DUALFASHION: DUAL-SIDED GARMENT-CONDITIONED MULTI-VIEW FASHION MODEL GENERATION VIA HIERARCHICAL STRUCTURED INPAINTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Most virtual try-on methods edit a supplied full-body person image, while garment-driven virtual dressing typically generates a single image from a frontal garment reference. Neither setting directly supports creating a coherent multi-view fashion presentation from complementary garment evidence. We introduce DualFashion, a framework for dual-sided garment-conditioned multi-view virtual dressing. Given paired front and back garment photographs and a head reference, DualFashion directly generates an ordered set of natural-presentation, front, side, and back views, without requiring a target body image, pose, or scene. The task requires accurate routing of front- and back-side garment details, consistent garment and subject appearance across views, and sufficient spatial capacity for fine-grained details. We address these challenges through hierarchical structured inpainting. A shared-canvas joint stage first synthesizes a coordinated four-view set within a single denoising process. A set-conditioned view-wise stage then reconstructs each view at higher resolution using the complete intermediate set together with the original garment and head references, improving local details while retaining the set-level context established by joint generation. To support this task, we introduce DVF-6K, a dataset of 6,011 product groups across 33 categories, pairing front and back retail garment photographs with natural-presentation, front, side, and back on-body views. Experiments demonstrate improved visual quality, garment fidelity, identity preservation, and viewpoint correctness, while the user study shows improved perceived cross-view consistency. The generated views can also serve as sparse conditions for downstream fashion-video synthesis.

1 INTRODUCTION

Online fashion retail requires complementary natural-presentation, front, side, and back views to display garment appearance and side-specific details. Existing virtual try-on methods (Fig. 1(a)) edit the clothing region of a supplied person image, whose identity, body layout, pose, and scene are already determined Han et al. (2018); Choi et al. (2021); Lee et al. (2022); Morelli et al. (2023); Choi et al. (2024). Garment-driven virtual dressing (Fig. 1(b)) removes this dependency and directly synthesizes a wearer from garment conditions Shen et al. (2025); Li et al. (2025), but existing approaches typically use a single frontal garment photograph and generate one image at a time. Consequently, a single garment reference provides only one-sided product evidence, while independently generated views may drift in garment structure, texture, subject appearance, and viewpoint. Existing multi-view generation methods mainly address camera-defined novel views, controlled capture, 3D reconstruction, or viewpoint transformation of a supplied person Shi et al. (2023); Liu et al. (2024); Long et al. (2024); Han et al. (2026); Laczkó et al. (2026). They do not directly address garment-driven generation of an ordered set of semantic catalog views from uncalibrated front and back retail photographs without a full-body target person image.

Existing fashion datasets likewise do not provide the supervision required by this task. Virtual try-on datasets pair garments with supplied target-person images, garment-driven datasets are generally organized as independent garment–image pairs, and multi-view datasets commonly rely on controlled

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.

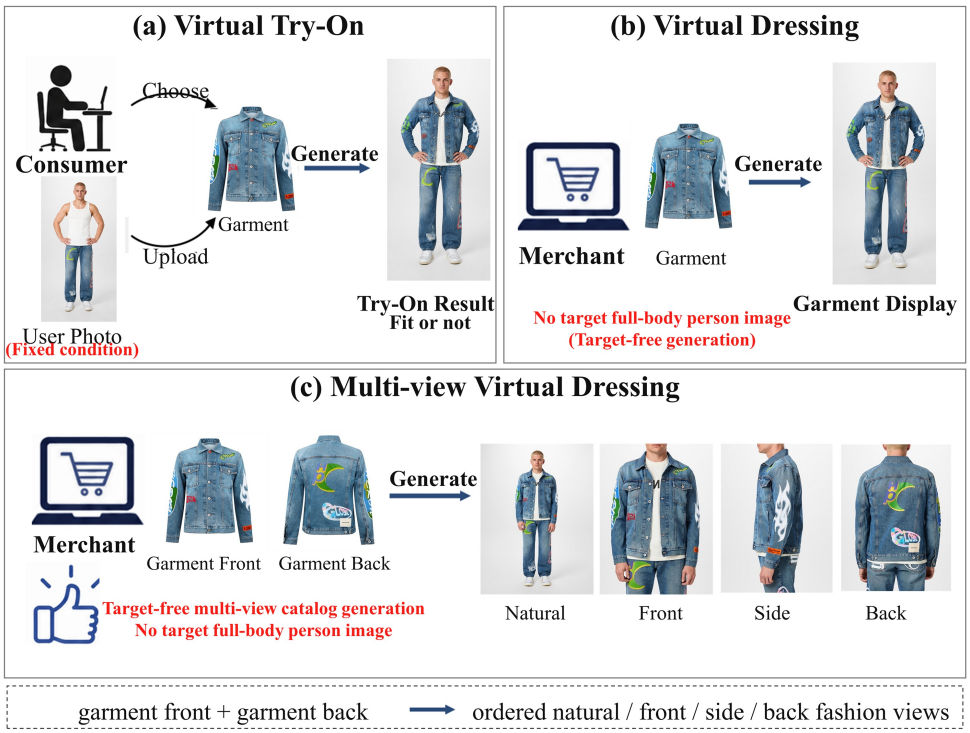

Figure 1: Fashion-generation settings. DualFashion generates ordered natural/front/side/back views from paired garments and a head reference, without a full-body target image.

cameras or synchronized capture. In particular, existing multi-view datasets provide either unstructured views, as in MVG Wang et al. (2025), or calibrated camera views, as in MV-Fashion Laczkó et al. (2026), rather than ordered natural/front/side/back semantic catalog views. We therefore introduce DVF-6K, a task-aligned dataset constructed from 155,412 online-retail images through multimodal garment–model matching, OpenPose-based orientation filtering Cao et al. (2017), and expert manual review. DVF-6K contains 6,011 six-image product groups, comprising 36,066 images across 33 garment categories. Each group links paired front and back garment photographs to a corresponding natural-presentation view and garment-focused front, side, and back on-body views, providing product-level supervision for side-specific garment transfer, semantic-view framing, and cross-view consistency.

Based on this formulation, we introduce DualFashion, a hierarchical framework for generating an ordered four-view fashion presentation from paired front/back garment photographs and a head reference, without requiring a target body image, pose, or scene (Fig. 1(c)). Built on FLUX.1-Fill Black Forest Labs (2024), DualFashion first performs shared-canvas joint inpainting, generating all four semantic views within a single denoising process to coordinate garment evidence, subject identity, and view organization. Since the shared canvas limits the spatial capacity of each output, a set-conditioned view-wise stage subsequently reconstructs each view at higher resolution using the complete intermediate set and the original references. This hierarchical design retains the common view-set context while recovering fine-grained facial, textual, and garment details. Experiments demonstrate improvements in visual quality, garment fidelity, identity preservation, and viewpoint correctness, with perceived cross-view consistency further evaluated in the user study. The generated front, side, and back views can further serve as sparse temporal anchors for fashion-video synthesis Wan et al. (2025), enabling coherent 360-degree presentations.

Our main contributions are:

- This work formulates dual-sided garment-conditioned multi-view virtual dressing, a new setting that directly generates an ordered set of natural-presentation, front, side, and back

views from paired garment photographs and a head reference, without requiring a full-body target person image.

- To support this task, we construct DVF-6K, a task-aligned dataset comprising 6,011 product groups, 36,066 images, and 33 garment categories. Its product-level supervision associates complementary front/back retail garment evidence with four semantic on-body views.
- Our hierarchical set-to-view structured inpainting framework combines shared-canvas joint synthesis for cross-view coordination with complete-set-conditioned high-resolution reconstruction for fine-grained garment and facial details. The resulting views can additionally serve as sparse visual anchors for downstream fashion-video synthesis.

2 Related Work

Virtual try-on and virtual dressing. Image-based virtual try-on edits a supplied person image while preserving the wearer's identity, pose, and scene Han et al. (2018); Choi et al. (2021); Lee et al. (2022); Morelli et al. (2023); Kim et al. (2024); Zhu et al. (2023); Choi et al. (2024); Chong et al. (2025). Pose-guided and multi-view variants support additional body configurations but still require a target person Dong et al. (2019); Wang et al. (2025). Garment-driven methods instead synthesize a wearer directly from garment conditions Chen et al. (2024); Lin et al. (2025); Shen et al. (2025); Li et al. (2025), yet typically generate a single image from frontal garment appearance. In contrast, DualFashion uses paired front/back garment references to generate an ordered and coordinated four-view presentation without a full-body target image.

Fashion datasets and multi-view supervision. Existing fashion datasets differ in their conditioning and view organization. DeepFashion and DeepFashion2 support recognition-oriented tasks rather than conditional generation Liu et al. (2016); Ge et al. (2019); VITON-HD, DressCode, and MPV pair garments with supplied target persons Choi et al. (2021); Morelli et al. (2022); Dong et al. (2019); and IGPair provides independent garment–dressed-image pairs without rear evidence or ordered views Shen et al. (2025). MVG contains multiple views of a supplied person, while MV-Fashion provides calibrated multi-camera sequences Wang et al. (2025); Laczkó et al. (2026). In contrast, DVF-6K links paired front/back retail garments to ordered, uncalibrated semantic on-body views, providing supervision for side-specific garment transfer, role-specific framing, and cross-view consistency.

Joint multi-view generation. Multi-view diffusion primarily targets camera-defined novel-view synthesis and 3D reconstruction Shi et al. (2024; 2023); Liu et al. (2024); Long et al. (2024); Li et al. (2024). Human-specific methods additionally rely on supplied person images, poses, or explicit body geometry He et al. (2025); Dong et al. (2019); Wang et al. (2025); Han et al. (2026). These formulations differ from our task, which generates a new wearer and an ordered set of semantic catalog views directly from uncalibrated garment photographs. DualFashion jointly establishes the complete view set on a shared canvas and then enhances each view under common set-level context.

3 The DVF-6K Dataset

Conventional garment–person pairs do not supervise how complementary front/back garment evidence should be routed across multiple semantic views. We therefore construct DVF-6K as product-level groups linking paired front/back retail garments to ordered natural-presentation, front, side, and back on-body views. This structure directly supports side-specific garment transfer, semantic-view organization, and cross-view consistency. Further details are provided in the supplementary material.

3.1 Task-Aligned Supervision Design

Each DVF-6K example is a six-image product group containing front and back photographs of the same garment and four corresponding on-body views. The paired garment references provide complementary evidence, including silhouettes, graphics, openings, seams, and closures, that cannot be recovered reliably from a frontal image alone. The natural-presentation view retains the primary

Table 1: Fashion datasets by supervision structure. DVF-6K couples paired retail garments with ordered semantic views. Sizes use native units; NVS denotes novel-view synthesis.

Dataset	Category	Garment reference	Supervision unit	View organization	Native scale
IGPair	virtual dressing	front garment	garment–dressed pair	unordered single view	324,857 pairs
MVG	multi-view VTON	front/back clothing	five-view person set	unstructured arbitrary views	1,009 groups; 5,045 images
MV-Fashion	multi-view VTON/NVS	paired flat/catalog	multi-camera sequence	calibrated video views	3,273 sequences; 72.5M frames
DVF-6K (ours)	multi-view virtual dressing	**front/back retail product**	**six-image product group**	**ordered semantic catalog views**	6,011 groups; 36,066 images

fashion pose and styling context, while the front, side, and back views are selected according to dominant body orientation and garment-focused framing. Unlike synchronized or calibrated multi-camera observations, these images represent ordered semantic catalog roles while preserving the pose and composition diversity of real e-commerce imagery. The front and back views primarily correspond to their respective garment references, whereas the side and natural-presentation views reconcile evidence from both. The six images therefore form one structured training example rather than independent garment–person pairs, as illustrated in Figure 6. DVF-6K contains 6,011 product groups and 36,066 images across 33 garment categories, including 3,068 female-presenting and 2,943 male-presenting groups. Table 1 compares its supervision structure with existing pairwise and multi-view fashion datasets.

3.2 Collection, Quality Control, and Splits

We collect 155,412 images from online retail catalog pages and organize them by product identifier. After source consolidation and file validation, 69,105 images in 11,662 product folders remain for candidate construction. A multimodal large language model evaluates within-product garment–model correspondence using color, silhouette, pattern, and local details, while OpenPose Cao et al. (2017) provides body-orientation cues for filtering images and assigning semantic view roles. These automated outputs only assist curation and are subsequently verified by human reviewers for garment correspondence, semantic-role correctness, framing, image quality, and six-image completeness. Further collection and annotation details are provided in the supplementary material. For identity-conditioned training and evaluation, a conservative detector-based filter retains 3,731 groups with clear head crops in the natural-presentation view. These groups are divided into 3,001/430/300 training, validation, and test examples containing 56/9/20 identities, respectively. Identities are mutually exclusive across splits, and all images of the same identity remain in one partition, preventing leakage and enabling evaluation on unseen subjects.

4 Method

4.1 Overview

Dual-sided garment-conditioned multi-view generation requires front- and back-specific garment evidence to be assigned to the appropriate semantic views while preserving garment appearance and subject identity across the complete presentation. Independent generation lacks explicit cross-view coordination, whereas directly generating all views at high resolution imposes a substantial spatial and optimization burden.

Built on FLUX.1-Fill Black Forest Labs (2024), DualFashion addresses these challenges through hierarchical structured inpainting. Shared-canvas joint inpainting first generates an ordered and coordinated four-view set in one denoising process. Set-conditioned view-wise inpainting then reconstructs each view at high resolution using the complete intermediate set and the original references. No full-body target person image, target pose, or target scene is required.

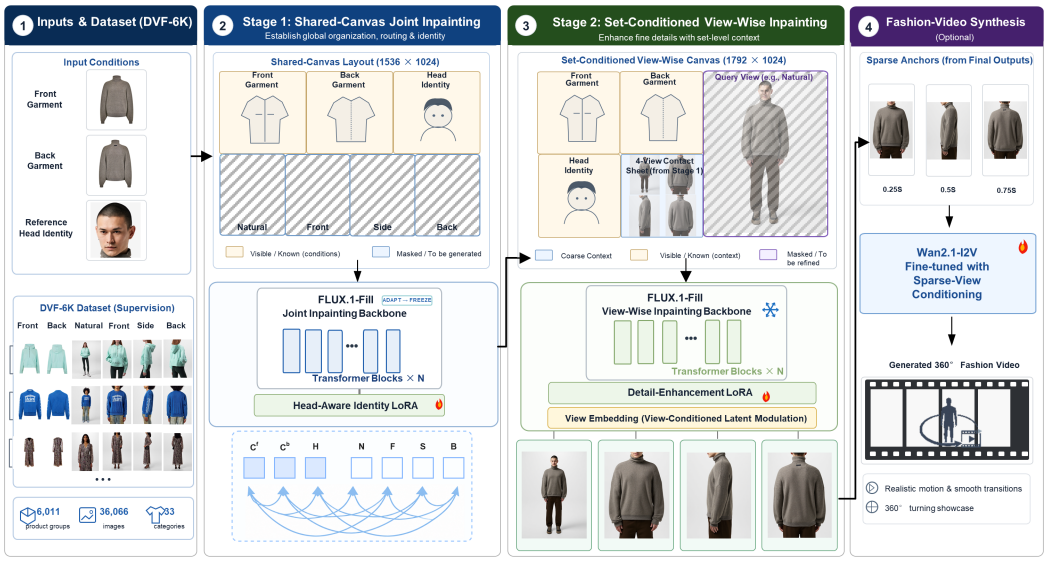

Figure 2: DualFashion overview. Shared-canvas joint inpainting coordinates the complete four-view presentation, after which set-conditioned view-wise inpainting reconstructs each view at high resolution using the original references and the complete intermediate set.

Let $\mathcal{V} = \{n, f, s, b\}$ denote the natural-presentation, front, side, and back roles. Given front and back garment photographs C^f and C^b, a head reference H, and a text condition p, the hierarchy is

$$\begin{aligned} \widetilde{\mathcal{Y}} &= \mathcal{G}_{\text{joint}}(C^f, C^b, H, p), \\ \widehat{Y}^v &= \mathcal{G}_{\text{view}}\Big(C^f, C^b, H, \Psi(\widetilde{\mathcal{Y}}), \widetilde{Y}^v, p_v, e_v\Big), \quad v \in \mathcal{V}, \end{aligned} \tag{1}$$

where $\widetilde{\mathcal{Y}} = \{\widetilde{Y}^v\}_{v\in\mathcal{V}}$ is the jointly generated intermediate set, $\Psi(\widetilde{\mathcal{Y}})$ is its set-level visual context, and p_v and e_v denote the role-specific text condition and learned view embedding.

4.2 Shared-Canvas Joint Inpainting

We organize each training group as a deterministic shared canvas:

$$X = \text{Canvas}\big(C^f, C^b, I; Y^n, Y^f, Y^s, Y^b\big), \tag{2}$$

where I is a neutral tile during initial task adaptation and the head crop H during identity specialization and inference. The condition regions contain the paired garment photographs and identity reference, while the target regions correspond to the ordered natural-presentation, front, side, and back views. The exact canvas layout and resolution are provided in the supplementary material.

During training, all target regions are masked, and the visible references enter through the native spatial-conditioning pathway of FLUX.1-Fill. No additional image encoder or modification to the transformer topology is required. Because all condition and target regions are processed within one transformer sequence and denoising trajectory, each output can attend to both garment references and exchange information with the other evolving views. This shared process encourages side-specific garment routing, semantic-view organization, and consistency in garment structure, texture, and subject appearance.

Training follows the rectified-flow objective Liu et al. (2023); Esser et al. (2024). Given $z_\sigma = (1-\sigma)z_0 + \sigma\epsilon$ and $u_\sigma = \epsilon - z_0$, we optimize

$$\mathcal{L}_{\text{joint}} = \mathbb{E}\Big[w(\sigma)\,\|v_\psi(q_\sigma, \sigma, p) - u_\sigma\|_2^2\Big]. \tag{3}$$

To exploit product groups without reliable head crops, we first adapt the full FLUX.1-Fill transformer without identity conditioning. This phase learns the shared-canvas structure, semantic-view

organization, and dual-sided garment transfer. The adapted backbone is subsequently frozen, and a head-aware identity LoRA is trained on groups with valid head references Hu et al. (2022). Detailed latent construction, identity supervision, and optimization settings are provided in the supplementary material.

4.3 Set-Conditioned View-Wise Inpainting

Although joint synthesis establishes a coordinated presentation, sharing one canvas limits the spatial capacity available to each output. We therefore arrange the intermediate predictions into a compact set-level condition, $\Psi(\widetilde{\mathcal{Y}}) = \mathrm{Grid}(\widetilde{Y}^n, \widetilde{Y}^f, \widetilde{Y}^s, \widetilde{Y}^b)$, and construct a view-specific canvas:

$$R^v = \mathrm{Canvas}_{\mathrm{view}}\Big(C^f, C^b, H, \Psi(\widetilde{\mathcal{Y}}); \widetilde{Y}^v\Big), \quad v \in \mathcal{V}. \tag{4}$$

The visible reference area contains the paired garments, head reference, and complete intermediate set, while only the view-specific target region is regenerated. The corresponding intermediate view preserves the silhouette and composition established during joint synthesis, the complete set provides common cross-view context, and the original references restore fine-grained garment and identity details. This formulation therefore performs set-conditioned high-resolution reconstruction rather than generic interpolation or super-resolution.

The view-wise module is initialized from pretrained FLUX.1-Fill, with its transformer backbone, VAE, and text encoders frozen while a shared detail-enhancement LoRA and role-specific view embeddings are optimized. To reduce the discrepancy between training and inference, training conditions approximate joint-stage intermediate outputs through GT-anchored mask-aware RF-SDE perturbations over the FlowMatch schedule. The training objective is

$$\begin{aligned}\mathcal{L}_{\mathrm{view}} = \mathcal{L}_{\text{view-FM}} + \lambda_{\mathrm{face}}\mathcal{L}^{x_0}_{\mathrm{face}} \\ + \lambda_{\mathrm{grad}}\mathcal{L}^{x_0}_{\mathrm{grad}} + \lambda_{\mathrm{lap}}\mathcal{L}^{x_0}_{\mathrm{lap}},\end{aligned} \tag{5}$$

where the auxiliary terms encourage identity preservation, boundary accuracy, and high-frequency garment details. Detailed source-state construction, masked flow matching, auxiliary losses, and weighting coefficients are provided in the supplementary material.

At inference, each reconstruction is initialized from its corresponding intermediate prediction and refined with mask-aware stochastic correction Singh & Fischer (2024), which recovers local details while retaining the silhouette and composition established by joint synthesis. The four view-wise passes share parameters and the same set-level context and can be executed in parallel. The resulting outputs combine semantic-view organization and cross-view coordination with improved garment details, facial fidelity, and role-specific framing, and can additionally serve as sparse semantic anchors for downstream fashion-video synthesis using a Wan image-to-video model Wan et al. (2025).

5 Experiments

We evaluate DualFashion on overall virtual-dressing quality, multi-view generation, the effectiveness of the hierarchical design, perceptual quality, and downstream sparse-anchor fashion-video synthesis.

5.1 Experimental Setup

Implementation details. The test set contains 300 product groups associated with 20 identities unseen during training. All images belonging to the same identity are restricted to one data split to prevent identity leakage. Training is conducted on three NVIDIA RTX PRO 6000 GPUs, while inference and evaluation use a single RTX PRO 6000 GPU. Unless otherwise specified, both stages use 28 denoising steps and a guidance scale of 30. View-wise reconstruction starts from $\sigma = 0.5$ with a mask-aware correction scale of $\alpha = 0.15$. Additional training and sampling details are provided in the supplementary material.

Baselines and metrics. We compare DualFashion with Magic Clothing Chen et al. (2024), DreamFit Lin et al. (2025), IMAGDressing-v1 Shen et al. (2025), and Qwen-Image-Edit-2511 Wu et al.

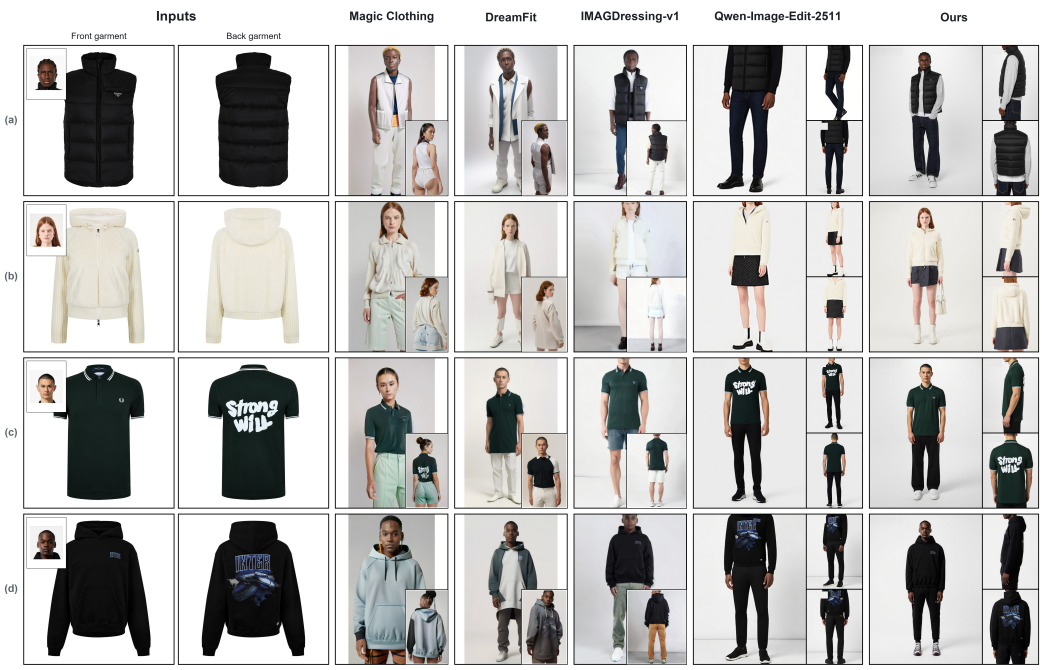

Figure 3: Qualitative comparison. Insets show rear-garment back-prompt outputs for frontal-garment baselines; Qwen-Image-Edit and DualFashion use paired garments.

(2025) using released checkpoints and recommended settings. The first three methods accept one garment image at a time, together with a text prompt and the same head reference through identity-conditioning plugins, whereas Qwen-Image-Edit and DualFashion use paired front/back garments and the head reference. For single-image quantitative evaluation, the first three baselines receive the frontal garment and are evaluated on the natural-presentation output, whose visible garment region is predominantly frontal. For the back-view diagnostic in Fig. 3, we instead provide the rear garment photograph together with a back-view prompt; these results are qualitative only. We measure identity preservation with ArcFace Deng et al. (2019), garment fidelity with garment-region CLIP-I Radford et al. (2021) and DINOv2-I Oquab et al. (2024), and distributional quality with FID Heusel et al. (2017) and KID Bińkowski et al. (2018). Multi-view evaluation further reports view-wise DINOv2-I and DINOv2/CLIP-based garment-side routing accuracy and margins. FID and KID use all 300 test groups, while garment-region metrics use the common 268-sample detected subset.

5.2 Comparison with Existing Methods

Overall virtual-dressing quality. Table 2 shows that DualFashion achieves the best identity similarity, garment fidelity, and distributional quality on the natural-presentation evaluation. The simultaneous gains in FID/KID and garment-region CLIP-I/DINOv2-I indicate that improved realism does not weaken correspondence to the supplied garments. Figure 3 shows that frontal-garment baselines often retain coarse category and color but alter silhouettes, layering, openings, and local patterns. When provided with the rear garment photograph and a back-view prompt, these baselines also frequently produce incorrect orientations or distorted structures. Qwen-Image-Edit is more flexible but may confuse front/back evidence and lose typography or texture, whereas DualFashion better preserves garment construction, local details, and identity.

Multi-view generation. We further compare DualFashion with Qwen-Image-Edit under matched paired garments, head references, and semantic-view prompts. Qwen-Image-Edit often mixes front/back details and varies the wearer, body proportions, or garment structure across views. In contrast, DualFashion produces clearer semantic roles, more accurate side-specific routing, and a more coherent four-view presentation. Table 3 confirms the gains in view-wise garment-reference similarity, garment-side routing, and natural-view identity preservation.

Table 2: Overall virtual-dressing quality. KID is scaled by 10^3; best values are bold.

Method	Arc.↑	CLIP↑	DINO↑
Magic	0.4791	0.5479	0.5973
DreamFit	0.4308	0.5284	0.5524
IMAG.	0.1620	0.6565	0.7249
Qwen	0.1101	0.6310	0.6810
DualFashion	**0.6162**	**0.6811**	**0.7505**

Method	FID↓	KID↓
Magic	107.5452	64.9
DreamFit	97.3419	51.0
IMAG.	67.1349	14.3
Qwen	71.3950	23.7
DualFashion	**45.6513**	**1.7**



Figure 4: Hierarchy ablation. Single-stage HR joint generation is over-smoothed, whereas DualFashion recovers details while preserving set coherence.

Table 3: Multi-view comparison using paired front/back garments. Best values are bold.

Method	Nat. DINO↑	Front↑	Back↑	DINO Δ ↑	CLIP Acc.↑	CLIP Δ ↑	Nat. ID↑
Qwen-Image-Edit	0.6719	0.7285	0.6894	0.1022	51.0	0.0559	0.2932
DualFashion	**0.7452**	**0.8307**	**0.7966**	**0.1705**	**81.3**	**0.1017**	**0.6177**

Table 4: Hierarchy ablation. "Single-stage HR joint" denotes single-stage high-resolution joint inpainting; best values are bold.

Variant	CLIP↑	DINO↑	FID↓	Nat.↑	Front↑	Back↑	Acc.↑	Nat. ID↑
Single-stage HR joint	0.5946	0.6873	71.4906	0.6627	0.7434	0.7241	61.0	0.3122
Full DualFashion	**0.6813**	**0.7509**	**45.6513**	**0.7452**	**0.8307**	**0.7966**	**81.3**	**0.6130**

5.3 Ablation Study

We compare the full hierarchy with a single-stage baseline that directly generates four 768×1024 views on one shared canvas using the same training data, FLUX.1-Fill initialization, output resolution. As shown in Figure 4 and Table 4, the baseline establishes the basic semantic roles but over-smooths faces, garment boundaries, typography, patterns, and structural details. A single oversized denoising process must simultaneously coordinate the complete view set and recover high-frequency content, creating competing global and local objectives that make optimization difficult. DualFashion instead separates compact joint coordination from set-conditioned high-resolution reconstruction, substantially improving garment fidelity, distributional quality, side-specific routing, and identity preservation. These results support the effectiveness of the complete hierarchy in balancing set-level coordination and high-resolution detail reconstruction.

5.4 Application to Sparse-Anchor Fashion Video

We use the final front, side, and back views as fixed sparse conditions for a Wan image-to-video model fine-tuned with sparse-view conditioning Wan et al. (2025). All examples use the same anchor placement, turning prompt, and sampling settings. The model generates 161 frames at 576×720 resolution and 15 fps.

As shown in Figure 5, the semantic anchors determine the main stages of the turning sequence, while the video model synthesizes the intermediate motion. The resulting videos transition smoothly from front to side and back while largely preserving garment appearance and subject identity near the supplied anchors. This demonstrates that the generated view set can also serve as structured conditions for downstream fashion-video synthesis.

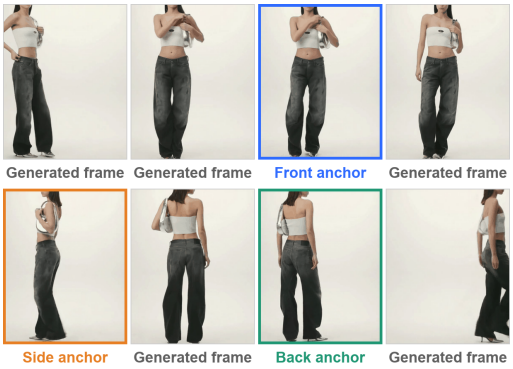

Figure 5: Sparse-anchor video synthesis. Colored borders mark supplied views; unmarked frames are generated.

Table 5: User study mean ratings (1–5; 43 participants). Best values are bold.

Single-image evaluation			
Method	Garment	Identity	Realism
Magic	2.228	2.293	2.660
DreamFit	2.102	2.493	2.526
IMAG.	3.177	2.502	2.935
Qwen	3.000	2.349	2.888
DualFashion	**4.558**	**4.516**	**4.512**
Multi-view evaluation			
Method	Viewpoint	Consistency	Overall
Qwen	3.214	3.107	3.070
DualFashion	**4.577**	**4.456**	**4.465**

5.5 User Study

We conduct a perceptual study on both single-image quality and complete multi-view presentation. The single-image module contains 20 examples and evaluates garment fidelity, identity preservation, and realism across five anonymous methods. The multi-view module contains 10 examples and compares DualFashion with Qwen-Image-Edit in terms of viewpoint correctness, cross-view consistency, and overall presentation quality. Each participant rates five randomly assigned examples from each module using a 1–5 Likert scale. We collect 43 valid responses, totaling 4,515 ratings. In particular, the multi-view module evaluates the generated views jointly as a complete presentation, providing a direct perceptual assessment of whether viewpoint-specific outputs remain coherent as a set.

Table 5 shows that DualFashion receives the highest score under every criterion, with particularly large gains in identity preservation and cross-view consistency. In the multi-view evaluation, the improvements extend to viewpoint correctness and overall presentation quality, indicating that the generated outputs are perceived as a more coherent semantic view set rather than merely a collection of individually plausible images. These perceptual results complement the reference-based metrics: ArcFace evaluates natural-view identity preservation, while view-wise DINO and garment-side routing measure garment-reference fidelity and evidence use. Together, the two evaluation protocols assess both correspondence to the supplied visual evidence and the perceptual quality of the generated presentation as a whole. Detailed score distributions, statistical significance tests, and sensitivity analysis are provided in the supplementary material.

6 Conclusion

We introduced dual-sided garment-conditioned multi-view virtual dressing, a setting that moves beyond person-preserving garment editing by directly generating an ordered fashion presentation from paired front/back garment photographs and a head reference, without requiring a full-body target person image, target pose, or target scene. To support this task, DVF-6K provides product-level supervision that links complementary retail-garment evidence to natural-presentation, front, side, and back on-body views. We further proposed hierarchical set-to-view structured inpainting, in which shared-canvas joint synthesis establishes semantic-view organization, side-specific garment routing, and a common set-level context, while set-conditioned view-wise reconstruction recovers high-resolution garment and facial details using the complete intermediate set. Experiments demonstrate improved visual quality, garment fidelity, identity preservation, and semantic-view correctness, with the multi-view user study further supporting the perceived cross-view consistency of the generated view sets. The generated views can also serve as sparse semantic anchors for downstream turn-around fashion-video synthesis. Together, these contributions provide a practical foundation for scalable generation of garment-faithful, view-consistent, and dynamic digital fashion content.

AI USE STATEMENT

In this work, we used generative AI tools to assist with portions of the software implementation of our methods, including code development and editing. We also used generative AI tools for language polishing and manuscript formatting. All AI-assisted code was reviewed, tested, and validated by the authors, and all AI-assisted manuscript edits were manually reviewed. We did not use generative AI tools for any other tasks requiring disclosure. The authors take full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

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

Black Forest Labs. Flux.1-fill [dev] model card. https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev, 2024.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7291–7299, 2017.

Weifeng Chen, Tao Gu, Yuhao Xu, and Chengcai Chen. Magic clothing: Controllable garment-driven image synthesis. *arXiv preprint arXiv:2404.09512*, 2024.

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

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9026–9035, 2019.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

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.

Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 932–940, 2017.

Seonghee Han, Minchang Chung, Gyeongsu Cho, Kyungdon Joo, and Taehwan Kim. Pose-diverse multi-view virtual try-on from a single frontal image via diffusion transformer. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3310–3319. IEEE, 2026.

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

Xu He, Zhiyong Wu, Xiaoyu Li, Di Kang, Chaopeng Zhang, Jiangnan Ye, Liyang Chen, Xiangjun Gao, Han Zhang, and Haolin Zhuang. Magicman: Generative novel view synthesis of humans with 3d-aware diffusion and iterative refinement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3437–3445, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

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.

Hunor Laczkó, Libang Jia, Loc-Phat Truong, Diego Hernández, Sergio Escalera, Jordi Gonzalez, and Meysam Madadi. Mv-fashion: Towards enabling virtual try-on and size estimation with multi-view paired data. *arXiv preprint arXiv:2603.08147*, 2026.

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

Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3260–3271, 2020.

Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, et al. Era3d: High-resolution multiview diffusion using efficient row-wise attention. *Advances in Neural Information Processing Systems*, 37:55975–56000, 2024.

Xinghui Li, Qichao Sun, Pengze Zhang, Fulong Ye, Zhichao Liao, Wanquan Feng, Songtao Zhao, and Qian He. Anydressing: Customizable multi-garment virtual dressing via latent diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23723–23733. IEEE, 2025.

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

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International conference on learning representations (ICLR)*, 2023.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *International conference on learning representations*, volume 2024, pp. 27676–27697, 2024.

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

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9970–9980, 2024.

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.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, 2024.

Qwen Team. Qwen-image-edit-2511. https://huggingface.co/Qwen/Qwen-Image-Edit-2511, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinhui Tang. Imagdressing-v1: Customizable virtual dressing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6795–6804, 2025.

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023.

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *International conference on learning representations*, volume 2024, pp. 39838–39859, 2024.

Saurabh Singh and Ian Fischer. Stochastic sampling from deterministic flow models. *arXiv preprint arXiv:2410.02217*, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Wan Team. Wan2.2: Open and advanced large-scale video generative models. https://github.com/Wan-Video/Wan2.2, 2025. Official repository.

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

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.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

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.

A Overview

This supplementary material provides implementation, dataset, evaluation, and qualitative details that complement the main paper. We first elaborate on the two-stage hierarchical framework in *Additional Method Details*, including the structured canvas layouts, Stage-I joint set generation, Stage-II set-conditioned reconstruction, view-role routing, auxiliary reconstruction objectives, mask-aware stochastic sampling, optimization settings, and Sparse-View-Conditioned Video Generation. We then describe the product-level organization, collection and quality control, identity-disjoint split, category distribution, and data-governance considerations of DVF-6K in *DVF-6K Dataset Details*. Next, *Evaluation Details* specifies the baseline identity-conditioning interfaces, view-specific prompts, metric preprocessing, and user-study protocol, together with confidence intervals, effect sizes, and statistical significance tests. Finally, *Additional Category Results* presents representative outputs across diverse garment categories, illustrating side-specific garment routing, cross-view coherence, and generalization to both common and long-tail garment types.

B Additional Method Details

B.1 Structured Inputs

Let C^{f} and C^{b} denote the front and back garment photographs, respectively, H the head reference, and $Y = (Y^{\mathrm{n}}, Y^{\mathrm{f}}, Y^{\mathrm{s}}, Y^{\mathrm{b}})$ the ordered natural-presentation, front, side, and back targets. Source images have a resolution of 1425×1425. Garment conditions are resized to 512×512. On-body images are first resized to 512×512 and then center-cropped to 384×512 for the joint stage. The detected head box is converted into a square crop using a context factor of 1.0 and resized to 512×512. Pixel values are mapped to $[-1, 1]$ after the composite canvas is constructed.

Stage-I shared canvas. The 1536×1024 canvas contains three visible 512×512 tiles—C^{f}, C^{b}, and H—in the upper row. The lower row contains four fully masked 384×512 target panels in the fixed order $(Y^{\mathrm{n}}, Y^{\mathrm{f}}, Y^{\mathrm{s}}, Y^{\mathrm{b}})$. Training and inference use identical coordinates and view ordering. All target panels therefore share one transformer sequence and one denoising trajectory, enabling information exchange across views while providing an explicit spatial assignment for each semantic role.

Stage-II view canvas. Each reconstruction pass uses a 1792×1024 canvas. The visible left 1024×1024 region is arranged as a 2×2 grid containing the two garment photographs, the head reference, and a contact sheet of all four Stage-I predictions. The contact sheet preserves the same semantic ordering. Only the right 768×1024 panel is masked and regenerated. A target view is thus reconstructed from the complete intermediate set, rather than from its low-resolution counterpart alone. The set-level condition preserves the common identity and garment structure, while the original references restore details weakened during compact joint generation. Both layouts use the native spatial-conditioning pathway of FLUX.1-Fill Black Forest Labs (2024). The clean latent, masked-condition latent, and downsampled binary mask follow the packing format of the released Fill Transformer.

B.2 Stage I: Joint Set Generation

The evaluated Stage-I model combines a task-adapted FLUX.1-Fill Transformer with a head-aware attention LoRA Hu et al. (2022). During identity specialization, the task-adapted Transformer, variational autoencoder (VAE), and text encoders are frozen. Rank-32 LoRA layers are inserted into the query, key, value, output, and added-attention projections. Separating task adaptation from identity specialization prevents the smaller head-conditioned subset from overwriting the shared-canvas organization and dual-sided garment routing learned by the backbone.

The head reference is supplied directly as a visible canvas tile, without an additional ArcFace, DINO, perceptual, or 3D identity loss. Identity preservation is instead learned from the reconstruction supervision of the head-conditioned groups, avoiding a separate identity objective that may compete with garment fidelity or view organization.

For a clean latent z_0, Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, and flow time σ, training follows the FLUX rectified-flow path

$$z_\sigma = (1-\sigma)z_0 + \sigma\epsilon, \qquad u^\star = \epsilon - z_0. \tag{6}$$

The velocity prediction is supervised with mean-squared error over the generated lower row, with flow times sampled uniformly. At inference, the four panels are generated through one trajectory and sliced at their fixed horizontal boundaries. Compared with independent sampling, this shared trajectory reduces drift in garment structure, identity, and semantic-view assignment.

B.3 Stage II: Set-Conditioned Reconstruction

Stage II is initialized from the original pretrained FLUX.1-Fill Transformer rather than the task-adapted Stage-I Transformer. Its backbone, VAE, and text encoders remain frozen. Training updates a rank-64 attention LoRA and a lightweight target-view router. This initialization retains the general high-resolution image prior of the pretrained model, while the Stage-I contact sheet supplies the task-specific structure and cross-view context.

The router contains one 64-dimensional embedding e_v for each of the four semantic roles. For packed latent token r_i and target-token set Ω, the target-view routing operation is defined as

$$\bar{r}^{v}_{i,c} = \begin{cases} r_{i,c} + e_{v,c}, & i \in \Omega,\ 1 \le c \le 64, \\ r_{i,c}, & \text{otherwise.} \end{cases} \tag{7}$$

The role embedding is broadcast only over the generated panel and does not modify the garment, head, or contact-sheet tokens. It therefore specifies the target semantic role without contaminating the visual conditions. The same LoRA parameters are shared across all four reconstruction tasks.

Training uses target-derived proxy conditions to approximate the intermediate states encountered at inference. Let M_{t} and M_{f} denote the target-panel and face masks, respectively. Given the predicted velocity, the clean-latent estimate is $\hat{z}_0 = z_\sigma - \sigma u_\theta(z_\sigma, \sigma)$. Stage II minimizes

$$\mathcal{L}_{\mathrm{II}} = \mathcal{L}_{\text{view-FM}} + \lambda_{\mathrm{face}}\mathcal{L}^{x_0}_{\mathrm{face}} + \lambda_{\mathrm{grad}}\mathcal{L}^{x_0}_{\mathrm{grad}} + \lambda_{\mathrm{lap}}\mathcal{L}^{x_0}_{\mathrm{lap}}, \tag{8}$$

where the Stage-II configuration uses $(\lambda_{\mathrm{face}}, \lambda_{\mathrm{grad}}, \lambda_{\mathrm{lap}}) = (0.20, 0.05, 0.02)$. The first term is the area-normalized flow-matching error inside M_{t}. For the remaining terms, define $\rho_\epsilon(a) = \sqrt{a^2+\epsilon^2} - \epsilon$, with $\epsilon = 10^{-3}$, and let $\langle\cdot\rangle_M$ denote a mean normalized by the support of mask M. The robust Charbonnier penalty limits the influence of isolated large residuals while retaining gradients around small reconstruction errors. The face term is

$$\begin{aligned} \mathcal{L}^{x_0}_{\mathrm{face}} &= \langle q(\sigma)\, \rho_\epsilon(\hat{z}_0 - z_0)\rangle_{M_{\mathrm{f}}}, \\ q(\sigma) &= \mathrm{clip}\left(\frac{0.25}{\max(\sigma, 10^{-4})}, 0.5, 4\right). \end{aligned} \tag{9}$$

Here M_{f} is the central 75% of the detected face box, rasterized in the target panel and downsampled to latent resolution. The area normalization prevents this small region from being overwhelmed by the much larger garment and background regions. It therefore encourages the clean estimate to retain facial shape and local identity cues. The bounded inverse-noise weight compensates for the scale of the clean-latent error while preventing low-noise samples from dominating the batch.

The structural terms compare signed first- and second-order responses over average-pooled clean-latent pyramids. Let D_s be average pooling by $s \in \{1, 2, 4\}$, with weights $a_s \in \{0.6, 0.3, 0.1\}$; G the normalized Sobel operator; Δ the normalized four-neighbor Laplacian; and M_s the eroded target mask at scale s. Their implemented definitions are

$$\mathcal{L}^{x_0}_{\mathrm{grad}} = \sum_s a_s \left\langle q(\sigma) w_s\, \rho_\epsilon(GD_s\hat{z}_0 - GD_s z_0)\right\rangle_{M_s}, \tag{10}$$

$$\mathcal{L}^{x_0}_{\mathrm{lap}} = \sum_s a_s \left\langle q(\sigma) w_s\, \rho_\epsilon(\Delta D_s\hat{z}_0 - \Delta D_s z_0)\right\rangle_{M_s}. \tag{11}$$

Specifically, $G_x = \frac{1}{8}[-1, 0, 1; -2, 0, 2; -1, 0, 1]$, $G_y = G_x^\top$, and $\Delta = \frac{1}{4}[0, 1, 0; 1, -4, 1; 0, 1, 0]$. The detached weight w_s lies in $[1, 2]$ and increases linearly with the target edge magnitude after normalization by twice its masked mean. The gradient term matches signed first-order changes, directly

supervising garment contours, seams, text strokes, and transitions at the face and hair boundaries. The Laplacian term matches second-order responses and penalizes the loss or invention of finer structures that can have similar average color but different local shape. Their multiscale evaluation covers both narrow texture elements and broader silhouette transitions. Because the operators are signed, the loss penalizes spurious markings as well as missing details; no additional perceptual or recognition network is introduced.

At inference, reconstruction begins from the corresponding Stage-I view at $\sigma_0 = 0.5$. Following stochastic sampling for deterministic flow models Singh & Fischer (2024), the score estimate and noise coefficient are

$$s_\theta(z_\sigma) = -\frac{z_\sigma + (1-\sigma)u_\theta}{\sigma}, \qquad g(\sigma) = 0.15\sqrt{\sigma(1-\sigma)}. \tag{12}$$

For a decreasing step $\Delta\sigma < 0$, the update is

$$z_{\sigma+\Delta\sigma} = z_\sigma + \Delta\sigma\left(u_\theta - \tfrac{1}{2}g(\sigma)^2 s_\theta\right) + g(\sigma)\sqrt{-\Delta\sigma}\,\epsilon. \tag{13}$$

The stochastic correction provides limited exploration around the Stage-I initialization, helping recover textures and local details that may remain over-smoothed under deterministic refinement. To avoid disturbing the visible conditions and the established set-level structure, the correction is restricted to packed tokens covered by the fill mask; unmasked tokens follow the deterministic scheduler update. The coefficient vanishes at both endpoints.

B.4 Optimization and Inference

The Stage-I identity adapter uses LoRA rank/alpha $32/32$, batch size 15 per GPU on three NVIDIA RTX PRO 6000 GPUs, and a learning rate of 5×10^{-5}, and is optimized for 2,500 steps. Stage II uses rank/alpha $64/64$, batch size 12 per GPU on two RTX PRO 6000 GPUs, a LoRA learning rate of 10^{-5}, a router learning rate of 3×10^{-6}, and 3,500 optimizer steps in total.

Both stages use fused AdamW with $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 10^{-4}, unit gradient clipping, and a constant learning-rate schedule after warm-up. Training uses bfloat16, TF32, and gradient checkpointing. Gradient accumulation is one. Neither stage applies condition dropout or stochastic image augmentation, since such transformations could disrupt the fixed correspondence between canvas positions and semantic-view roles. The warm-up lasts 100 steps for Stage I and 10 steps for Stage II. At inference, Stage I and each Stage-II view pass use 28 sampling steps with guidance scale 30; the four Stage-II passes share parameters and can be executed in parallel.

B.5 Sparse-View-Conditioned Video Generation

The downstream demonstration uses `Wan2.2-I2V-A14B` Wan et al. (2025); Wan Team (2025). The final front, side, and back views are inserted as three sparse anchors at normalized temporal positions 0.25, 0.50, and 0.75, respectively. The high- and low-noise DiT branches load their corresponding DualFashion video LoRAs with merge strength 1.0, switching branches at 0.875. We use 50 sampling steps, guidance scale 5.0, and a fixed walking–posing–turning prompt. The resulting video contains 161 frames at 576×720 and is exported at 15 fps.

C DVF-6K Dataset Details

C.1 Product-Level Structured Supervision

DVF-6K is organized as product-level groups rather than independent garment–person pairs. This organization is aligned with the proposed task: the model must associate complementary front- and back-side garment evidence with an ordered set of semantic on-body views. Each group therefore contains six RGB images: front and back retail photographs of the same garment, together with natural-presentation, front, side, and back on-body images. The complete dataset contains 6,011 product groups and 36,066 images.

The four output labels denote semantic catalog roles rather than calibrated camera poses. The natural-presentation image retains the primary fashion pose and styling context, while the front, side,

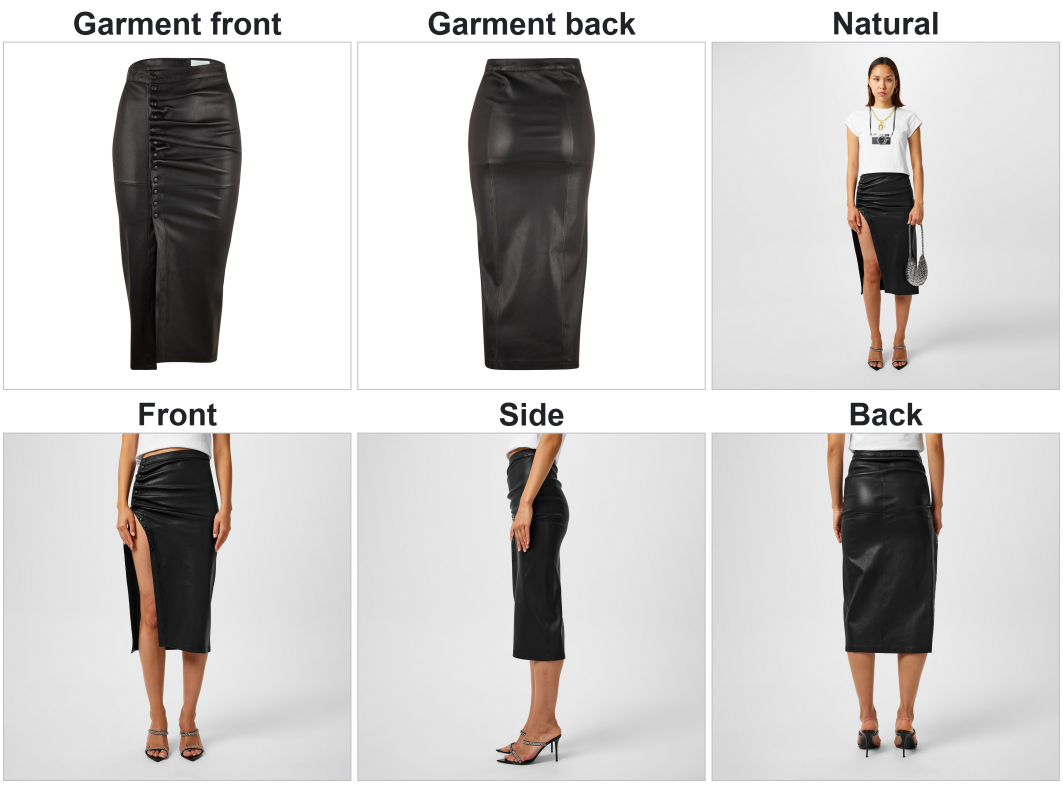

Figure 6: A DVF-6K group: paired front/back garment photographs and natural, front, side, and back on-body views with garment-focused directional framing.

and back images are assigned according to their dominant garment-facing orientation and garment-focused framing. These photographs are not required to be synchronized, and no camera calibration or pixel-level correspondence is assumed. Each six-image group thus provides structured supervision for side-specific garment routing, role-specific framing, and cross-view appearance consistency, rather than metric 3D reconstruction.

C.2 Collection, Matching, and Quality Control

We initially collect 155,412 images from publicly accessible online retail catalog pages and organize them by product identifier. After source consolidation and file validation, 69,105 images in 11,662 product folders remain as candidates. Within each folder, a multimodal matcher evaluates garment–model correspondence using color, silhouette, pattern, and distinctive local details. OpenPose keypoints Cao et al. (2017) additionally provide body-visibility and dominant-orientation cues.

The automatic predictions are used only for candidate filtering and preliminary semantic-role assignment. All retained groups are subsequently reviewed manually for product correspondence, front/back assignment, semantic-view order, garment visibility, framing, image quality, and six-image completeness. A candidate group is discarded if the retail and on-body images depict different garments, either garment side is missing, a required view is ambiguous, the garment is severely occluded, or the image lacks usable garment-focused body framing. This process leaves 6,011 complete product groups.

Product membership is preserved throughout collection and filtering. Consequently, the six images belonging to one product remain a single structured example and are never treated as independent garment–person pairs during data splitting.

Table 6: DVF-6K collection and identity-disjoint experimental split. Every group contains two garment images and four on-body views.

Partition	Groups	Identities	Images
Full DVF-6K	6,011	–	36,066
Female-presenting	3,068	–	18,408
Male-presenting	2,943	–	17,658
Strict head-valid subset	3,731	85	22,386
Train	3,001	56	18,006
Validation	430	9	2,580
Test	300	20	1,800

C.3 Identity-Disjoint Experimental Split

Identity-conditioned training requires a reliable head reference. A conservative head-validity filter retains 3,731 groups whose natural-presentation images contain a usable identity crop. We extract L2-normalized ArcFace features and construct connected components using a cosine-similarity threshold of 0.55. Complete components, rather than individual product groups, are assigned to the data partitions.

The resulting split contains 3,001 training groups from 56 inferred identities, 430 validation groups from 9 identities, and 300 test groups from 20 identities. No inferred identity appears in more than one partition. This protocol greatly reduces the risk of identity leakage and supports evaluation on identities not matched to the training split.

C.4 Dataset Composition and Category Coverage

The complete dataset contains 3,068 female-presenting and 2,943 male-presenting catalog groups across 33 recorded garment-category labels. Table 7 reports the full category distribution. T-shirt is the largest category with 1,065 groups, followed by jacket (732), pants (599), dress (568), sweater (486), and hoodie (402).

The distribution is naturally long-tailed, with twelve labels containing fewer than ten groups. For transparency and reproducibility, we retain the stored normalized category labels and do not merge rare or semantically overlapping categories after model evaluation.

A garment category is also recorded for every test group. The additional qualitative results cover seven representative labels—T-shirt, jacket, pants, shorts, dress, skirt, and sports bra—spanning upper-body, lower-body, and full-body garments. These examples complement the aggregate evaluation by examining the method across garments with different silhouettes, coverage ranges, and front/back structures.

C.5 Ethical considerations and data governance.

The source images were collected from publicly accessible online retail catalog pages and may depict identifiable individuals, trademarks, and copyrighted garment designs. We do not claim ownership of the underlying source images; the associated copyrights, trademark rights, and privacy or publicity rights remain with their respective rights holders. Public accessibility is not treated as consent from depicted individuals or as authorization for unrestricted use or redistribution.

ArcFace features and inferred identity clusters are used solely to construct identity-disjoint experimental splits. They are not linked to real-world names and are not used to determine the real-world identities of depicted individuals. The current anonymous submission package does not include the full source-image collection, face embeddings, or identity-cluster assignments; only a limited number of illustrative examples are included for academic presentation. Any access to or distribution of dataset materials is subject to documented terms of use, applicable source-specific restrictions, and a procedure through which rights holders or depicted individuals may request content removal.

Table 7: Product-group counts for all 33 normalized garment-category strings in DVF-6K. Counts sum to the 6,011 complete groups. Rare, semantically overlapping labels are retained as recorded rather than silently merged for this audit.

Category	Groups	Category	Groups
T-shirt	1,065	Jacket	732
Pants	599	Dress	568
Sweater	486	Hoodie	402
Shirt	392	Shorts	316
Polo shirt	294	Skirt	196
Tank top	176	Jeans	144
Coat	122	Vest	121
Swimsuit	98	Cardigan	81
Blouse	59	Jumpsuit	45
Long-sleeve T-shirt	32	Sports bra	23
Top	18	Crop top	8
Robe	8	Turtleneck	7
Corset	5	Leggings	4
Pajama set	3	Bra	2
Cape	1	Long sleeve	1
Long-sleeve top	1	Poncho	1
Trench coat	1		
		Total	6,011

Table 8: Overall user-study distributions. Each overall score pools the three criteria within its module. Confidence intervals resample both participants and evaluated cases; the last column reports the proportion of ratings equal to 4 or 5. Per-criterion means are reported in the main paper.

Module	Method	Overall mean [95% CI]	Median [Q1, Q3]	Ratings 4–5
Single image	Magic Clothing	2.394 [2.064, 2.764]	2 [1, 3]	21.2%
	DreamFit	2.374 [2.055, 2.715]	2 [1, 3]	18.3%
	IMAGDressing-v1	2.871 [2.538, 3.216]	3 [2, 4]	36.7%
	Qwen-Image-Edit	2.746 [2.423, 3.081]	3 [2, 4]	34.9%
	DualFashion	**4.529 [4.315, 4.715]**	**5 [4, 5]**	**90.7%**
Multi-view	Qwen-Image-Edit	3.130 [2.714, 3.551]	3 [2, 4]	44.8%
	DualFashion	**4.499 [4.244, 4.713]**	**5 [4, 5]**	**89.3%**

D Evaluation Details

D.1 Test Protocol and Baselines

All methods are evaluated on the same test groups using identical head references. We use the released checkpoints of Magic Clothing Chen et al. (2024), DreamFit Lin et al. (2025), and IMAGDressing-v1 Shen et al. (2025) with their recommended inference settings. For the single-image comparison, these methods receive the frontal garment photograph and are evaluated against the corresponding natural-presentation target. Qwen-Image-Edit-2511 Wu et al. (2025); Qwen Team (2025) receives the head reference together with both the front and back garment photographs. DualFashion uses the same three visual conditions. No method receives a ground-truth on-body image, target pose, or target scene as input.

Identity conditioning is retained for all three garment-driven baselines. The comparison uses the official SD1.5 IP-Adapter FaceID PlusV2 checkpoint Ye et al. (2023), rather than representing identity through text or omitting it. Magic Clothing uses its released garment-plus-FaceID inference pathway, including the accompanying FaceID LoRA. IMAGDressing-v1 uses its official identity-

Table 9: Inferential results for the user study. Q denotes the Friedman statistic, V the Wilcoxon signed-rank statistic, W Kendall's concordance, and r_{rb} rank-biserial correlation. Positive Δ favors DualFashion. Pairwise p-values are Holm corrected within each criterion.

Module	Criterion	Omnibus statistic	p	Effect	Δ [95% CI]	Pairwise p_{H}
Single	Garment fidelity	$Q(4) = 110.12$	6.9×10^{-23}	$W = .640$	1.381 [1.056, 1.712]	1.2×10^{-7}
	Identity preservation	$Q(4) = 74.12$	3.1×10^{-15}	$W = .431$	2.014 [1.619, 2.395]	1.5×10^{-7}
	Image realism	$Q(4) = 80.19$	1.6×10^{-16}	$W = .466$	1.577 [1.237, 1.912]	1.7×10^{-7}
Multi	Viewpoint correctness	$V = 21.5$	4.1×10^{-7}	$r_{\mathrm{rb}} = .942$	1.363 [1.000, 1.753]	4.1×10^{-7}
	Cross-view consistency	$V = 72.5$	2.1×10^{-6}	$r_{\mathrm{rb}} = .839$	1.349 [0.949, 1.781]	2.1×10^{-6}
	Overall multi-view quality	$V = 56.5$	1.2×10^{-6}	$r_{\mathrm{rb}} = .869$	1.395 [1.005, 1.791]	1.2×10^{-6}

conditioning branch without pose conditioning. DreamFit natively contains IP-Adapter attention layers, through which we load the same official PlusV2 parameters. FaceID conditioning strengths remain fixed throughout the evaluation rather than being tuned for individual examples. Thus, all methods receive the same head crop, while identity encoding follows the interface supported by each model.

D.2 View-Specific Prompt Adaptation

All prompt-driven baselines are instructed to generate a single uncropped, full-body studio photograph against a plain white background. We use four role-specific prompt templates. The natural-presentation template requests a natural three-quarter fashion pose while keeping the garment front visible; the remaining templates specify a strict straight-on front view, a strict 90° side profile, or a strict straight-on back view.

For Qwen-Image-Edit, the prompt explicitly identifies which input image shows the garment front and which shows the garment back. The natural and front prompts emphasize frontal evidence, the back prompt explicitly requires the rear garment appearance, and the side prompt refers to both garment photographs. The prompt additionally excludes contact sheets, split-screen layouts, textual labels, and cropped heads or feet.

The rear-view insets in the main paper provide an additional diagnostic for methods designed around a single garment image. In this setting, Magic Clothing, DreamFit, and IMAGDressing-v1 receive the rear garment photograph instead of the frontal photograph, together with the same head reference. The prompt requires the generated person to face fully away from the camera, showing only the back of the head and garment. Prompt wording is fixed across examples; only the recorded gender and garment category are substituted.

D.3 Metric Preprocessing

All generated outputs are converted to a common $3{:}4$ frame before evaluation. Images narrower than this aspect ratio are padded symmetrically with white pixels on the left and right, whereas wider images are symmetrically cropped along the horizontal axis. The resulting images are resized to 768×1024 without vertical cropping or anisotropic stretching.

A SCHP ResNet-101 model trained on LIP Li et al. (2020); Gong et al. (2017) is used to extract category-specific garment masks. Each detected garment bounding box is expanded by 8%, resized while preserving its aspect ratio, and centered on a white 512×512 canvas. Garment-level comparison across the five methods is conducted on the common intersection of 268 test groups for which

valid garment masks are available for every method. FID, KID, and identity evaluation use all 300 test groups.

Identity similarity is measured using ArcFace features Deng et al. (2019) extracted with the InsightFace toolkit. Garment CLIP-I Radford et al. (2021) and DINOv2-I Oquab et al. (2024) are computed on the normalized garment crops. FID and KID compare all 300 generated natural-presentation images against their ground-truth counterparts. KID is estimated using 1,000 random subsets, each containing 100 samples.

D.4 User Study

Design and response audit. The study evaluates selected qualitative cases rather than a uniformly random sample of the complete test set. The single-image pool contains 20 natural-presentation cases and the multi-view pool contains 10 complete four-view sets. Each participant receives five cases from each pool. Method identities are hidden, and method order is randomized independently within each case. The single-image module rates garment fidelity, identity preservation, and realism for five methods; the multi-view module rates viewpoint correctness, cross-view consistency, and overall presentation quality for Qwen-Image-Edit and DualFashion. All criteria use a five-point Likert scale.

We received 43 structurally complete responses. Each contains 75 single-image ratings and 30 multi-view ratings, giving 4,515 valid scores in total. The audit found no duplicated responses, missing ratings, or scores outside the permitted range. Owing to random assignment, each single-image case was rated by 7–14 participants and each multi-view case by 19–24 participants. Reported mean-score intervals use 10,000 iterations of a participant-by-sample two-way cluster bootstrap.

Statistical analysis. Table 9 reports the tests omitted from the main paper. For the five-method single-image module, participant-level method means are compared with Friedman tests, with Kendall's W as the omnibus effect size. For the two-method multi-view module, we use paired Wilcoxon signed-rank tests and rank-biserial correlation. The final two columns give the paired mean advantage of DualFashion over the strongest corresponding baseline—IMAGDressing-v1 for single images and Qwen-Image-Edit for multi-view sets—and Holm-corrected pairwise p-values. Difference intervals are obtained by paired participant bootstrap.

Aggregating the three criteria, DualFashion scores 4.529 for single images, exceeding the strongest baseline by 1.657 points (95% CI [1.330, 1.984], $p_{\mathrm{H}} = 6.0{\times}10^{-8}$, $r_{\mathrm{rb}} = .978$). Its multi-view aggregate is 4.499, a 1.369-point gain over Qwen-Image-Edit (95% CI [0.989, 1.749], $p = 1.0{\times}10^{-6}$, $r_{\mathrm{rb}} = .887$). Across all method–criterion combinations, the proportion of 4–5 ratings for DualFashion ranges from 87.9% to 92.1%.

E Additional Category Results

Figure 7 presents representative results across seven garment categories: T-shirt, jacket, pants, shorts, dress, skirt, and sports bra. These examples span upper-body, lower-body, and full-body garments with diverse silhouettes, coverage ranges, and front/back constructions. Within each row, the first three images show the front garment, back garment, and head conditions, followed by the generated natural-presentation, front, side, and back views in their fixed semantic order.

Across these diverse categories, DualFashion consistently produces coherent and garment-faithful view sets. In particular, the directional front, side, and back outputs preserve garment evidence appropriate to their semantic roles: frontal structures remain visible in the front view, rear graphics and constructions are recovered in the back view, and the side view reconciles complementary cues from both garment photographs. This side-specific routing benefits from the product-level supervision of DVF-6K, which explicitly associates paired front/back garment evidence with ordered semantic on-body views. Notably, visually plausible and structurally consistent results are also obtained for long-tail categories with relatively few training groups in Table 7, suggesting that the learned representation generalizes beyond the dominant garment categories.



Figure 7: Additional results organized by garment category. Each row shows the two garment inputs, the head reference, and the four generated semantic views.