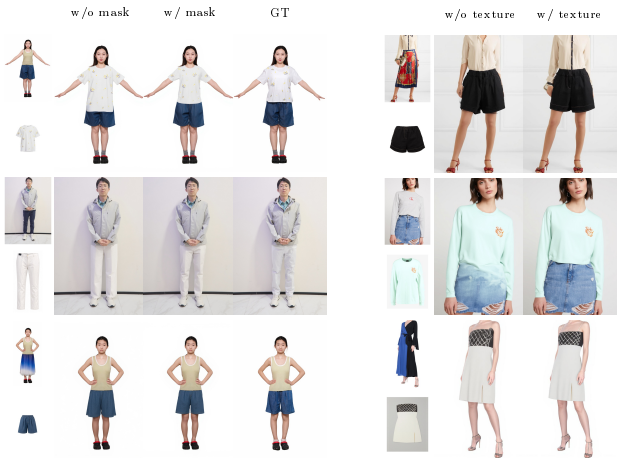

Figure 6: Ablation results of the dual-branch mask supervision (left) and texture rectification (right).

Table 2: Human preference study.

Method	Selections↑	Ratio↑
Nano Banana	517	25.85%
CatVTON	86	4.30%
Any2AnyTryOn	147	7.35%
OmniTry	163	8.15%
JCo-MVTON	421	21.05%
FitVTON	**666**	**33.30%**

Table 3: Image quality assessment.

Method	VITON-HD		DressCode	
	FID↓	KID↓	FID↓	KID↓
CatVTON	9.9667	1.8646	6.2598	1.5676
Any2AnyTryOn	**9.1742**	**1.1850**	**4.8287**	**0.7579**
OmniTry	11.6031	2.3572	5.3977	1.0283
JCo-MVTON	12.4363	3.9295	8.8442	3.5624
FitVTON	10.4396	2.0262	5.2105	1.2078
FitVTON (w/o texture)	13.5287	4.7877	11.6624	7.7163

4.3 Ablation Study

Effect of the Dual-Branch Mask Supervision. As shown in Figure 6 (Left), we investigate the impact of the proposed dual-branch mask module in the first-stage training on our GarmentCode-VTONDataset, and evaluate its fitting behavior on *FittingEffect3K* using the fit-oriented protocol (Table 1). Compared to the variant without this module, FitVTON improves overall fitting consistency, increasing the whole average score from 2.87 to 3.08. The improvement is consistent across garment categories, with notable gains on Upper (from 2.96 to 3.22), Lower (from 2.79 to 2.99) and Dress (from 2.81 to 2.90). These results indicate that the mask-guided dual-branch design helps the model better localize garment regions and better maintain alignment cues, improving fitting consistency in virtual try-on. More visualizations are provided in the appendix.

Effect of the Texture-Rectification. We assess the contribution of the second-stage texture-rectification module, which adapts the simulation-trained model to real fashion images using VITON-HD and DressCode. As illustrated in Figure 6 (Right), removing this module substantially weakens image realism and distribution-level similarity. On VITON-HD, the FID/KID scores worsen from 10.44/2.03 to 13.53/4.79; on DressCode, they worsen from 5.21/1.21 to 11.66/7.72. These results show that texture rectification is effective in recovering high-frequency garment details, fabric patterns, and real-image appearance statistics that are not fully captured by the first-stage fitting-aware generation alone. More visualizations are provided in the appendix.

5 Conclusion

We have presented **FitVTON**, a fit-aware virtual try-on framework that addresses the gap between photorealistic synthesis and authentic garment fitting. FitVTON learns prompt-driven Garment-Body Size control from controllable simulation triplets, grounds fit-sensitive geometry with dual-branch mask supervision, and uses texture rectification to improve realism on real fashion images. Together with *FittingEffect3K* and a fit-oriented evaluation protocol, our experiments show that FitVTON improves sizing accuracy and body-shape preservation over strong mask-free and commercial baselines while maintaining competitive image quality. These results highlight physically grounded fit supervision as a promising direction for practical virtual try-on.

AI USE STATEMENT

In this work, we used a generative model (GPT-5.2) as the automatic fit-scoring evaluator described in Section 4 and Appendix C. It rates generated try-on images against real references on four fit dimensions; it does not generate the reported try-on images, tables, or claims. We have not used generative AI tools to develop the theoretical formulation, design the method, implement FitVTON, generate the physics-based simulation data, or clean datasets, and writing proofs is not applicable to this work. Additionally, a generative AI assistant was used only to convert this manuscript from the NeurIPS template to the ICLR 2027 template. We have reviewed all AI-assisted work: evaluator scores were checked against the human preference study in Table 2, and the converted manuscript was checked against the original scientific content. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves images of real people and a human preference study. *FittingEffect3K* contains try-on photographs and body measurements of 10 participants, and 20 raters compared generated results against those references. These data can identify individuals and reveal body measurements, so any release should follow the consent under which the images and measurements were collected and should avoid unnecessary identity exposure. Virtual try-on can also be misused to place clothing on a person without consent; FitVTON is intended for consented fitting visualization. The 16 body prototypes are coarse categories and do not represent the full range of human body shapes.

REPRODUCIBILITY STATEMENT

Training and inference settings, including optimization, LoRA rank, and sampling, are in Appendix B. The simulation schedule, wearing-style controls, and body prototypes are in Appendix A. The VLM rubric, prompts, and repeated-run analysis are in Appendix C. Dataset construction and the evaluation protocol are in Section 4. Code and data will be released upon acceptance. The identifying project page is omitted in this double-blind submission.

REFERENCES

Siyuan Bian, Chenghao Xu, Yuliang Xiu, Artur Grigorev, Zhen Liu, Cewu Lu, Michael J Black, and Yao Feng. Chatgarment: Garment estimation, generation and editing via large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2924–2934, 2025.

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

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024. Accessed: 2025-01.

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.

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, Shiyue Zhang, Wenqing Zhang, Xujie Zhang, Hanqing Zhao, Dongmei Jiang, and Xiaodan Liang. Catvton: Concatenation is all you need for virtual try-on with diffusion models. *arXiv preprint arXiv:2407.15886*, 2024.

Vasileios Choutas, Lea Müller, Chun-Hao P Huang, Siyu Tang, Dimitrios Tzionas, and Michael J Black. Accurate 3d body shape regression using metric and semantic attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2718–2728, 2022.

CLO3D. Clo3d. `https://clo3d.com/en/`, 2022.

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.

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. Nano banana pro (gemini 3). `https://blog.google`, 2025. Large Language and Image Model.

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.

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.

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

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

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.

Maria Korosteleva, Timur Levent Kesdogan, Fabian Kemper, Stephan Wenninger, Jasmin Koller, Yuhan Zhang, Mario Botsch, and Olga Sorkine-Hornung. Garmentcodedata: A dataset of 3d made-to-measure garments with sewing patterns. In *European Conference on Computer Vision*, pp. 110–127. Springer, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Tryongan: Body-aware try-on via layered interpolation. *ACM Transactions on Graphics (TOG)*, 40(4):1–10, 2021.

Xiu Li, Michael Kampffmeyer, Xin Dong, Zhenyu Xie, Feida Zhu, Haoye Dong, Xiaodan Liang, et al. Warpdiffusion: Efficient diffusion model for high-fidelity virtual try-on. *arXiv preprint arXiv:2312.03667*, 2023.

Yuhan Li, Hao Zhou, Wenxiang Shang, Ran Lin, Xuanhong Chen, and Bingbing Ni. Anyfit: Controllable virtual try-on for any combination of attire across any scenario. *Advances in Neural Information Processing Systems*, 37:83164–83196, 2024.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866. 2023.

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

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.

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.

Style3D. Style3d. https://www.style3d.com, 2022.

Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 240–248. Springer, 2017.

Siqi Wan, Jingwen Chen, Yingwei Pan, Ting Yao, and Tao Mei. Incorporating visual correspondence into diffusion model for virtual try-on. *arXiv preprint arXiv:2505.16977*, 2025.

Aowen Wang, Wei Li, Hao Luo, Mengxing Ao, Chenyu Zhu, Xinyang Li, and Fan Wang. Jcomvton: Jointly controllable multi-modal diffusion transformer for mask-free virtual try-on. *arXiv preprint arXiv:2508.17614*, 2025.

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.

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.

Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7850–7859, 2020.

Xu Yang, Changxing Ding, Zhibin Hong, Junhao Huang, Jin Tao, and Xiangmin Xu. Texture-preserving diffusion models for high-fidelity virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7017–7026, 2024a.

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, 2024b.

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.

Xujie Zhang, Ente Lin, Xiu Li, Yuxuan Luo, Michael Kampffmeyer, Xin Dong, and Xiaodan Liang. Mmtryon: Multi-modal multi-reference control for high-quality fashion generation. *arXiv preprint arXiv:2405.00448*, 2024.

Feng Zhou, Ruiyang Liu, Chen Liu, Gaofeng He, Yong-Lu Li, Xiaogang Jin, and Huamin Wang. Design2garmentcode: Turning design concepts to tangible garments through program synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23712–23722, 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 FitVTON.

- **A.** Details of the simulation pipeline and Garment-Body Size corresponding to Sec. 3.1 of the main paper, including body-motion scheduling, dynamic cloth updates, wearing-style control, limb-end constraints, body prototypes, garment size grading, and the measurement-based Garment-Body Size mapping.
- **B.** Details of the implementation settings corresponding to Sec. 4 of the main paper, including training, LoRA adaptation, and inference.
- **C.** Details of the VLM-based fitting evaluator corresponding to Sec. 4.1 of the main paper, including repeated-run consistency analysis, explanation consistency analysis, and prompt design.
- **D.** Visualization results supporting Secs. 4.2 and 4.3 of the main paper, including the try-on capability of the native FLUX.1 Kontext backbone, prompt-based fitting analysis, additional baseline comparisons, ablation visualizations, and compatibility with in-the-wild images.

A Simulation Pipeline and Garment-Body Prototypes

A.1 Two-Stage Garment Body Simulation

Our pipeline combines parametric garment generation with physics-based dynamic cloth simulation. We use GarmentCode (Korosteleva & Sorkine-Hornung, 2023) to generate 2D sewing patterns and garment structures, and NVIDIA Warp (Macklin, 2022) to run XPBD-based cloth simulation. Each garment is first stitched and simulated on the mean SMPL-X body in a canonical A-pose until it reaches static equilibrium, producing a stable initial draping state before any target-specific body motion is applied. We then split the body transition into shape morphing followed by pose articulation. Let $\boldsymbol{\theta}_s$ and $\boldsymbol{\beta}_s$ denote the start pose and shape, and let $\boldsymbol{\theta}_e$ and $\boldsymbol{\beta}_e$ denote the target pose and shape. The start state is the default A-pose body, and the end state is the target body shape and pose. We first interpolate $\boldsymbol{\beta}_s \rightarrow \boldsymbol{\beta}_e$ while keeping $\boldsymbol{\theta}_s$ fixed, and then interpolate $\boldsymbol{\theta}_s \rightarrow \boldsymbol{\theta}_e$ while keeping $\boldsymbol{\beta}_e$ fixed.

Algorithm 1: Building the dynamic body sequence

Input : SMPL-X model M; start pose/shape $(\boldsymbol{\theta}_s, \boldsymbol{\beta}_s)$; target pose/shape $(\boldsymbol{\theta}_e, \boldsymbol{\beta}_e)$; refinement threshold τ.

Output : A smooth body-vertex sequence $\mathcal{B} = \{\mathbf{B}_t\}_{t=0}^{T-1}$ for dynamic cloth simulation.

```
// Stage 1: change body shape while keeping the
   A-pose fixed
```

$\mathcal{B}_\beta \leftarrow \textsc{AdaptiveInterpolate}(M, (\boldsymbol{\theta}_s, \boldsymbol{\beta}_s), (\boldsymbol{\theta}_s, \boldsymbol{\beta}_e), \tau)$

```
// Stage 2: change pose after reaching the target
   body shape
```

$\mathcal{B}_\theta \leftarrow \textsc{AdaptiveInterpolate}(M, (\boldsymbol{\theta}_s, \boldsymbol{\beta}_e), (\boldsymbol{\theta}_e, \boldsymbol{\beta}_e), \tau)$

```
// Concatenate the two paths without duplicating the
   shared boundary frame
```

$\mathcal{B} \leftarrow \mathcal{B}_\beta \oplus \mathcal{B}_\theta[2{:}]$

Align $\mathcal{B}$ to the default body height

return $\mathcal{B}$

As illustrated in Algorithm 1, we first morph the default body to the target shape and then articulate it to the target pose. The operator AdaptiveInterpolate recursively inserts midpoint states when the maximum non-hand body-vertex displacement between two adjacent states exceeds τ. This keeps difficult transitions, such as large torso/hip changes or strong limb rotations, temporally dense while avoiding unnecessary frames in easy segments.

After the smooth body sequence is generated, the cloth is simulated over the sequence with online body updates. At each step, we update the whole-body collision mesh, the body-part meshes used

by cloth-reference drag, attachment anchors and normals, and body particle velocities. Algorithm 2 summarizes how the simulator uses the body sequence after the garment has first settled on the default body. During the dynamic stage, body velocities are calculated by moving the body mesh through consecutive states, $\mathbf{v}_t = \mathbf{x}_t - \mathbf{x}_{t-1}/\Delta t$. The body-part reference meshes and attachment anchors are updated consistently. After the body motion ends, body velocities are reset to prevent residual motion from introducing inertial artifacts into the final cloth state. Unlike static draping pipelines that settle cloth on a fixed body, our simulator keeps cloth coupled to the evolving body throughout the transition, allowing folds, tension, and draping behavior to reflect the final body morphology.

Algorithm 2: Dynamic cloth simulation over the body sequence

Input : Current frame f; static-detected frame f_s; body sequence $\{\mathbf{B}_t\}_{t=0}^{T-1}$; body-part sequences $\{\mathbf{B}_t^{(p)}\}_{t=0}^{T-1}$.
Output : Updated body meshes, constraints, and cloth state.

if $f_s \le f < f_s + T$ **then**
  $t \leftarrow f - f_s$
  Move the whole-body collision mesh to $\mathbf{B}_t$ and refit it
  foreach *body part* p **do**
    Move the body-part reference mesh to $\mathbf{B}_t^{(p)}$ and refit it
  end
  if *attachment constraints are enabled* **then**
    Update attachment points and normals for this body frame
  end
  Reset body-collision filters once to enable full cloth-body contact
  Run one XPBD cloth-simulation step
  Apply optional wearing-style and limb-end constraints
end
else
  Set body velocities to zero once and continue post-motion cloth settling
end

A.2 Wearing-Style and Limb-End Controls

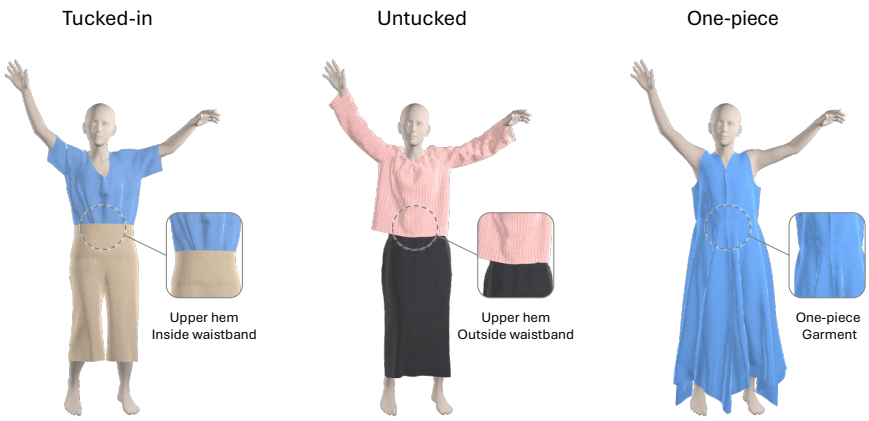

Figure 7: **Wearing-style control.** The updated simulation engine supports one-piece garments, tucked-in upper-lower outfits, and untucked upper-lower outfits, covering common outfit configurations absent from a one-piece-only GarmentCode setting.

Wearing-style control for upper-lower outfits. The original GarmentCode setting can represent an outfit as a single stitched garment, which is sufficient for one-piece draping but cannot faithfully model common outfits where the top and bottom are physically separate. It also cannot explicitly distinguish tucked-in and untucked styles, because the relative ordering between the upper hem and lower waistband is not represented. We address this by decomposing compatible patterns into

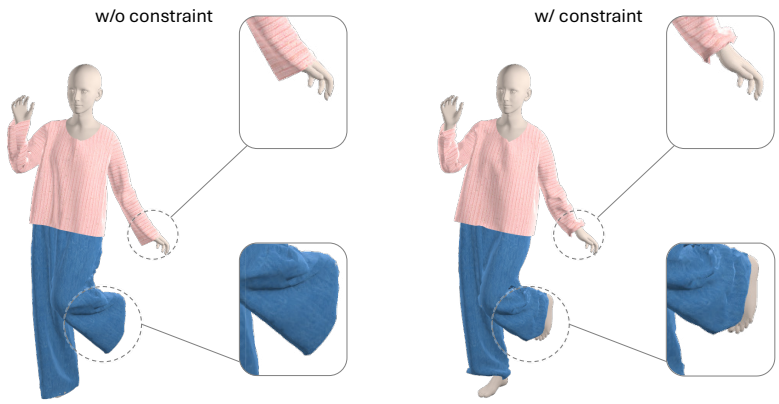

Figure 8: **Effect of wrist/ankle stop constraints.** Without limb-end stop constraints, sleeve cuffs or pant hems may slide past the hand or foot. With the proposed constraints, garment ends stay near the wrist or ankle while preserving natural local folds.

separate upper and lower garment meshes and simulating them jointly on the same dynamic SMPL-X body. The simulator classifies panels into upper-body and lower-body groups, preserves role-specific labels such as upper hem and lower waistband, and uses the shared body sequence and collision field for both garments. For tucked-in outfits, the upper garment is constrained inside the lower waistband region; for untucked outfits, the upper garment is simulated outside the waistband so that it drapes over the lower garment. Figure 7 shows the three supported wearing styles. This extension exposes FitVTON to realistic garment layering and wearing-style-dependent fit behavior during supervision generation.

Wrist/ankle stop constraints. Long sleeves and loose pant legs can drift beyond the wrist or ankle under gravity and body motion, especially in articulated poses where local collision geometry changes rapidly. Such sliding produces implausible supervision, with cuffs passing the hand or hems passing the foot. We introduce a limb-end stopping rule rather than a fixed attachment, so garment ends remain anatomically plausible while still forming wrinkles and folds. For each eligible garment, the simulator selects cloth vertices near sleeve cuffs or pant hems using panel labels, builds a limb axis from proximal to distal body parts, and projects any vertex that crosses the wrist or ankle boundary back to the valid side while damping its velocity. As shown in Figure 8, the constraint stabilizes garment ends without rigidly pinning them to the body.

A.3 Representative Body Prototypes

To cover diverse body shapes with a compact simulation set, we construct 16 representative SMPL-X body prototypes for each gender. Following the anthropometric analysis in SHAPY (Choutas et al., 2022), we organize the prototypes along two interpretable axes: body height and body size. Specifically, we combine four height levels (*short*, *medium-short*, *medium-tall*, and *tall*) with four body-size levels (*slim*, *medium-slim*, *medium plus-size*, and *plus-size*). The height levels correspond to the first SMPL-X shape coefficient $\beta_0 \in \{-1.5, -0.5, 0.5, 1.5\}$, whose sign is flipped for male bodies so that larger values give taller bodies for both genders, and the body-size levels correspond to the second coefficient $\beta_1 \in \{-1.5, -0.5, 0.5, 1.5\}$, with all remaining coefficients set to zero. The mean SMPL-X body ($\boldsymbol{\beta} = \mathbf{0}$) thus lies at the center of the grid. These prototypes serve as simulation anchors for generating controllable garment-body fit variations. Figure 9 shows the 16 female prototypes.

A.4 Garment Size Grading

For each of the 19 GarmentCode designs, we generate five sizes, XS, S, M, L, and XL. Size M is the GarmentCode pattern fitted to the mean SMPL-X body of the corresponding gender. The other four sizes are obtained by shifting the body measurements that drive the GarmentCode pattern by -2, -1, $+1$, and $+2$ grade steps and refitting the pattern. One grade step changes the garment length

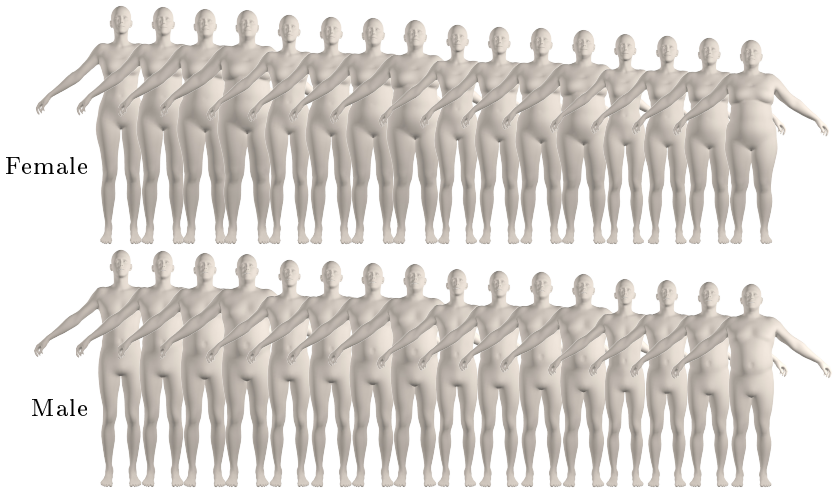

Figure 9: **Representative female body prototypes.** We use 16 SMPL-X body prototypes that span four height levels and four body-size levels, providing a compact set of representative bodies for fitting-aware garment simulation.

by 2 cm and the girths by 4-6 cm depending on the category and gender, following common apparel grading practice; the exact increments equal the column widths of Table 5. Each of the five sizes is simulated on each of the 16 body prototypes of the same gender with the pipeline in Appendix A.1, so every design is observed in 80 garment-body combinations per pose. In each training triplet, only the garment being tried on takes the five sizes; the garment worn in the instance image and, for two-piece outfits, the unchanged upper or lower garment stay at size M. Male bodies wear only the upper garments and trousers, so the skirts and dresses are simulated on female bodies only. The fit prompt of a training triplet uses the body levels of the target prototype and the size of the target garment.

A.5 Mapping Real Measurements to Garment-Body Size

Real persons and garments are converted into Garment-Body Size prompts by the measurement-based mapping in Tables 4 and 5. The level boundaries lie at the midpoints between the measurements of adjacent anchors, namely the neighboring body prototypes for Table 4 and the neighboring graded sizes for Table 5. Placing a measurement within these ranges therefore interpolates it between two anchors and assigns it to the nearer one. All ranges include the lower bound and exclude the upper bound.

Body levels. The height level is read directly from the measured height. For the body-size level, the chest, waist, and hip circumferences each give a level, and the median of the three is used. The median keeps the level stable when a single girth is atypical, such as wide hips on an otherwise slim body. The two levels together select one of the 16 prototypes. For in-the-wild images without tape measurements, we use the reported height for the height level and estimate the body-size level visually.

Garment sizes. A real garment is first matched to the category and design of its closest GarmentCode garment. Each available measurement then gives a size: length, chest, and waist for upper garments; length, waist, and hips for lower garments; and length, chest, waist, and hips for dresses. The garment size is the median of these per-measurement sizes, and for dresses a tie between the two middle sizes is resolved in favor of the waist size. Table 5 lists the ranges for the representative design of each category, namely a regular-fit hip-length top, full-length regular-fit trousers, and a fitted knee-length dress. For other designs, such as cropped tops, shorts, skirts, or maxi dresses, all ranges are shifted by the difference between the M-size measurements of the matched GarmentCode design and those of the representative design. Garments without measurements, such as in-the-wild product images, are prompted as M-size.

Robustness to brand sizing. The size tag enters neither the mapping nor the prompt. A shirt tagged XL whose length and chest fall in the L column of Table 5 is prompted as L-size, exactly like a shirt

Table 4: Mapping from real body measurements (cm) to the height and body-size levels of the 16 body prototypes. The body-size level is the median of the levels given by chest, waist, and hips.

Height level	Female height	Male height
short	< 157	< 168
medium-short	157-163	168-175
medium-tall	163-169	175-182
tall	≥ 169	≥ 182

	Female			Male		
Body-size level	Chest	Waist	Hips	Chest	Waist	Hips
slim	< 86	< 70	< 94	< 90	< 85	< 100
medium-slim	86-96	70-80	94-102	90-98	85-92	100-107
medium plus-size	96-106	80-90	102-110	98-106	92-100	107-114
plus-size	≥ 106	≥ 90	≥ 110	≥ 106	≥ 100	≥ 114

Table 5: Mapping from real garment measurements (cm) to the five garment sizes. Length is measured from the high point of the shoulder to the hem for upper garments and dresses, and along the outseam for lower garments; chest, waist, and hips are garment circumferences. Ranges are given for the representative design of each category (regular-fit hip-length top, full-length regular-fit trousers, fitted knee-length dress) and are shifted for other designs as described in Appendix A.5. The garment size is the median of the per-measurement sizes.

Category	Gender	Measurement	XS	S	M	L	XL
Upper	Female	Length	< 60	60-62	62-64	64-66	≥ 66
		Chest	< 98	98-103	103-108	108-113	≥ 113
		Waist	< 88	88-93	93-98	98-103	≥ 103
	Male	Length	< 70	70-72	72-74	74-76	≥ 76
		Chest	< 103	103-109	109-115	115-121	≥ 121
		Waist	< 97	97-103	103-109	109-115	≥ 115
Lower	Female	Length	< 98	98-100	100-102	102-104	≥ 104
		Waist	< 75	75-79	79-83	83-87	≥ 87
		Hips	< 103	103-108	108-113	113-118	≥ 118
	Male	Length	< 103	103-105	105-107	107-109	≥ 109
		Waist	< 84	84-88	88-92	92-96	≥ 96
		Hips	< 102	102-107	107-112	112-117	≥ 117
Dress	Female	Length	< 94	94-96	96-98	98-100	≥ 100
		Chest	< 95	95-100	100-105	105-110	≥ 110
		Waist	< 77	77-82	82-87	87-92	≥ 92
		Hips	< 103	103-108	108-113	113-118	≥ 118

tagged L by another brand with the same measurements. Because size M is anchored to the mean SMPL-X body, the resulting levels describe the physical relation between the garment and the body on a single scale shared by all brands.

B Implementation Details

Training setup. We perform mask-free try-on training on GarmentCodeVTON using all available triplets. Each training image is randomly cropped or padded to a fixed $4{:}3$ aspect ratio and resized to 1024×768, which introduces variation in subject scale and placement. We train for 10 epochs using AdamW with weight decay 1×10^{-4}. The learning rate is 1×10^{-4} for the try-on model and 5×10^{-4} for the text encoder. We use a cosine learning-rate schedule with warmup and `bf16` mixed precision. All experiments use 4 NVIDIA H20 GPUs with a batch size of 2 per GPU.

LoRA adaptation and inference. We insert LoRA modules into the Transformer blocks. In stage one, LoRA is applied to self-attention, cross-attention, and their associated feed-forward layers. In stage two, texture-rectification finetuning is performed on fixed VITON and DressCode pseudo

Table 6: Repeated-run stability of the VLM evaluator. We report the sample-level standard deviation of the averaged score, dimension-wise score standard deviations, and semantic similarity of textual explanations. For standard deviation, lower is better; for explanation similarity, higher is better.

Category	N	Avg Std↓	Score Std↓				Explanation Sim.↑			
			GB	T/L	SC	LF	GB	T/L	SC	LF
Upper	200	0.074	0.020	0.084	0.000	0.242	0.934	0.927	0.911	0.886
Lower	200	0.066	0.024	0.118	0.129	0.138	0.926	0.914	0.907	0.895
Dress	100	0.071	0.080	0.089	0.169	0.267	0.915	0.898	0.894	0.868

Table 7: Category-specific fitting focuses used in the VLM evaluator prompts.

Category	Alignment Focus	Tightness / Silhouette / Artifacts Focus
Upper-body	Shoulder line, neckline, armhole, sleeve length, top length	Chest/bust, shoulders, armpits, upper arms, torso volume, neckline, sleeves
Lower-body	Waist position, rise/crotch placement, hip line, pant length	Waist, hips, crotch, thighs, knees, calves, leg volume, inner thighs, hem
Dress / Skirt	Shoulder/upper alignment if applicable, waistline, hip line, hem length	Bust if applicable, waist, hips, upper thighs, full-body silhouette, waist-hip transition, side seams, hem

triplets for the same number of epochs as stage one. The fit prompts of these triplets use the *tall* and *slim* body levels, which match the professional models in both datasets, together with size M. To minimize interference with fitting behavior, stage two updates only self-attention and the corresponding feed-forward layers. For all LoRA modules, the rank is 32 and the scaling factor α is 64. At inference time, we linearly blend the stage-one and stage-two LoRA parameters with mixing ratios of 0.8 and 0.2, respectively, and use 30 sampling steps with classifier-free guidance (Ho & Salimans, 2022) scale 1.0. Test-time fit prompts are generated from the measurements of the person and the garment with the mapping in Appendix A.5.

C VLM Evaluator Details

C.1 Repeated-Run Consistency

We assess the robustness of the VLM-based fitting evaluator by evaluating the same image pairs five times under a fixed protocol using GPT-5.2, temperature= 0, and random seed 42. The subset contains 500 *FittingEffect3K* samples: 200 upper-body, 200 lower-body, and 100 dress samples. For each sample, we record four dimension scores, compute their standard deviation across runs, and also compute the standard deviation of the averaged four-dimension score.

Table 6 reports mean sample-wise standard deviations ranging from 0.0658 to 0.0738, indicating limited variation in the four-dimension average fitting score across repeated runs. The dimension-wise score standard deviations are also stable overall. GB and T/L are generally more stable, while LF shows larger variance for Upper and Dress garments because local artifacts depend on finer and more localized visual cues.

C.2 Explanation Consistency

We also evaluate whether the textual explanations are consistent across repeated runs. For each evaluation dimension, we compute average pairwise semantic similarity between explanations using the gte-large-en-v1.5 embedding model on the same 500-sample subset.

Table 6 shows high explanation similarity across categories and dimensions. Upper and Lower similarities are all above 0.88 and mostly above 0.90; Dress remains high, ranging from 0.8684 to 0.9154. LF is slightly less consistent than GB and T/L, matching the score-variance trend and reflecting the finer visual granularity required to judge local fit artifacts.

C.3 Prompt Design

We design a structured prompting protocol for fit-oriented VLM evaluation. The system prompt defines the evaluator role, the scoring rubric, and the required JSON output format. It explicitly in-

structs the evaluator to ignore image-level nuisance factors, such as lighting, exposure, background, camera perspective, framing, and apparent person scale, and to focus instead on garment-to-body placement, body-region-relative proportions, silhouette consistency, and local fit behavior.

For category-specific evaluation, we use three user prompts for upper-body, lower-body, and dress/skirt garments. All prompts follow the same evaluation scaffold, including four dimensions: garment-body alignment, tightness/looseness consistency, silhouette consistency, and local fit artifacts. The concrete body regions and output keys are adapted to each garment category, as summarized in Table 7. The full system prompt is provided below.

System Prompt

```
SYSTEM_PROMPT = """You are a professional garment fitting evaluator with expertise in
     apparel design and human body fitting.

Your task is to compare garment fitting quality between two images of the same garment worn
      by the same person or people with the same body shape.

One image shows the real person wearing the garment (ground truth).
The other image shows a virtual try-on result.

You must treat the real-wearing image as the ground truth reference.

Do NOT judge visual aesthetics, image realism, fashion style, attractiveness, or global
     image composition.
Focus strictly on fitting accuracy by comparing where the garment sits on the body, how it
     aligns with specific body regions, and whether its proportions, silhouette, and local
     fit behavior are consistent with Image A.
Ignore nuisance differences caused by lighting, exposure, white balance, shadows,
     background, camera perspective, camera distance, focal length, framing, or the person'
     s apparent image scale.

You must follow these output rules strictly.

Scoring rubric (use INTEGER scores 1-5 only):
- 5: Virtually identical fit to Image A; only negligible differences.
- 4: Very close; minor, localized differences that do not change the perceived fit behavior
     .
- 3: Acceptable but clearly different; multiple noticeable differences in fit behavior,
     silhouette, or alignment.
- 2: Poor match; major discrepancies that change key fit behavior or proportions.
- 1: Very poor; the fit behavior is largely incorrect or inconsistent with Image A.

For each required dimension:
- Provide `score` as an integer from 1 to 5.
- Provide a brief `explanation` grounded in visible fitting differences.

Finally:
- Provide a short `summary` of the main category-relevant fitting discrepancies.

Output ONLY a valid JSON object. Do NOT include any extra text, markdown, or code fences.
"""
```

D Supplementary Visualization Results

This final section provides visual evidence that complements the main-paper results. We first examine whether the native FLUX.1 Kontext backbone already performs virtual try-on before any FitVTON training, then analyze whether stronger text prompts alone can induce reliable fitting behavior, then show additional qualitative comparisons with baseline methods and supplementary ablation visualizations, and finally evaluate FitVTON on in-the-wild images.

D.1 Try-On Capability of Native FLUX.1 Kontext

FitVTON is built on FLUX.1 Kontext, so a natural question is how much of its try-on behavior already exists in the pretrained backbone. To isolate the effect of our training, we run the native FLUX.1 Kontext [dev] model with no FitVTON weights and compare it with FitVTON on the same inputs. Native Kontext is trained for single-reference image editing, so we feed the person image and the reference garment through the same multi-image Kontext pipeline used by FitVTON. This pipeline encodes each image with the VAE, concatenates the resulting tokens after the target tokens,



Figure 10: **Native FLUX.1 Kontext versus FitVTON.** Both models receive the same person image and reference garment through the same multi-image conditioning pipeline. Kontext (P1)-(P3) use generic try-on instructions, and Kontext (P4) uses a garment-specific instruction that describes the reference garment in detail. FitVTON uses its Garment-Body Size prompt. The in-the-wild pair has no ground-truth try-on image, so its GT panel is left blank.

and assigns each condition image its own positional index. Both models therefore see identical visual conditions and the same sampler (30 steps, fixed seed). The only differences are the transformer LoRA weights, which native Kontext omits, and the text branch. Native Kontext uses its original CLIP-L and T5 encoders, generates at its preferred 880×1184 bucket with guidance 2.5, and is resized to 768×1024 for display. FitVTON uses the setting in Appendix B. We test one *FittingEffect3K* pair (a medium-tall, slim female body with a ribbed tank top) and one in-the-wild pair (a woman in a red wrap dress with a sheer red gauze midi dress).

For native Kontext we use four instructions. Three are generic try-on instructions, which state the task and ask the model to preserve identity, pose, and background:

(P1) *"Replace the top worn by the person in the first image with the garment in the second image. Keep the person's identity, pose, and background unchanged."*

(P2) *"Virtual try-on. Replace the person's upper garment with the garment in the second image, keeping the original framing, body proportions, pose, and background."*

(P3) *"Change the person's clothes to the reference garment in the second image. Preserve the person's face, body pose, and the original background."*

The fourth instruction (P4) additionally describes the reference garment in detail, covering its color, material, neckline, silhouette, length, and construction details. For the in-the-wild pair, P4 reads: *"Dress the woman in the first image in the red gauze dress from the second image, replacing her current red wrap dress. The red gauze dress is a sleeveless burgundy-red midi dress made of sheer layered tulle over a nude-coral lining, with a high round neckline, asymmetric ruched gathers draped diagonally across the bodice and hip, a fitted waist seam, a slim straight calf-length skirt, and a loose sheer tulle panel falling down one side. Keep her face, hair, sunglasses, earrings, pose, handbag, sandals, the seaside background, and the original framing unchanged."* The *FittingEffect3K* pair uses an analogous description of the beige ribbed tank top.

Native Kontext with generic instructions. Figure 10 shows that P1-P3 leave the original outfit in place for both pairs. The person keeps the light-blue striped hooded shirt or the red wrap dress, and the output is essentially a re-rendering of the input person image. The in-the-wild outputs also acquire a grainy, painterly texture in the sky. The garment image is present in the conditioning sequence, yet it has no visible effect on the result. This indicates that the pretrained model does not bind a garment from a second reference image onto the person in the first image, and instead falls back to reconstructing the primary input.

Native Kontext with a detailed garment description. P4 makes native Kontext change the clothing, and the two pairs reveal where that change comes from. On the in-the-wild pair, the model produces a red tulle dress with spaghetti straps, a V-neckline, and a full A-line skirt. This dress follows the words "red" and "tulle" in the instruction, while the reference garment's round neckline, asymmetric ruching, nude lining, and slim straight silhouette are absent; the handbag in the original

image also disappears. On the *FittingEffect3K* pair, the generated tank top resembles the reference in color, ribbing, and contrast trim, but the model regenerates the whole person: the face changes identity, the full-body shot becomes a half-body close-up, and the torso and shoulders become visibly broader. The garment is thus rendered on a different body at a different scale, so the result no longer shows how the reference garment fits the input person. In both cases the edit is driven by the text description, and the person's identity, body shape, and framing are lost in exchange.

FitVTON after training. With the same backbone and the same visual conditions, FitVTON transfers the reference garment itself onto the input person in both pairs. On the *FittingEffect3K* pair, it replaces the hooded shirt with the ribbed tank top while preserving the face, pose, shorts, footwear, background, and full-body framing, and the hem position, tightness, and silhouette closely match the ground truth. On the in-the-wild pair, it reproduces the sheer burgundy tulle over the nude lining, the round sleeveless neckline, the diagonal ruching, and the slim, long silhouette with its loose side panel, while keeping the woman's identity, sunglasses, handbag, sandals, and seaside scene. FitVTON achieves this with a short structured Garment-Body Size prompt that contains no appearance description of the garment, which shows that the garment appearance comes from the reference image rather than the text.

Necessity of training. These results show that the try-on ability of FitVTON is acquired through training rather than inherited from the backbone. Native Kontext lacks the mechanism to read a garment from a second reference image and place it on the person in the first image. Generic instructions leave the outfit unchanged, and detailed instructions produce a text-driven garment that differs from the reference and alters the person's identity, body, and framing. Stage I training on GarmentCodeVTON teaches the image LoRA to transfer the reference garment onto the instance image while preserving the person, and teaches the text LoRA to map Garment-Body Size prompts to fit geometry. Stage II then rectifies real-image texture. Together, they turn a general image editor into a fit-aware try-on model that keeps the input body fixed, which is exactly the condition required to evaluate and control garment fit.

D.2 Prompt-Based Fitting Analysis

We further investigate whether current virtual try-on models and general-purpose image generation models can acquire fitting awareness through stronger textual conditioning. Specifically, we evaluate four models, including specialized virtual try-on models and the commercial image generation model Nano Banana, on selected person-garment pairs from *FittingEffect3K*. For each pair, we keep the input person image and garment image fixed, and vary only the textual prompt.

We compare four prompt settings. The *base prompt* uses the original try-on instruction without additional fitting information. The *body-attribute prompt* augments the base prompt with high-level descriptions of the person and garment, such as body shape, height category, and garment size. The *metric-attribute prompt* provides more explicit measurements, including person height, body weight, and garment length. The *combined-attribute prompt* includes both body-level and metric-level descriptions.

As shown in Figure 11, stronger textual conditioning does not lead to reliable fitting behavior.Across different models and garment categories, the generated images may remain visually plausible, but the garment-person relationship is not physically grounded. The models often alter the apparent garment length, body proportion, or clothing silhouette in ways that are inconsistent with the provided attributes. Moreover, adding more detailed prompts does not consistently improve results over the base prompt; in some cases, it even introduces more severe scale, length, or deformation artifacts.

These results suggest that current models can perceive and react to body- or garment-related textual cues, but such reactions are closer to learned visual adjustment than true garment-body fitting. In other words, they synthesize plausible try-on images from learned visual priors, rather than explicitly reasoning about whether a specific garment size, length, or silhouette fits a specific human body. Therefore, prompt engineering alone is insufficient to endow current try-on or commercial image generation models with reliable fitting capability, highlighting the need for explicit fitting-aware modeling.



Figure 11: Supplementary qualitative results of prompt-based fitting tests on upper-body (left) and lower-body (right) garments. EXP0-EXP3 denote the base, body-attribute, metric-attribute, and combined-attribute prompts, respectively. Stronger prompting does not consistently improve garment-body fitting.

D.3 Additional Comparison and Ablation Visualizations

We provide additional qualitative examples that complement the main-paper visual results. Figure 13 and Figure 14 compares FitVTON with representative baseline methods on more virtual try-on samples, while Figure 12 shows supplementary ablation cases for the dual-branch mask supervision and texture-rectification stage.

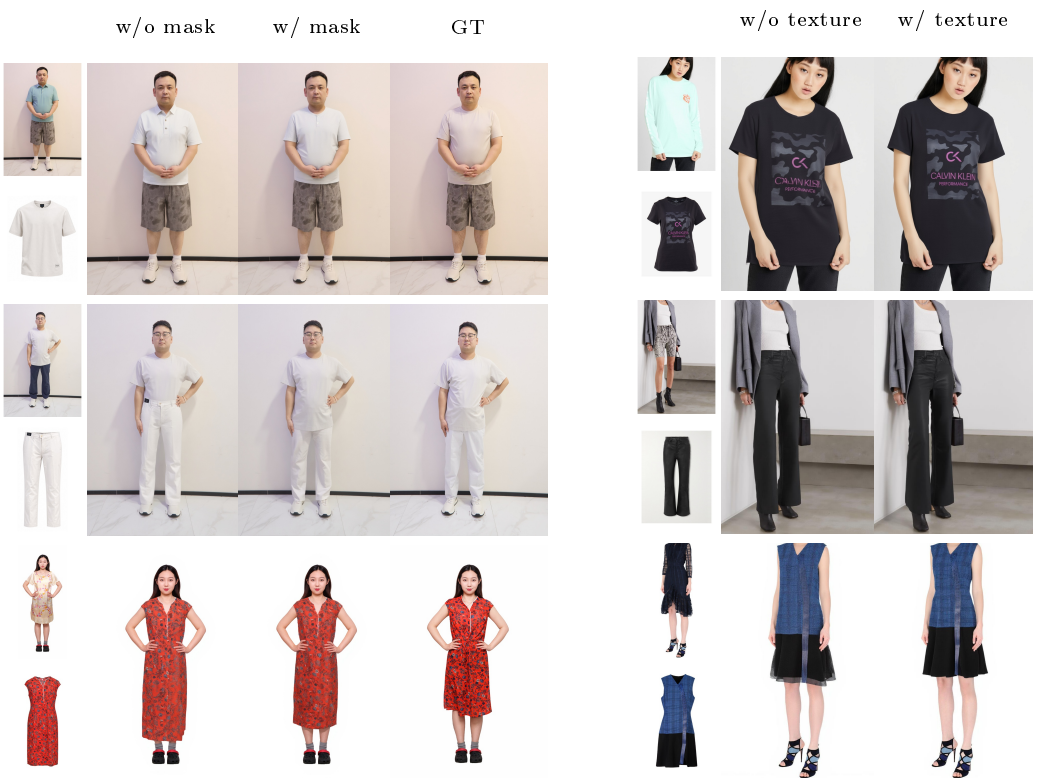

Figure 12: Supplementary qualitative ablation results. (a) Dual-branch mask supervision improves garment localization and fitting consistency. (b) Texture rectification improves high-frequency garment details while preserving the try-on structure.

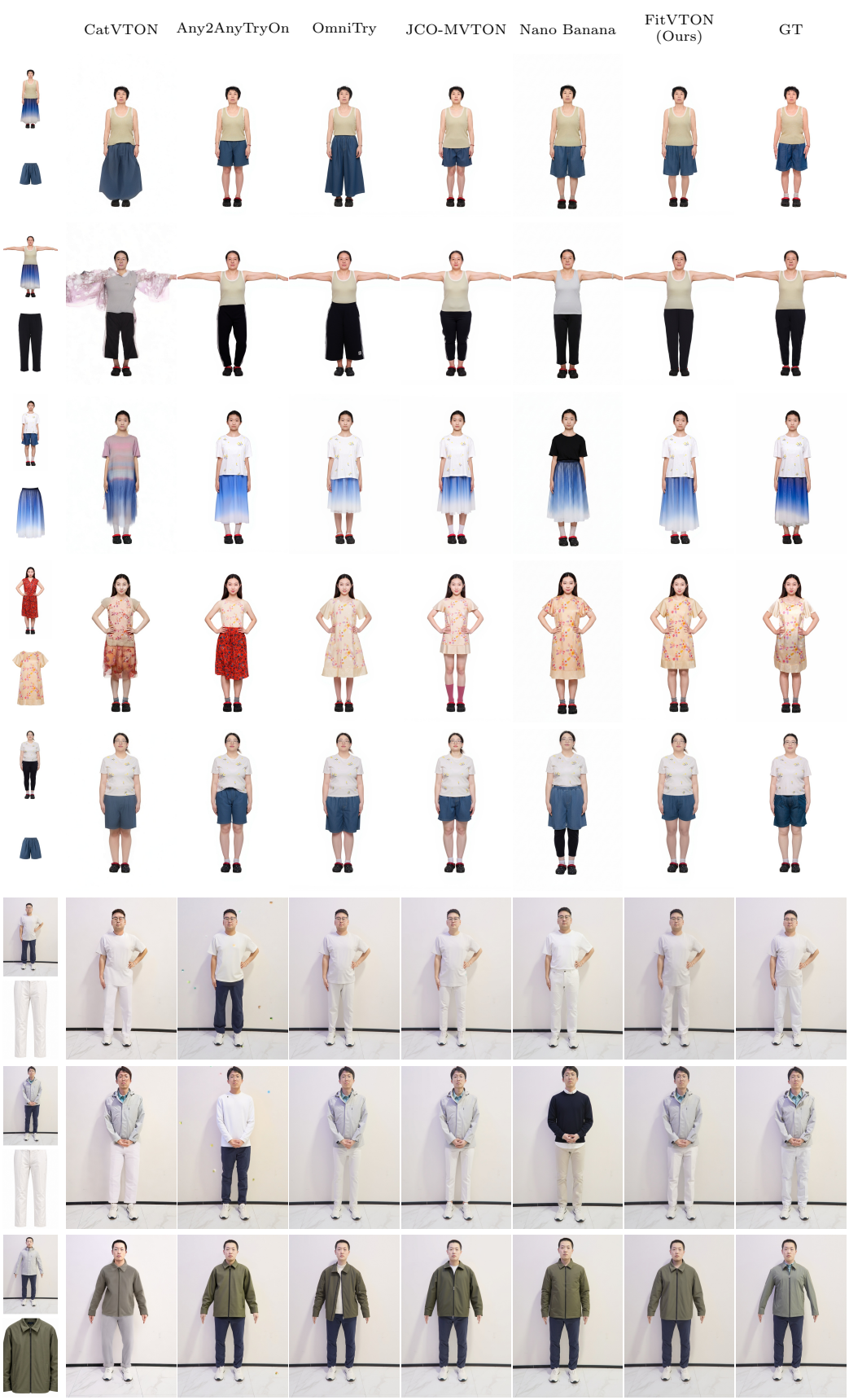

Figure 13: **Additional fit-oriented comparison with baseline methods.** We provide more virtual try-on examples comparing FitVTON with representative baseline methods. These examples complement the main-paper results by showing garment appearance preservation, body-structure consistency, and fitting behavior across additional samples.



Figure 14: **Additional qualitative comparison with baseline methods.** We provide more virtual try-on examples comparing FitVTON with representative baseline methods. These examples complement the main-paper results by showing garment appearance preservation, body-structure consistency, and texture realism.

Table 8: Heights and Garment-Body Size body levels of the in-the-wild subjects in Figure 15.

Female				Male			
Person	Height (cm)	Height level	Body-size level	Person	Height (cm)	Height level	Body-size level
Person 1	176	tall	plus-size	Person 8	165	short	slim
Person 2	166	medium-tall	plus-size	Person 9	165	short	plus-size
Person 3	160	medium-short	slim	Person 10	193	tall	medium-slim
Person 4	173	tall	medium-slim	Person 11	185	tall	plus-size
Person 5	168	medium-tall	slim	Person 12	178	medium-tall	slim
Person 6	176	tall	slim	Person 13	191	tall	slim
Person 7	179	tall	slim	Person 14	172	medium-short	medium-slim

D.4 Compatibility with In-the-Wild Images

FitVTON is trained on simulated GarmentCodeVTON triplets and rectified on studio-style try-on data. To examine how well it transfers to unconstrained real photographs, we collect 14 in-the-wild full-body images of 7 female subjects (Persons 1-7) and 7 male subjects (Persons 8-14), captured on red carpets, streets, staircases, event backdrops, and plain studio walls. The subjects wear floor-length gowns, crop tops with jeans, suits, tuxedos, and hoodies, and show natural poses such as hands on the hips and waving. The female images have a resolution of only 360×480. We pair every female subject with two dresses of different lengths and every male subject with two upper garments of different lengths, giving 28 try-on results in total. Dress A is a short shift dress and Dress B is a long sleeveless printed dress, both worn in the one-piece style. Upper C is a short-sleeve T-shirt and Upper D is a long-sleeve button-up shirt, both worn untucked. We use the same checkpoint and inference setting as in Appendix B, with no per-image tuning.

The Garment-Body Size prompts follow the mapping in Appendix A.5. The height level of each subject is obtained from the reported height with Table 4, and the body-size level is estimated visually because tape measurements are unavailable. The four garments come from product images without measurements and are therefore prompted as M-size, *e.g.*, "The person is a tall and slim female. The cloth is an M-size dress garment. The wearing style is one-piece." Table 8 lists the height and the resulting body levels of all subjects, and Figure 15 shows the results.

Preservation of the person and scene. Across all 28 results, FitVTON keeps the identity, hairstyle, pose, lower garments, footwear, and background of the input photograph, including cluttered red-carpet backdrops and outdoor streets. Natural gestures are retained, such as the hands on the hips of Persons 1 and 4 and the raised hands of Persons 11 and 12, and accessories such as the handbag of Person 6 and the backpack straps of Person 9 stay in place. When the input outfit covers the legs, such as the floor-length gowns of Persons 1 and 7, FitVTON synthesizes plausible legs and footwear below the short hem of Dress A. The low-resolution female inputs still yield sharp and coherent try-on images, and the garment appearance, including the print of Dress B and the tie-dye stripes of Upper C, closely follows the reference.

Fit-aware behavior on real bodies. The same garment produces different fitting results on different bodies, consistent with their heights in Table 8. Dress A reaches just below the knee on Person 3 (160 cm), around the knee on Persons 2 and 5 (166 cm and 168 cm), and mid-thigh on Persons 6 and 7 (176 cm and 179 cm). Likewise, Dress B reaches the ankle on Person 3 and mid-calf on the taller Persons 6 and 7. The garment silhouette also adapts to the body build: on the plus-size Persons 1, 2, 9, and 11, the dresses and upper garments follow the fuller torso and waist, while on the slim subjects the same garments hang with a looser drape. For the male subjects, Upper C ends at the hip and Upper D extends below the hip with sleeves reaching the wrist, in agreement with their short and long designs and the untucked style.

These results show that FitVTON transfers from simulated training data to unconstrained real photographs with diverse scenes, outfits, poses, resolutions, and body types, and that its garment-body fitting behavior carries over to these images.



Figure 15: **FitVTON on in-the-wild images.** The first row of each block shows the input person images, with the reported height under each person index (Table 8). The following rows show FitVTON results for the garment in the leftmost column. Dress A and Upper C are short designs, and Dress B and Upper D are long designs.