DRESSING THE CROWD: MULTI-HUMAN VIRTUAL TRY-ON VIA LOW-SUBSTITUTABILITY PREFERENCE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Virtual try-on (VTON) has achieved impressive fidelity in single-person settings, but extending it to multi-human scenes requires resolving person–garment assignments, inter-person occlusions, spatial interactions, and non-target preservation, for which paired supervision is difficult to scale. We formulate Multi-Human VTON and introduce MultiHuman-VTON, a 55K-image dataset and benchmark covering one-to-many *Partial* and many-to-many *Full* editing. We further propose CES-VTON, a low-substitutability preference optimization framework that post-trains an image-editing foundation model using group-relative reward signals. Conventional additive rewards permit a strong criterion, such as source preservation, to compensate for failed editing, enabling reward-hacking solutions. To restrict such compensation, our Hierarchical CES Utility (HCU) applies nested CES aggregation: it first combines source-relative binding gain with source-garment replacement and then couples the editing utility with non-target preservation. Dynamic Fréchet Credibility (DFC) maintains synchronized source and generated feature queues to estimate each candidate's marginal effect on the source–generation feature-distribution gap, asymmetrically penalizing candidates that increase it. Experiments show CES-VTON outperforms specialized and general-purpose baselines in complex multi-human scenes, improving target assignment, garment replacement, and non-target preservation while its components generalize to single-person VTON and multi-image editing.

1 INTRODUCTION

Virtual try-on (VTON) renders target garments while preserving identity and visual realism. Despite diffusion advances, existing VTON methods and benchmarks focus largely on single subjects Morelli et al. (2022); Choi et al. (2021); Cui et al. (2025) or one-to-many garment–person relations Liang et al. (2026). Many-to-many try-on poses additional challenges in multi-reference processing, explicit person–garment binding, occlusion handling, and non-target preservation. As illustrated in Fig. 1, we formulate this as Multi-Human Virtual Try-On. Unlike generic multi-image editing, its correctness strictly requires explicit person-garment assignment, genuine replacement, and localized preservation amid complex interactions.

The validity structure of multi-human try-on is inherently hierarchical. At the editing level, success requires both accurate person-garment binding and sufficient source garment replacement; merely overlaying textures over original clothing constitutes an invalid edit. At the task level, this edit must coexist with preserving non-target people and backgrounds. Thus, binding and replacement dictate editing validity, which, alongside preservation, defines overall task success. Because each factor is indispensable, a large improvement in one should not compensate for a critical failure in another. This low-substitutability makes conventional additive rewards vulnerable to cross-objective compensation.

Supervised VTON methods Lee et al. (2022); Yang et al. (2024) typically rely on paired data. Extending this paradigm to multi-human scenes requires either costly paired targets or inpainting-based pseudo-targets Yang et al. (2026c). However, severe inter-person occlusions make pseudo-target construction unreliable, introducing annotation noise and a mismatch between synthetic supervision

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.

and real inference. To enable unpaired learning, we introduce **MultiHuman-VTON**, a large-scale dataset of 55,308 real-world images with person-level keypoints, descriptions, instance masks, and refined agnostic masks. During training, scenes are combined with independent garment references, eliminating the need for ground-truth edits. Its test split benchmarks one-to-many *Partial* and many-to-many *Full* editing from only source scenes, garments, and spatial assignments.



Figure 1: From conventional Single-Person VTON to Multi-Human VTON. Our formulation supports both partial and full editing settings.

Leveraging the multi-reference editing capability of a pretrained foundation model Labs (2025), we propose **CES-VTON**, an unpaired framework for group-relative diffusion preference optimization Shao et al. (2024); Zheng et al. (2026). CES-VTON realizes low-substitutability preference optimization by designing its reward around the hierarchical validity structure of multi-human try-on. In contrast, standard weighted sums permit improvements in one component to compensate linearly for failures in another. Such compensation creates trivial high-reward solutions: a model may preserve the source scene without executing the requested edit, or copy target-garment textures into the designated region without coherently removing the source clothing.

To suppress these solutions, we draw inspiration from the Constant Elasticity of Substitution (CES) utility function in microeconomics Arrow et al. (1961), which explicitly controls the substitutability among inputs. We adopt a low-substitutability configuration and formulate the task reward as a Hierarchical CES Utility (HCU). HCU first aggregates person–garment binding and source-garment replacement into an editing utility, and combines this editing utility with non-target preservation to obtain the task utility. This nested construction mirrors the task hierarchy and restricts cross-component compensation, preventing strong preservation from concealing binding failure or high target-garment similarity from concealing incomplete replacement.

Although HCU promotes balanced task correctness, its similarity-based signals may still be exploited through mechanical texture copying, producing locally favorable scores at the expense of global realism. We therefore introduce Dynamic Fréchet Credibility (DFC), a source-synchronized, one-sided distributional anchor. Using dynamic source and generated feature queues, DFC estimates each candidate's marginal contribution to the source–generation Fréchet gap. A gap-reducing candidate receives neither an additional bonus nor a penalty, leaving its task advantage unchanged and avoiding an incentive for conservative source reconstruction. In contrast, a gap-increasing candidate has its positive task advantage attenuated and receives an additional one-sided penalty. DFC therefore suppresses locally rewarded but distributionally degrading solutions without directly competing with the requested edit.

Our main contributions are:

- We formulate the Multi-Human Virtual Try-On task and introduce MultiHuman-VTON, a 55K-image dataset and benchmark for evaluating assignment accuracy, localized editing, non-target preservation, and distributional fidelity.
- We propose CES-VTON, an unpaired framework that realizes low-substitutability preference optimization through a Hierarchical CES Utility, organizing binding, replacement, and preservation according to the intrinsic task hierarchy.
- We introduce Dynamic Fréchet Credibility, a source-synchronized, one-sided anchor that estimates each candidate's marginal distributional effect and asymmetrically penalizes globally degrading solutions.

- Extensive experiments establish state-of-the-art performance in multi-human scenes, validate the proposed components, and demonstrate their generalization to generic multi-image editing and single-person VTON.

2 Related Work

2.1 Virtual Try-On

Early methods, such as VITON Han et al. (2018) and CP-VTON Wang et al. (2018), rely on explicit garment warping and image synthesis but often distort textures under complex poses and occlusions. Recent diffusion-based methods improve generation fidelity: IDM-VTON Choi et al. (2024) injects multi-level garment features through separate attention pathways, CatVTON Chong et al. (2025) adopts spatial concatenation for efficient conditioning, and OmniVTON Yang et al. (2025) provides a training-free framework for multi-human try-on. VTEdit-Bench Liang et al. (2026) evaluates one-to-many garment–person relations, whereas MultiHuman-VTON additionally covers many-to-many assignments with a large-scale benchmark and unpaired preference optimization.

2.2 Multi-Image Editing

Reference-guided image models enable flexible customization and editing. DreamO Mou et al. (2025) unifies diverse conditions within a diffusion transformer for multi-concept customization, while Qwen-Image-Edit Wu et al. (2025) combines semantic and reconstructive representations for instruction-guided editing. These models do not explicitly optimize person–garment assignment or localized preservation in multi-human scenes, risking cross-person misbinding, incomplete garment replacement, and unintended background changes. CES-VTON reinforces person–garment assignments and hierarchically optimizes binding, replacement, and preservation.

2.3 Reward-Based Post-Training

Reward-based post-training can adapt pretrained diffusion models to task-specific objectives without paired targets. Diffusion-DPO Wallace et al. (2024) learns from pairwise preferences, GRPO-style methods Shao et al. (2024) optimize online rewards through group-relative comparisons, and DiffusionNFT Zheng et al. (2026) incorporates rewards through a negative-aware forward-process objective. Their effectiveness, however, depends on reward design when multiple criteria interact. Additive rewards permit direct compensation among criteria, allowing success in one, such as preservation, to conceal critical failures in another. Without modifying the optimization backbone, CES-VTON introduces HCU to restrict such compensation and DFC to suppress global feature-distribution degradation, thereby mitigating reward hacking.

3 Methodology

Given a multi-human image I and K garment references $\mathcal{G} = \{g_k\}_{k=1}^{K}$ assigned to K target persons, CES-VTON post-trains FLUX.2-klein Labs (2025) for joint multi-human try-on through unpaired preference optimization. We leverage group-relative preference optimization Shao et al. (2024); Zheng et al. (2026) to enhance person–garment binding and localized editing without paired supervision. However, directly combining multiple rewards permits cross-component compensation (*e.g.*, preserving the source image to inflate fidelity scores while ignoring the edit). To address this issue (Fig. 2), CES-VTON introduces a Hierarchical CES Utility (HCU) that couples binding, replacement, and preservation, together with Dynamic Fréchet Credibility (DFC) to mitigate global distributional degradation.

3.1 Multi-Human Try-On Condition Construction

To construct unpaired training conditions, we build a category-aware training garment pool from VITON-HD Choi et al. (2021) and DressCode Morelli et al. (2022), while using a separate OpenVTON Li et al. (2026) garment pool for evaluation. We extract detailed visual descriptions using a VLM Bai et al. (2025). We then dynamically assign these garments to selected individuals in I and

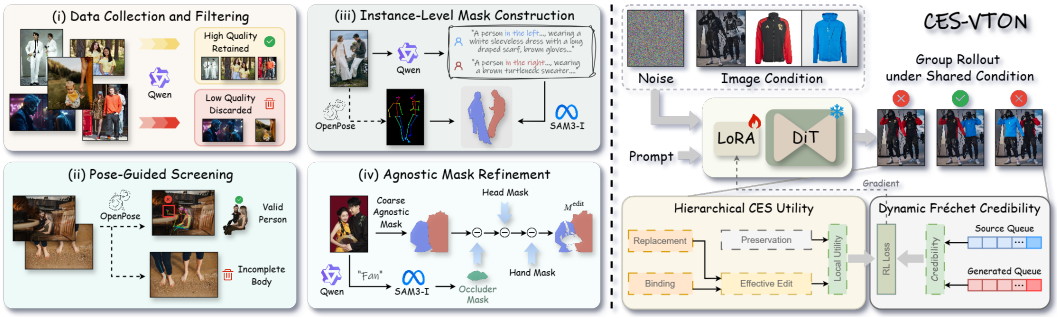

Figure 2: Overview of MultiHuman-VTON construction and CES-VTON. Left: A semi-automated pipeline produces person-level annotations and refined agnostic masks. Right: HCU evaluates binding, replacement, and preservation under a low-substitutability hierarchy, while DFC penalizes distributionally degrading candidates during group-relative preference optimization.

encode their correspondences into spatially aware textual instructions. The final prompt integrates relative person positions, garment slots, and garment descriptions to specify the target mappings. Details of garment-pool construction and instruction formulation are provided in the supplement.

3.2 Hierarchical CES Utility

As discussed above, a valid multi-human edit requires effective local editing, comprising accurate person–garment binding and sufficient garment replacement, together with consistent non-target preservation. Linear reward combinations permit cross-component compensation, allowing a strong component to mask failure in another. Inspired by low-substitutability CES models, we formulate HCU as a nested weak-link-sensitive utility that limits such compensation.

Binding and Replacement Utilities. To reduce background interference and reward leakage, we restrict the editing reward to the precomputed agnostic masks M_k^{edit} introduced during data construction. Let $C(\cdot,\cdot)$ denote frozen-encoder patch-level Chamfer similarity Siméoni et al. (2025). For candidate $\hat{I}_i$, target person k, and garment g_l, the absolute similarity $s_{i,k,l}^{\text{gen}} = C(\hat{I}_i \odot M_k^{\text{edit}}, g_l)$ cannot reliably quantify binding because different garments may be visually similar. We therefore measure the improvement in the assigned garment's relative matching advantage over the source image I, where $s_{k,l}^{\text{src}} = C(I \odot M_k^{\text{edit}}, g_l)$:

$$B_{i,k} = \left(s_{i,k,k}^{\text{gen}} - \max_{l \neq k} s_{i,k,l}^{\text{gen}}\right) - \left(s_{k,k}^{\text{src}} - \max_{l \neq k} s_{k,l}^{\text{src}}\right). \tag{1}$$

Averaging $B_{i,k}$ across targets yields the scene-level binding gain B_i. We calibrate B_i into $Q_i^{\text{bind}} \in (0,1)$ using historical median/MAD statistics Hampel (1974) updated via bounded EMA. The same robust calibration procedure is applied to all subsequent task scores.

To encourage source-garment replacement rather than superficial overlay, we jointly score target-feature acquisition and source-feature removal:

$$R_{i,k}^{\text{rep}} = \frac{1}{2}\left[\left(s_{i,k,k}^{\text{gen}} - s_{k,k}^{\text{src}}\right) + \left(s_k^{\text{self}} - s_{i,k}^{\text{gs}}\right)\right], \tag{2}$$

where $s_{i,k}^{\text{gs}} = C(\hat{I}_i \odot M_k^{\text{edit}}, I \odot M_k^{\text{edit}})$ and $s_k^{\text{self}} = C(I \odot M_k^{\text{edit}}, I \odot M_k^{\text{edit}})$. After calibrating each $R_{i,k}^{\text{rep}}$, their geometric mean defines Q_i^{rep}, reducing compensation between targets.

Preservation and Hierarchical Aggregation. To penalize modifications to unedited bodies or background structures, we compute the Charbonnier loss Charbonnier et al. (1994) between $\hat{I}_i$ and I within the non-target preservation region $M^{\text{keep}} = 1 - \bigcup_{k=1}^{K} M_k^{\text{edit}}$:

$$E_i^{\text{keep}} = \frac{\sum_p M^{\text{keep}}(p)\mathcal{L}_C\left(\hat{I}_i(p), I(p)\right)}{\sum_p M^{\text{keep}}(p) + \epsilon}, \tag{3}$$

where $\epsilon > 0$ ensures numerical stability. Calibrating $-E_i^{\text{keep}}$ yields the preservation utility Q_i^{keep}.

To reflect the task hierarchy while limiting compensation, we use an equally weighted Constant Elasticity of Substitution (CES) function with $\rho = -1$ at each aggregation level. Under this low-substitutability configuration, CES reduces to the harmonic mean:

$$\text{CES}_{-1}(x, y) = \frac{2xy}{x + y + \epsilon}. \tag{4}$$

We first combine binding and replacement into an editing utility and then couple it with preservation:

$$\begin{aligned} U_i^{\text{edit}} &= \text{CES}_{-1}\left(Q_i^{\text{bind}}, Q_i^{\text{rep}}\right), \\ U_i^{\text{task}} &= \text{CES}_{-1}\left(U_i^{\text{edit}}, Q_i^{\text{keep}}\right). \end{aligned} \tag{5}$$

This nested formulation is sensitive to weak components: binding and replacement must jointly establish an effective edit, which is subsequently balanced against non-target preservation. The hierarchy is deliberate: a flat three-way aggregation would treat preservation as interchangeable with the two criteria that determine whether the requested edit is actually executed. By resolving edit validity first, HCU prevents preservation from compensating for a failed edit before the task-level trade-off is considered. For a condition containing only one garment reference, no inter-garment competitor exists; we therefore omit binding and use $U_i^{\text{edit}} = Q_i^{\text{rep}}$. Following Zheng et al. (2026), we convert U_i^{task} into the task advantage A_i^{task} through group-wise relative comparison and centering.

3.3 Dynamic Fréchet Credibility

While HCU promotes balanced region-level correctness, the model may still exploit feature similarities by copying target textures into the edit region. This may introduce seams and structural distortions while obtaining favorable local rewards. We therefore introduce Dynamic Fréchet Credibility (DFC), a source-synchronized, one-sided anchor that penalizes locally favorable but globally degrading candidates.

Inspired by Yang et al. (2026a), we maintain dynamic source ($\mathcal{Q}_r$) and generated ($\mathcal{Q}_f$) feature queues using a frozen AlexNet Krizhevsky et al. (2012), distinct from the Inception-v3 encoder Szegedy et al. (2016) used for benchmark FID, to prevent direct optimization of the evaluation feature space. After updating $\mathcal{Q}_r$ once with the current source feature, we evaluate all candidates against the same queue snapshot. For each $\hat{I}_i$, we compute the Fréchet Distance before and after provisionally replacing the oldest feature in $\mathcal{Q}_f$ with the candidate feature:

$$\Delta\text{FD}_i = \text{FD}_i^{\text{before}} - \text{FD}_i^{\text{after}}. \tag{6}$$

The shared snapshot prevents order-dependent bias and makes ΔFD_i directly comparable across candidates in the same rollout group, so the score reflects each candidate's marginal distributional effect rather than queue-update order. After scoring, we replace the oldest $\mathcal{Q}_f$ entry with a uniformly sampled candidate feature.

A negative ΔFD_i indicates that the candidate exacerbates the global distribution shift. Because directly rewarding positive FD improvements would implicitly encourage trivial reconstruction of the source image, DFC operates as a one-sided constraint. We exclusively penalize negative shifts, converting them into a global credibility score:

$$\begin{aligned} P_i^{\text{DFC}} &= \max\left(\frac{-\Delta\text{FD}_i}{\tau_d s_d + \epsilon}, 0\right), \\ Q_i^{\text{DFC}} &= \exp\left(-P_i^{\text{DFC}}\right), \end{aligned} \tag{7}$$

where τ_d is a temperature parameter and s_d is the EMA-updated historical standard deviation of ΔFD. Consequently, non-degrading candidates maintain full credibility ($Q_i^{\text{DFC}} = 1$), whereas severe global degradation yields exponentially decayed scores.

Finally, we dynamically modulate the local task advantage A_i^{task} using this credibility factor:

$$S_i = \min\left(A_i^{\text{task}}, 0\right) + Q_i^{\text{DFC}} \max\left(A_i^{\text{task}}, 0\right) - \left(1 - Q_i^{\text{DFC}}\right). \tag{8}$$

Method	Upper-Full					Upper-Partial				
	FID↓	KID↓	PA-GAC↑	GC↑	MTEC↑	FID↓	KID↓	PA-GAC↑	GC↑	MTEC↑
CatVTON Chong et al. (2025)	18.371	6.143	0.467	2.864	2.401	9.641	1.808	0.525	3.375	2.755
IDM-VTON Choi et al. (2024)	19.156	7.168	0.505	3.779	2.806	10.535	2.172	0.546	3.805	2.840
OmniVTON Yang et al. (2025)	19.202	7.829	0.481	3.517	2.556	10.174	2.432	0.536	3.527	2.693
OmniVTON++ Yang et al. (2026b)	21.238	8.780	0.484	3.393	2.428	13.000	3.498	0.521	3.362	2.588
DreamO Mou et al. (2025)	99.527	75.964	0.439	3.848	2.080	79.510	57.322	0.516	3.857	2.080
Qwen-Image-Edit-2511 Wu et al. (2025)	40.660	16.300	0.364	2.535	2.068	31.462	9.378	0.397	2.662	2.125
FLUX.2-klein Labs (2025)	16.533	3.845	0.479	3.860	3.054	15.516	2.974	0.523	3.893	2.875
CES-VTON	**13.976**	**2.574**	**0.514**	**3.897**	**3.111**	**8.728**	**0.793**	**0.564**	**3.916**	**3.009**

Table 1: Quantitative comparisons on the MultiHuman-VTON Benchmark under the Upper-Full and Upper-Partial settings.



Figure 3: Qualitative comparisons on upper-body Multi-Human VTON. The top two rows show the *Full* setting, while the bottom two rows show the *Partial* setting. Matching numbers indicate the assigned person–garment pairs.

This leaves negative advantages ungated, attenuates positive ones by global credibility, and explicitly penalizes distributional degradation. Consequently, reduced credibility cannot improve a candidate's score, *i.e.*, $S_i \leq A_i^{\text{task}}$. The resulting S_i is group-wise normalized, clipped, and used for policy optimization. HCU and DFC are used only during post-training; inference retains only the learned LoRA parameters and therefore requires neither reward encoders nor agnostic masks.

4 Experiments

4.1 Main Comparisons

We compare CES-VTON with methods from two categories. The first category comprises inpainting-based VTON approaches, including CatVTON Chong et al. (2025), IDM-VTON Choi et al. (2024), OmniVTON Yang et al. (2025), and OmniVTON++ Yang et al. (2026b). The second category consists of general-purpose multi-image editing models, including DreamO Mou et al. (2025), Qwen-Image-Edit-2511 Wu et al. (2025), and FLUX.2-klein. FID/KID measure distributional fidelity, while PA-GAC, GC, and MTEC assess person-aware garment alignment, garment consistency, and multi-target editing compliance. For a fair comparison, all methods are evaluated on the same source scenes, garment assignments, and output resolution. Single-person VTON baselines are adapted to joint multi-human inference, as detailed in Appendix C.4.

Qualitative Comparison. Figs. 3 and 4 compare upper- and lower-body try-on under *Full* and *Partial* settings. Mask-guided methods preserve backgrounds through explicit spatial constraints but remain limited in multi-human scenes. CatVTON and IDM-VTON, designed for single-person try-on, exhibit texture degradation and incorrect binding, while the training-free OmniVTON and

Method	Lower-Full					Lower-Partial				
	FID↓	KID↓	PA-GAC↑	GC↑	MTEC↑	FID↓	KID↓	PA-GAC↑	GC↑	MTEC↑
CatVTON Chong et al. (2025)	16.025	4.140	0.472	1.778	2.360	10.930	1.194	0.485	2.085	2.662
IDM-VTON Choi et al. (2024)	19.900	4.568	0.470	2.965	2.512	12.828	1.611	0.484	3.012	2.668
OmniVTON Yang et al. (2025)	24.862	10.007	0.472	2.009	2.298	13.170	2.484	0.486	2.222	2.537
OmniVTON++ Yang et al. (2026b)	26.554	9.859	0.469	1.983	2.250	15.964	3.065	0.483	2.166	2.462
DreamO Mou et al. (2025)	98.077	76.749	0.385	1.300	1.999	73.002	48.431	0.412	1.281	2.004
Qwen-Image-Edit-2511 Wu et al. (2025)	47.774	16.904	0.369	1.474	2.037	39.153	10.395	0.388	1.670	2.083
FLUX.2-klein Labs (2025)	20.161	6.268	0.473	3.372	2.750	16.873	4.419	0.483	3.468	2.867
CES-VTON	**15.058**	**2.968**	**0.485**	**3.413**	**2.878**	**10.280**	**0.981**	**0.498**	**3.493**	**2.895**

Table 2: Quantitative comparisons on the MultiHuman-VTON Benchmark under the Lower-Full and Lower-Partial settings.

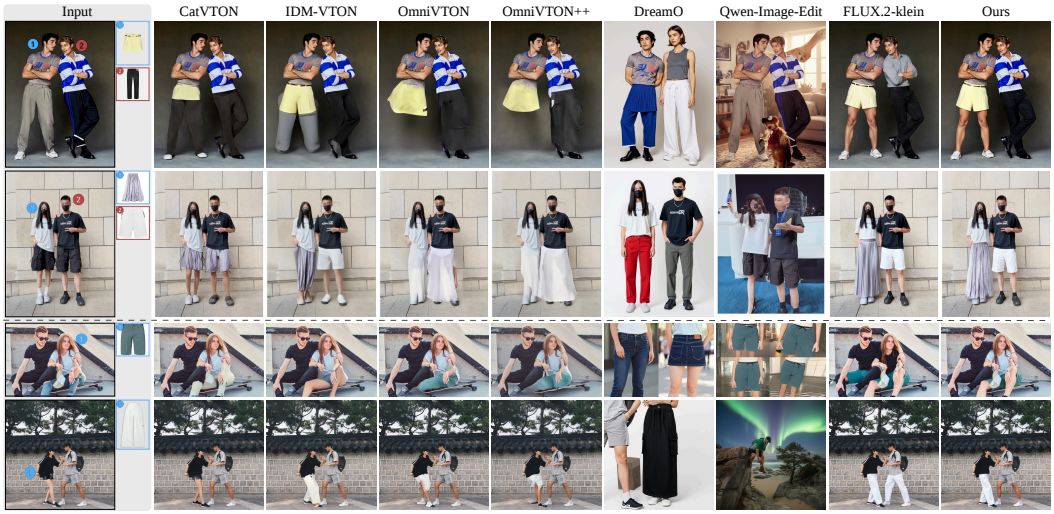

Figure 4: Qualitative comparisons on lower-body Multi-Human VTON. The top two rows show the *Full* setting, while the bottom two rows show the *Partial* setting. Matching numbers indicate the assigned person–garment pairs.

OmniVTON++ struggle with fine-grained garment details and complex occlusions despite supporting multi-human inputs.

General-purpose editors generate rich textures but struggle with source preservation. Lacking explicit spatial constraints, both DreamO and Qwen-Image-Edit prioritize subject customization, inadvertently altering original poses and backgrounds. While FLUX.2-klein offers the strongest baseline visual quality, it remains susceptible to over-editing and category errors (*e.g.*, turning an upper-body garment into a dress in the second row). Conversely, CES-VTON ensures accurate person–garment assignment and high-fidelity replacement while preserving non-target regions, effectively mitigating misbinding and edit leakage.

Quantitative Comparison. Tables 1 and 2 show that CES-VTON achieves state-of-the-art performance across all metrics and settings, particularly in global distributional realism. Under Upper-Full, it improves the FID and KID of the strong FLUX.2-klein backbone from 16.533 and 3.845 to 13.976 and 2.574, respectively. It further achieves an FID of 8.728 under Upper-Partial, with consistent improvements across both lower-body settings.

The results reveal a trade-off between mask-guided methods and general-purpose editors. Explicit spatial constraints provide the former with competitive PA-GAC and MTEC but weaker GC, whereas FLUX.2-klein better preserves garment details but suffers on preservation-sensitive metrics due to over-editing and misbinding. By jointly promoting binding, replacement, and preservation through HCU, CES-VTON overcomes this trade-off and achieves the best GC, PA-GAC, and MTEC across all settings. These results demonstrate a favorable balance among garment replacement, non-target preservation, and global visual consistency.



Figure 5: Qualitative ablation study of the proposed utility components.

Variant	FID↓	KID↓	PA-GAC↑	GC↑	MTEC↑
Base	16.533	3.845	0.479	3.860	3.054
+ Binding	15.867	3.169	0.495	3.868	3.066
+ Replacement	130.912	112.480	0.442	3.902	2.039
+ Preservation (HCU)	17.578	5.524	**0.552**	**3.921**	3.098
Arithmetic Aggregation	15.962	4.055	0.540	3.917	3.083
CES-VTON (+ DFC)	**13.976**	**2.574**	0.514	3.897	**3.111**

Table 3: Progressive ablation on the Upper-Full split of MultiHuman-VTON. Arithmetic Aggregation replaces HCU with a linear reward combination while retaining DFC.



Figure 6: Agreement (%) between the proposed metrics and human-majority rankings, excluding ties (38, 38, and 37 valid pairs for GC, MTEC, and PA-GAC, respectively).

4.2 Ablation Studies

We add reward components to examine their roles. As shown in Fig. 5, the 'Base' model occasionally misassigns garments. '+Binding' improves assignment but may confuse garment categories or mix textures. '+Replacement' strengthens source removal but, without preservation, drifts from the source scene, revealing reward hacking. '+Preservation (HCU)' restores non-edited content and resolves texture mixing, yet may copy-paste flat references into editing regions. 'Arithmetic Aggregation' yields similar artifacts, showing linear combinations permit local similarity to compensate for implausible synthesis. The full 'CES-VTON (+DFC)' suppresses these degeneracies, achieving correct assignment, coherent replacement, and consistent preservation.

Table 3 corroborates these observations. Without preservation, '+Replacement' increases FID and KID to 130.912 and 112.480. Adding preservation yields the highest PA-GAC and GC of 0.552 and 3.921, but its FID of 17.578 reveals copy-paste artifacts despite these local gains. Arithmetic Aggregation exhibits a similar imbalance and remains inferior to CES-VTON in FID, KID, and MTEC. CES-VTON achieves the best FID and MTEC, confirming that HCU limits cross-component compensation while DFC suppresses locally favorable but distributionally harmful solutions.

4.3 Extended Evaluations

Component Generalizability. To evaluate component generalizability, we integrate both HCU and DFC into OmniGen2 Wu et al. (2026) for multi-image editing and CatVTON Chong et al. (2025) for single-person VTON. Following OmniGen2, we evaluate on the Multiple split of OmniContext using Prompt Following (PF), Subject Consistency (SC), and their geometric-mean Overall score. For single-person VTON, we evaluate CatVTON on VITON-HD Choi et al. (2021), reporting FID and KID under the unpaired setting and SSIM under the paired setting. As shown in Table 4, HCU and DFC consistently improve all metrics across both tasks, demonstrating their applicability to different architectures and editing settings.

User Study. To evaluate the alignment of the proposed metrics with human perception, we recruit 56 participants to assess 40 anonymized output pairs evenly sampled from the Upper-Full and

Method	PF ↑	SC ↑	Overall ↑
OmniGen2	7.60	8.34	7.96
OmniGen2 + Ours	**7.76**	**8.45**	**8.10**
Method	FID ↓	KID ↓	SSIM ↑
CatVTON	9.015	1.091	0.870
CatVTON + Ours	**8.648**	**0.788**	**0.879**

Table 4: Component generalizability across multi-image editing (top) and single-person VTON (bottom).



Figure 7: Failure cases under structurally conflicting or incomplete garment conditions.



Figure 8: Flexible person–garment assignment via point- or box-guided attention routing.

Upper-Partial settings. Each participant answers two binary preference questions per pair, covering target garment fidelity and editing-and-preservation compliance, yielding 4,480 judgments in total. Majority votes provide the human pairwise rankings for GC and MTEC, while averaged preference margins induce the PA-GAC ranking. As shown in Fig. 6, GC, MTEC, and PA-GAC achieve agreements of 94.7%, 86.8%, and 78.4%, respectively, demonstrating strong alignment with human perception.

Flexible Assignment Control. Beyond text-based assignment, CES-VTON supports point- or box-guided person–garment control. As shown in Fig. 8, SAM Kirillov et al. (2023) converts each prompt into a person mask paired with the corresponding garment slot at the attention resolution. The mask then performs attention routing: attention between the selected region and its assigned garment slot is retained, while competing garment slots are suppressed within that region and the assigned slot is blocked outside it. We further apply the same routing to the base model. As shown by 'Base + Attn.', attention control alone causes over-editing and texture blending, indicating that spatial routing cannot ensure reliable replacement and preservation. In contrast, 'Ours + Attn.' benefits from the capabilities learned through HCU and DFC, enabling flexible spatial assignment without retraining.

5 Limitations

As shown in Fig. 7, our method remains limited when the input conditions are structurally conflicting or visually incomplete. In the left example, preserving a hand-in-pocket pose becomes implausible when the target garment does not contain a compatible pocket. In the right example, only the back view of the target garment is provided while the person faces forward; the model must therefore infer the unobserved front appearance, which may differ from the actual garment design. Garment-aware pose adaptation and multi-view or 3D garment representations could alleviate these limitations.

6 Conclusion

Diffusion-based VTON has advanced garment transfer, but most methods remain limited to single-person scenes. To address multi-human scenes, we formulated Multi-Human VTON and introduced MultiHuman-VTON, a 55,308-image dataset and benchmark supporting *Partial* and *Full* editing. We proposed CES-VTON, an unpaired low-substitutability preference optimization framework in which HCU hierarchically couples binding, replacement, and preservation, while DFC penalizes distributionally degrading candidates. Experiments establish state-of-the-art performance in complex multi-human scenes and show that the components generalize to single-person VTON and multi-image editing.

AI USE STATEMENT

In this work, we used generative AI tools for translation, language polishing, expression refinement, and the detection of typographical and grammatical errors. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our study concerns human-centered image editing and uses public fashion datasets and web-sourced images, together with the 56-participant user study of anonymized output pairs described in Sec. 4.3. Because the work involves images of people and human evaluation, dataset provenance, privacy, licensing, and participant-study procedures are relevant ethical considerations. Any release of MultiHuman-VTON will be subject to the licenses and usage terms of the underlying sources, with the applicable conditions documented alongside the dataset.

REPRODUCIBILITY STATEMENT

To facilitate verification and extension of our work, we provide detailed descriptions of the model architecture, training settings, and evaluation protocols in the paper and supplementary materials. The complete source code and trained model weights will be made publicly available after publication to support reproducibility and further research.

REFERENCES

Kenneth J Arrow, Hollis B Chenery, Bagicha S Minhas, and Robert M Solow. Capital-labor substitution and economic efficiency. *REStat*, pp. 225–250, 1961.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *IEEE ICIP*, volume 2, pp. 168–172. IEEE, 1994.

Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *CVPR*, 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 *ECCV*, 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. In *ICLR*, 2025.

Aiyu Cui, Jay Mahajan, Viraj Shah, Preeti Gomathinayagam, Chang Liu, and Svetlana Lazebnik. Street tryon: Learning in-the-wild virtual try-on from unpaired person images. In *WACV*, pp. 8235–8239, 2025.

Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *CVPR*, pp. 7543–7552, 2018.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pp. 4015–4026, 2023.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 25, 2012.

Black Forest Labs. Flux.2. https://github.com/black-forest-labs/flux2, 2025.

Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *ECCV*, pp. 204–219. Springer, 2022.

Jin Li, Tao Chen, Kai Wen, Siqi Yin, Shuai Jiang, Weijie Wang, Jingwen Luo, and Chenhui Wu. Openvton-bench: A large-scale high-resolution benchmark for controllable virtual try-on evaluation. *arXiv preprint arXiv:2601.22725*, 2026.

Xiaoye Liang, Zhiyuan Qu, Mingye Zou, Jiaxin Liu, Lai Jiang, Mai Xu, and Yiheng Zhu. Vtedit-bench: A comprehensive benchmark for multi-reference image editing models in virtual try-on. In *ECCV*, 2026.

Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *CVPR*, pp. 2231–2235, 2022.

Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. Dreamo: A unified framework for image customization. In *SIGGRAPH Asia*, pp. 1–12, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pp. 2818–2826, 2016.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *CVPR*, pp. 8228–8238, 2024.

Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, pp. 589–604, 2018.

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

Chenyuan Wu, Jiahao Wang, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, et al. Omnigen2: Towards instruction-aligned multimodal generation. In *CVPR*, pp. 21964–21975, 2026.

Jiawei Yang, Zhengyang Geng, Xuan Ju, Yonglong Tian, and Yue Wang. Representation fr\'echet loss for visual generation. *arXiv preprint arXiv:2604.28190*, 2026a.

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 *ECCV*, 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 *ICCV*, pp. 16702–16711, 2025.

Zhaotong Yang, Yong Du, Shengfeng He, Yuhui Li, Xinzhe Li, Yangyang Xu, Junyu Dong, and Jian Yang. Omnivton++: Training-free universal virtual try-on with principal pose guidance. *arXiv preprint arXiv:2602.14552*, 2026b.

Zhaotong Yang, Ying Tai, Jiahui Zhan, Yu Zheng, Jianjun Qian, and Jian Yang. Orthotryon: Geometric orthogonalization for conflict-free unified fashion generation. 2026c.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *ICLR*, 2026.

APPENDIX: SUPPLEMENTARY MATERIAL

A DATASET CONSTRUCTION DETAILS

A.1 IMAGE COLLECTION

We collect real-world multi-human fashion images from public web sources using automated retrieval queries. Each query combines a multi-person subject descriptor (*e.g.*, couples, best friends, duos, or groups of friends) with fashion attributes such as visual style, garment category, outfit description, color theme, and scene context. Representative templates include *[style] [subject] [outfit description]*, *[subject] wearing [garment item]*, and *[style] [scene] [group descriptor]*, with optional terms emphasizing real-person, multi-human, and full-body imagery.

A.2 GARMENT POOL AND ASSIGNMENT CONSTRUCTION

We construct the training garment pool from the official flat-lay garment images of VITON-HD Choi et al. (2021) and DressCode Morelli et al. (2022). VITON-HD provides upper-body garments, while DressCode covers upper-body, lower-body, and dress categories. For evaluation, we use a separate OpenVTON pool Li et al. (2026), so training and test garments come from different dataset sources. All garments are organized by their annotated categories.

For each garment, Qwen2.5-VL-7B-Instruct Bai et al. (2025) generates a concise English noun phrase describing reliable attributes, including color, garment type, sleeve length, neckline, pattern, fabric appearance, and distinctive details. It is instructed to describe only the garment and exclude the background, composition, mannequin, and other irrelevant content. The resulting category labels and descriptions are stored with the images for condition sampling and prompt construction.

Following the *Full* and *Partial* definitions in the main paper, we determine the target people eligible for the selected clothing category and sort them by their display indices. Category-compatible garments are sampled without replacement from the corresponding training or evaluation pool and assigned in the same order. Reindexing the selected people by this order, target k has a unique assigned garment g_k and multimodal garment slot.

A.3 DATASET STATISTICS AND SPLIT INTEGRITY

MultiHuman-VTON contains 55,308 real-world images, including 52,266 training images and 3,042 test images. Table 5 summarizes the distribution of group sizes across the two splits, with most images containing two or three valid persons after pose-guided filtering. We additionally examine the separation between the training and test sets to reduce the risk of cross-split duplication.

Table 5: Distribution of valid person counts in the training and test splits of MultiHuman-VTON after pose-guided filtering.

Statistic	Train	Test
Images	52,266	3,042
Two-person images	40,195	2,780
Three-person images	10,076	200
Four or more persons	1,995	62

A.4 PROMPT CONSTRUCTION

Each prompt begins with the global instruction:

> Replace the clothing of the specified people in the source image with the provided garments. Preserve the background, faces, bodies, and people who are not specified. Use each provided garment only for its assigned person.

For each target person, we combine its annotated spatial position with its discriminative person description. Spatial annotations such as *left*, *right*, *center*, *far left*, *center-left*, and n*-th from left* are

converted into natural-language references. When no valid spatial phrase is available, we use the explicit identifier `person_`k. The instruction for garment slot j is:

> Garment slot j is assigned to [person reference]. Replace only the [upper-body/lower-body] clothing of this person with garment slot j: [garment description].

Only target people appear in the slot-specific clauses; all remaining people are protected by the global preservation instruction.

A.5 Data Provenance, Privacy, and Release

MultiHuman-VTON is constructed from publicly accessible real-world multi-human fashion images. We plan to publicly release all images and annotations for which redistribution is permitted, subject to the applicable licensing conditions. To mitigate privacy risks, we will apply face anonymization to distributable images where appropriate.

B Evaluation Metric Details

B.1 PA-GAC Definition

Person-Anchored Garment Assignment Consistency (PA-GAC) jointly evaluates assigned-garment correspondence, target-person consistency, non-target-person preservation, and background preservation. Let $\mathcal{P}$ contain all detected people, with $\mathcal{T}_v$ and $\mathcal{N}_v$ denoting the valid target and non-target subsets. We encode each region into an ℓ_2-normalized global representation $\phi(\cdot)$ using a frozen DINOv3 ViT-S/16 Siméoni et al. (2025) and define $c(X, Y) = \phi(X)^\top \phi(Y)$. Here, M_k^{person} is the individual mask of person k. Preservation similarities are linearly rescaled from $[-1, 1]$ to $[0, 1]$ using $(c + 1)/2$.

Assigned-Garment Correspondence. For each target person k, the same edit mask M_k^{edit} is applied to the source I and generation $\hat{I}$. If the edit mask is unavailable, the individual mask M_k^{person} is used. Pixels outside the mask are replaced with a constant background, and the masked regions are cropped to the same bounding box before encoding. We compute

$$\begin{aligned} s_{k,l}^{\text{gen}} &= c\Big(\hat{I} \odot M_k^{\text{edit}}, g_l\Big), \\ s_{k,l}^{\text{src}} &= c\big(I \odot M_k^{\text{edit}}, g_l\big). \end{aligned} \tag{9}$$

We use a source- and competitor-anchored reference:

$$a_k = \max\left(s_{k,k}^{\text{src}}, \max_{l \neq k} s_{k,l}^{\text{gen}}\right). \tag{10}$$

When $K = 1$, the competing-garment term is omitted and $a_k = s_{k,k}^{\text{src}}$. The normalized assignment score is

$$Q_k^{\text{gar}} = \text{clip}\left(\frac{[s_{k,k}^{\text{gen}} - a_k]_+}{\max(1 - a_k, \epsilon)}, 0, 1\right), \tag{11}$$

where $[x]_+ = \max(x, 0)$ and $\epsilon > 0$. The image-level garment score Q^{gar} averages Q_k^{gar} over $k \in \mathcal{T}_v$. It is high only when the generated region matches the assigned garment more strongly than both the source region and every competing garment.

Target-Person Consistency. Using the same masked-crop procedure, Q_k^{tp} is the rescaled cosine similarity between $I \odot M_k^{\text{person}}$ and $\hat{I} \odot M_k^{\text{person}}$, reflecting feature-level identity and pose consistency. Averaging over $k \in \mathcal{T}_v$ gives Q^{tp}.

Non-Target-Person Preservation. Similarly, Q_j^{nt} is the rescaled cosine similarity between the source and generated regions of each $j \in \mathcal{N}_v$. Their average gives Q^{nt}, which is omitted when $|\mathcal{N}_v| = 0$.

Background Preservation. The background mask is $M^{\text{bg}} = 1 - \bigcup_{j \in \mathcal{P}} M_j^{\text{person}}$. Unlike person regions, the masked background retains the full-image layout. Its preservation score Q^{bg} is the rescaled cosine similarity between $I \odot M^{\text{bg}}$ and $\hat{I} \odot M^{\text{bg}}$.

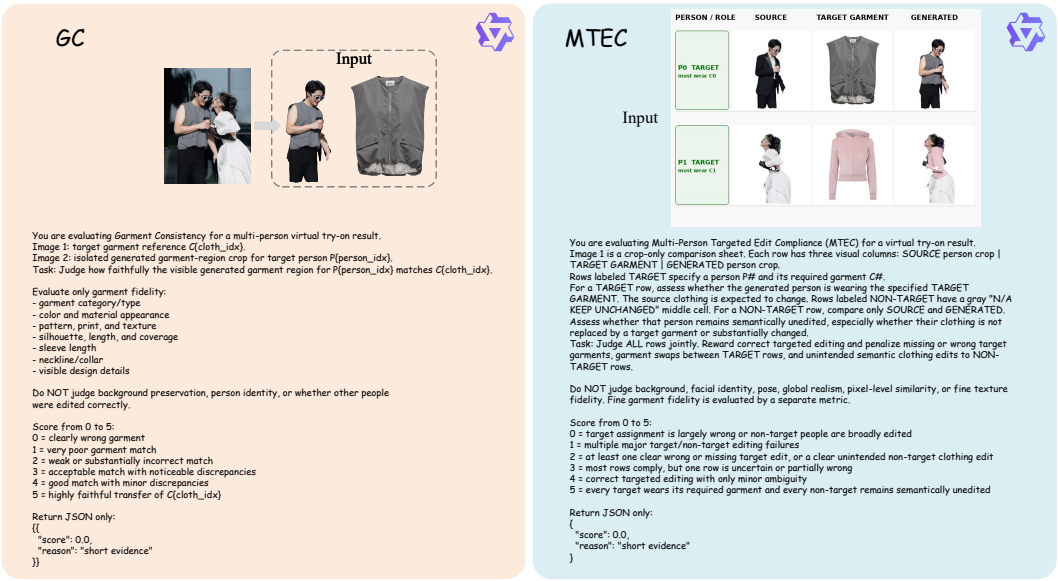

Figure 9: Visual inputs and user prompts for the reference-aware semantic metrics.

PA-GAC Aggregation. Let $\mathcal{R}_{\mathrm{PA}} = \{\mathrm{gar}, \mathrm{tp}, \mathrm{bg}\}$ denote the available component labels, augmented with nt when $|\mathcal{N}_v| > 0$. The image-level score is

$$\text{PA-GAC}(I, \hat{I}) = \frac{1}{|\mathcal{R}_{\mathrm{PA}}|} \sum_{r \in \mathcal{R}_{\mathrm{PA}}} Q^r. \tag{12}$$

This formulation naturally accommodates *Full* settings, for which the non-target-person component is unavailable. We report the mean PA-GAC over all valid evaluation samples.

B.2 Reference-Aware Semantic Metrics

As illustrated in Fig. 9, we use frozen Qwen2.5-VL-7B-Instruct Bai et al. (2025) evaluators with metric-specific visual inputs and prompts for Garment Consistency (GC) and Multi-Human Targeted Edit Compliance (MTEC). Each evaluator returns a JSON object containing a score from 0 to 5 and a short justification; only the score is retained.

Garment Consistency. For each target person k, the evaluator receives the assigned garment g_k and generated garment-region crop $\hat{I} \odot M_k^{\mathrm{edit}}$. The crop uses the same mask and fallback rule as PA-GAC. The prompt on the left of Fig. 9 evaluates garment category, appearance, structure, and visible design details while excluding unrelated regions. Image-level GC is the arithmetic mean over valid targets; we then report its evaluation-set mean.

Multi-Human Targeted Edit Compliance. MTEC receives a crop-only comparison sheet containing the role, source crop, target garment, and generated crop of each person. Target rows specify the required person–garment assignment, whereas non-target rows use an N/A–Keep Unchanged indicator. Within each row, the same individual mask is used for the source and generated person crops. The prompt on the right of Fig. 9 jointly evaluates correct assignment, complete source-garment replacement, and non-target-person preservation. We report the evaluation-set mean of the image-level scores.

C Implementation Details

C.1 Training Configuration

We fine-tune FLUX.2-klein Labs (2025) with LoRA using DiffusionNFT Zheng et al. (2026), while freezing the backbone. HCU uses a frozen DINOv3 ViT-S/16 Siméoni et al. (2025) at 224×224,

Table 6: Implementation configuration used for CES-VTON.

Configuration	Setting
Backbone	FLUX.2-klein
Fine-tuning	LoRA
LoRA rank / alpha	128 / 128
Optimizer	AdamW
Learning rate	3×10^{-4}
Optimization steps	20
Rollout candidates	24
Denoising steps	4
Assignment sampling (Full / Partial)	0.7 / 0.3
HCU encoder	DINOv3 ViT-S/16
HCU encoder resolution	224×224
DFC encoder	AlexNet
DFC queue capacity	1,024 per queue
Fréchet approximation	Diagonal
Numerical precision	bfloat16
Hardware	4×RTX A6000

whereas DFC uses a frozen AlexNet Krizhevsky et al. (2012) and two 1,024-entry feature queues with the diagonal Fréchet approximation. Each condition produces 24 rollout candidates with four Euler denoising steps. Other settings are summarized in Table 6.

C.2 Resolution Bucketing

To preserve diverse aspect ratios while enabling efficient batched training, we use the following resolution buckets, expressed as (H, W):

$$\begin{aligned}\mathcal{R} = \{&(768, 512), (768, 576), (768, 640), (768, 704),\\ &(768, 768), (896, 768), (1024, 768), (512, 768),\\ &(576, 768), (640, 768), (704, 768)\}.\end{aligned} \tag{13}$$

For each source image, we select the bucket closest in aspect ratio and overall resolution, resizing images and masks with bicubic and nearest-neighbor interpolation, respectively. Garment references are scaled to the bucket height with a fixed $3{:}4$ aspect ratio and a width aligned to 16 pixels. To avoid cross-sample padding, samples are batched by identical spatial resolutions and person–garment slot counts.

C.3 Inference Configuration

At inference, HCU, DFC, and all reward statistics are discarded; only the trained LoRA parameters are retained. We use the same preprocessing, resolution-bucket selection, and four-step Euler denoising as the base model. Garment references are ordered by target indices and formatted with the training prompt template. Standard generation therefore requires neither agnostic masks nor auxiliary reward-network evaluation.

C.4 Baseline Adaptation Protocol

To adapt single-person VTON baselines, including CatVTON Chong et al. (2025) and IDM-VTON Choi et al. (2024), to multi-human scenes, we adopt a unified joint-inference protocol rather than sequentially editing individual persons. Specifically, we construct a scene-level agnostic mask by taking the union of the agnostic masks of all target persons. All assigned garment references are provided jointly, together with the same spatially aware prompt used by our method, which specifies the correspondence between each target person and their assigned garment. The baselines then perform a single joint inference run on the complete scene using the unified mask. This protocol avoids sequential error accumulation and ensures that the compared methods receive consistent editing regions and person–garment assignment instructions.

Table 7: Inference efficiency with one and two garment references under identical resolution and denoising settings.

	1 Garment		2 Garments	
Method	MACs (T)↓	Time (s)↓	MACs (T)↓	Time (s)↓
IDM-VTON	280.15	14.08	381.28	19.07
OmniVTON	563.35	20.01	700.24	24.37
FLUX.2-klein	138.29	5.97	179.13	7.89
Ours	138.91	6.05	179.94	8.03

Table 8: Comparison of optimization strategies and HCU substitutability on Upper-Full. GRPO and the final DiffusionNFT model both use $\rho = -1$; the remaining DiffusionNFT variants change only ρ.

Optimizer	HCU setting	FID↓	KID↓	PA-GAC↑	GC↑	MTEC↑
Base model	–	16.533	3.845	0.479	3.860	3.054
GRPO Shao et al. (2024)	$\rho = -1$	14.316	2.991	0.502	3.891	**3.113**
DiffusionNFT	$\rho = 1$	15.962	4.055	0.540	**3.917**	3.083
DiffusionNFT	$\rho \to 0$	15.544	3.991	**0.555**	3.904	3.097
DiffusionNFT	$\rho \to -\infty$	15.703	3.523	0.511	3.893	3.108
DiffusionNFT (Ours)	$\rho = -1$	**13.976**	**2.574**	0.514	3.897	3.111

D Inference Efficiency

HCU and DFC are used only during training, and the learned LoRA parameters can be merged into the backbone. Thus, the method introduces no inference-specific module and retains nearly the same efficiency as FLUX.2-klein. As shown in Table 7, the measured difference is below 1.8% in both MACs and inference time for one- and two-garment inputs. Moving from one to two garments increases MACs by about 30% and time by 32–33% due to the additional visual-conditioning tokens. The resulting model remains substantially more efficient than the specialized baselines, reducing inference time by approximately 57–58% relative to IDM-VTON and using about one quarter of OmniVTON's MACs. We measure batch-one inference time after a fixed warm-up period and average it over the complete efficiency test subset.

E Optimization and CES Substitutability

We compare the policy optimizer and the substitutability controlled by ρ. The same value of ρ is used at both levels of HCU. Under equal weights, $\rho = 1$, $\rho \to 0$, $\rho = -1$, and $\rho \to -\infty$ correspond to arithmetic, geometric, harmonic, and minimum aggregation, respectively. All preference-trained variants use the same backbone, training data, LoRA configuration, rollout budget, reward components, and DFC settings.

Optimization strategy. As Table 8 shows, applying GRPO with HCU and DFC improves every metric over the base model, indicating that the proposed preference signal is not specific to one optimizer. Under the same $\rho = -1$ setting, DiffusionNFT further improves FID, KID, PA-GAC, and GC; GRPO is only 0.002 higher in MTEC. We therefore use DiffusionNFT for its better overall trade-off.

CES substitutability. Arithmetic ($\rho = 1$) and geometric ($\rho \to 0$) aggregation achieve the highest GC and PA-GAC, respectively, but substantially worsen KID. This indicates that more substitutable utilities can favor local metric gains at the expense of distributional quality. At the other extreme, minimum aggregation ($\rho \to -\infty$) is governed entirely by the weakest component and can be sensitive to noisy reward estimates. Harmonic aggregation ($\rho = -1$) yields the best FID and KID while retaining competitive task metrics, providing the most balanced configuration.

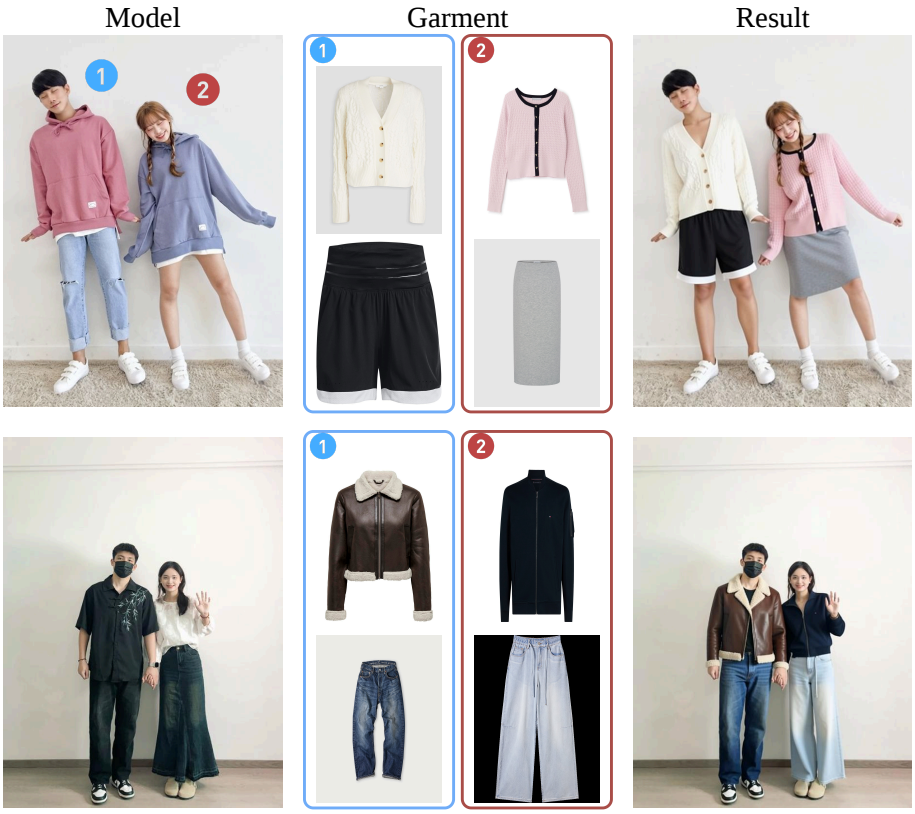

Figure 10: Qualitative results of multi-garment editing.

F MULTI-GARMENT EDITING

Fig. 10 presents additional results in which upper- and lower-body garments are assigned simultaneously. For each example, the garment references are grouped by their target person to make the person–garment correspondence explicit. The results demonstrate that our method can jointly transfer both garment categories while preserving their assignments and the remaining scene content.