DUAL-GEOMETRY CARRIER ATLAS TRANSPORT FOR EFFICIENT AND CONSISTENT SINGLE-REFERENCE 3D HUMAN GARMENT EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing 3D human garment editing methods typically follow a paradigm that repeatedly performs diffusion inference on multi-view images as references and updates 3D Gaussians accordingly. Repeatedly and independently applying diffusion models leads to substantial computational cost and cross-view inconsistency. In this paper, we propose Dual-geometry Carrier Atlas Transport (DCAT), which edits a single reference view and propagates the appearance to all views for 3DGS supervision. DCAT anchors the edited garment to a garment-adaptive carrier and transports its appearance through a shared UV atlas with side-view feedback, thereby coupling geometry adaptation with cross-view propagation. Specifically, Garment-Adaptive Dual-geometry Carrier deforms an SMPL-X shell according to the edited silhouette for normal garments and constructs an independent skirt proxy for skirt-like garments, thus reducing silhouette mismatch and texture distortion. Built on the carrier, UV Atlas Transport with Side-View Feedback bakes the reference appearance into a shared atlas and propagates it to the extreme side views. Boundary-locked diffusion then complements the side-view regions unobserved in the reference view, after which the results are harmonized and fed back into the atlas for propagation to the remaining views. Extensive experiments on diverse garments and editing instructions demonstrate that DCAT significantly improves editing efficiency and cross-view consistency while maintaining high visual quality.

1 INTRODUCTION

3D human garment editing has been widely applied in fields like digital avatar creation Qiu et al. (2025); Zhuang et al. (2025), film production Taubner et al. (2025); Chen et al. (2023), and personalized content generation Liu et al. (2025a); Cha et al. (2024). Text-driven 3D editing is an intuitive way to modify reconstructed scenes and has shown potential for digital humans Xu et al. (2025a); Zhang et al. (2023), virtual content creation Chen et al. (2026a); Li et al. (2025), and personalized avatars Xi et al. (2026); Liu et al. (2025a). Recent methods render 3D scene from multiple viewpoints, independently edit each view using a pretrained diffusion model, and then optimize the 3D representation using the edited views Khalid et al. (2024); Wang et al. (2024a). Although effective for general scene editing, this paradigm is inefficient and unreliable for human garment editing. Garments exhibit substan-

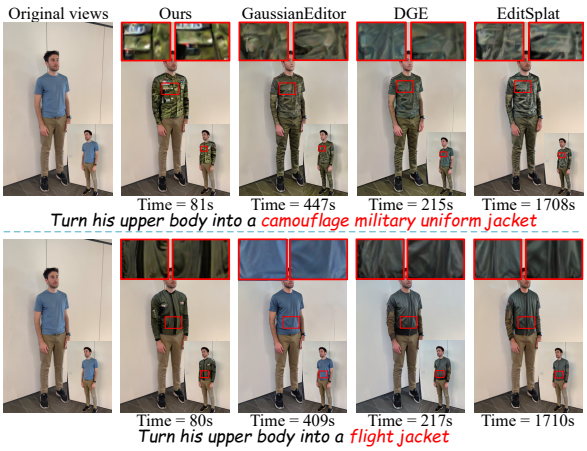


Figure 1: Comparison of text-driven 3D human garment editing. Red boxes highlight representative differences.

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.

tial diversity in topology, thickness, and silhouette, together with complex non-rigid deformations across viewpoints Li et al. (2024). Independently edited views usually exhibit inconsistent garment colors, textures, silhouettes, and local details, even when conditioned on the same text instruction. Such conflicting supervision for Gaussian primitives can lead to blurred textures, unstable geometry, and view-dependent artifacts. Moreover, given the high computational cost of diffusion inference, performing it repeatedly over a large number of views leads to considerable cost, as shown in Figure 1. Therefore, improving editing efficiency as well as cross-view consistency while accommodating diverse garment geometry remains a fundamental challenge for text-driven 3D human garment editing.

A natural alternative is to edit one reference view and propagate its result to the remaining viewpoints. However, this single-reference paradigm is non-trivial. On the one hand, the edited garment information in the reference view must be transferred to the remaining views under viewpoint changes and self-occlusion. This requires establishing reliable cross-view correspondences while faithfully preserving the edited garment's silhouette, color, and texture. On the other hand, a single reference view observes only a limited portion of the garment surface. Extreme side views may reveal garment regions that are invisible in the reference view and lack corresponding edited appearance information. To address these two challenges, we propose to construct a carrier space that maps both the multi-view images and the reference image onto a unified domain, and complete the garment regions that remain invisible from the reference viewpoint.

In this paper, we propose Dual-geometry Carrier Atlas Transport (DCAT), a single-reference framework for text-driven 3D human garment editing. DCAT first performs semantic editing on a single reference view and then transports the edited appearance through a garment-adaptive dual-geometry carrier. For general garments, the fitted SMPL-X Pavlakos et al. (2019) surface is deformed according to the edited silhouette to construct a garment shell that better matches loose clothing. For skirts, an independent closed proxy is built with regular UV parameterization to alleviate texture distortion. DCAT further introduces UV Atlas Transport with Side-View Feedback. With the adapted carriers serving as transport surfaces, reference-view textures are baked into UV space and propagated to the extreme side views. Boundary-locked diffusion then complements the missing regions while preserving the reliable propagated content. After color harmonization, the completed side appearances are baked back into the UV atlas and propagated to all remaining views. Finally, the multi-view images jointly supervise the update of the original 3DGS representation. In this way, our framework requires only one full reference-view editing rather than independent per-view editing.

The contributions of this paper are summarized as follows:

- We propose a single-reference framework for efficient and consistent text-driven 3D human garment editing. It performs the semantic edit only once and transports the result to all views, thereby reducing repeated diffusion inference and avoiding the appearance inconsistencies caused by independent per-view editing.
- We introduce a garment-adaptive dual geometry carrier, consisting of a silhouette-aligned SMPL-X garment shell and an independent skirt proxy. The two complementary carriers accommodate different garment structures, alleviating silhouette mismatch and UV distortion during appearance transport.
- We develop UV Atlas Transport with Side-View Feedback, which propagates the reference appearance to the extreme side views, completes the missing regions, and feeds the color-harmonized results back into the atlas for consistent propagation to the remaining views.

2 Related Work

Text-Driven 3D Editing. The development of pretrained diffusion models has advanced text-driven image editing, motivating its extension from 2D images to neural 3D representations Sheynin et al. (2024); Mou et al. (2024); Xu et al. (2025c). Instruct-NeRF2NeRF Haque et al. (2023) iteratively edits rendered views with a diffusion model and uses the resulting images to update the underlying NeRF Mildenhall et al. (2021). With the emergence of 3D Gaussian Splatting Kerbl et al. (2023), its explicit representation and efficient rasterization have further encouraged text-driven editing directly on Gaussian scenes Wang et al. (2024b). GaussianEditor Chen et al. (2024c) identifies text-relevant Gaussian primitives and regulates their optimization through hierarchical splatting and anchor con-

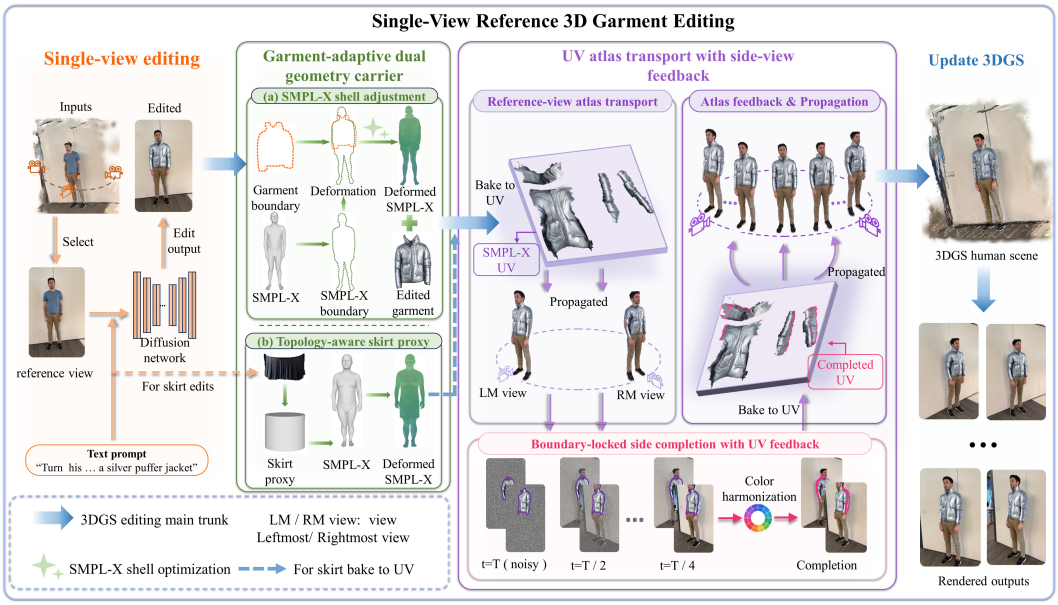


Figure 2: Framework of DCAT, which consists of four stages: (1) single-view editing, (2) garment-adaptive dual-geometry carrier, (3) UV atlas transport with side-view feedback, and (4) 3DGS updating.

straints. GaussCtrl Wu et al. (2024) jointly edits rendered views using depth-conditioned diffusion and cross-view attention alignment. DGE Chen et al. (2024b) separates the process into geometry-aware multi-view image editing and direct Gaussian fitting, avoiding incremental 3D updates. EditSplat Lee et al. (2025) combines multi-view fusion guidance with attention-guided Gaussian trimming to improve consistency and optimization efficiency. Despite recent advances in these methods, complex garment structures and textures remain difficult to preserve consistently across views.

3D Human Gaussian and Garment Editing. Recent human Gaussian methods employ parametric body priors such as SMPL or SMPL-X to provide pose correspondence and deformation guidance Liu et al. (2025b); Xu et al. (2025b); Xiao et al. (2025). GauHuman Hu et al. (2024) initializes articulated Gaussians around an SMPL body and learns pose-dependent deformation from monocular videos. 3DGS-Avatar Qian et al. (2024) represents a clothed human with canonical Gaussians and uses SMPL-guided non-rigid deformation for novel-pose animation. Despite progress in reconstruction and animation, dedicated text-driven garment editing for reconstructed human Gaussian scenes remains largely unexplored. The closest related direction is image-exemplar-driven 3D virtual try-on. GaussianVTON Chen et al. (2024a) uses a garment image as the editing prompt and refines the 3DGS result through a multi-stage try-on process. GS-VTON Cao et al. (2026) personalizes a pretrained 2D virtual try-on model with the input garment image and transfers the garment to a human 3DGS while improving cross-view consistency. These methods perform 3D garment editing using a reference garment image. In contrast, DCAT enables open-ended garment editing from natural-language instructions without an external garment image, while ensuring efficient and consistent cross-view propagation.

Overall, text-driven 3D editing methods are primarily designed for general scenes and often rely on repeated multi-view diffusion, causing high computational cost and inconsistent garment appearance across views. To address these limitations, DCAT performs editing on a single reference view and propagates the result through a garment-adaptive dual-geometry carrier and a shared UV atlas, enabling efficient and consistent 3D human garment editing.

3 METHOD

DCAT takes a reconstructed human 3DGS, a fitted SMPL-X model, and a text instruction as inputs, and edits only one reference view. As illustrated in Figure 2, Garment-Adaptive Dual Geometry

Carrier first constructs either a silhouette-aligned SMPL-X garment shell or a topology-aware skirt proxy according to the edited silhouette. The edited garment appearance in the reference view is then baked into the carrier UV atlas and propagated to two extreme side views. By Boundary-Locked Side-View Completion with Atlas Feedback, the side regions unobserved in the reference view are complemented while the reliable propagated appearance is preserved. The resulting side-view appearances are subsequently harmonized and fed back into the shared atlas. The enriched atlas is finally rendered to all remaining views, which jointly supervise the update of the original 3DGS.

3.1 Garment-Adaptive Dual Geometry Carrier

A unified carrier provides a common surface for transferring the edited garment appearance from the reference view to the remaining views. The fitted SMPL-X body model cannot accurately capture the silhouettes of loose garments, while its separated leg topology distorts the continuous silhouette and texture of skirts. To address these limitations, DCAT introduces a garment-adaptive dual-geometry carrier comprising a silhouette-aligned SMPL-X garment shell and a topology-aware skirt proxy. These complementary carriers preserve geometric alignment and stable texture parameterization prior to appearance transport.

Silhouette-Aligned SMPL-X Garment Shell. A frontal image is selected as the reference view, where the garment is edited once. The garment masks after and before editing are merged to form the target silhouette M_g, while the fitted SMPL-X is rendered using the same camera to obtain the body silhouette M_b. The two silhouettes are compared row by row to determine the pixel expansion of M_b. To convert these pixel offsets into 3D distances, the SMPL-X boundary pixels x_y^L and x_y^R are mapped to the mesh through rasterized face indices and barycentric interpolation, thereby yielding the corresponding 3D points $\mathbf{P}_y^L$ and $\mathbf{P}_y^R$. Accordingly, the 3D spatial distance κ_y per pixel at row y is estimated as

$$\kappa_y = \frac{\left\|\mathbf{P}_y^R - \mathbf{P}_y^L\right\|_2}{x_y^R - x_y^L}. \tag{1}$$

Then, each expansion pixel is assigned to the nearest body part, and the pixels on row y are projected back onto the mesh to determine the active vertices. These vertices are displaced along their outward surface normals, while the vertices of the head, hands, and feet remain fixed. Finally, the initialized position of an active vertex $\mathbf{v}_i$ located at row y is given by

$$\widehat{\mathbf{v}}_i = \mathbf{v}_i + \kappa_{y_i} \Delta x_i \mathbf{n}_i, \tag{2}$$

where y_i denotes the image row associated with vertex i, κ_{y_i} is the corresponding 3D spatial distance per image-space pixel, Δx_i denotes the number of expansion pixels assigned to vertex i, and $\mathbf{n}_i$ is its outward surface normal. For both sleeves, the displacements at the outer and inner contours are estimated separately and smoothly interpolated across the sleeve, thereby preventing the deformed arms from intersecting the torso.

Since the metric initialization can be affected by segmentation and projection errors, the garment shell is refined through differentiable silhouette optimization. First, the silhouette loss $\mathcal{L}_{\text{sil}}$ aligns the rendered soft silhouette $\widetilde{M}_b$ with the target silhouette M_g. The gradient magnitude of the rendered soft silhouette is defined as $C_b(\mathbf{x}) = \|\nabla \widetilde{M}_b(\mathbf{x})\|_2$, and the silhouette loss is formulated as

$$\mathcal{L}_{\text{sil}} = \frac{\sum_{\mathbf{x}\in\Omega} C_b(\mathbf{x}) D_g(\mathbf{x})}{\sum_{\mathbf{x}\in\Omega} C_b(\mathbf{x})}, \tag{3}$$

where $D_g(\mathbf{x})$ denotes the Euclidean distance from pixel $\mathbf{x}$ to the target silhouette boundary, and Ω denotes the image domain. Then, the row loss $\mathcal{L}_{\text{row}}$ minimizes the pixel differences between the expanded and target silhouette boundaries at each valid row. Finally, the Laplacian regularizer Wang et al. (2018) $\mathcal{L}_{\text{lap}}$ suppresses abrupt changes between neighboring vertices. The final objective is

$$\mathcal{L}_{\text{shell}} = \lambda_{\text{sil}}\mathcal{L}_{\text{sil}} + \lambda_{\text{row}}\mathcal{L}_{\text{row}} + \lambda_{\text{lap}}\mathcal{L}_{\text{lap}}. \tag{4}$$

Topology-Aware Skirt Proxy. To address the texture distortion introduced by the separated leg topology of SMPL-X in skirt regions, DCAT constructs a skirt proxy. The proxy is modeled as a

skirt-like frustum whose upper and lower cross-sections are elliptical. First, a vector from the hip center to the spine center defines the longitudinal body axis and the orientation of its cross-sectional planes. Next, the original lower garment mask is projected back onto the SMPL-X surface, where the highest valid 3D point along this axis determines the position of the upper elliptical plane. SMPL-X vertices near the upper ellipse are projected onto this plane to estimate the semi-axis lengths of the upper ellipse and the lateral and front-to-back unit axes $\mathbf{e}_{\mathrm{lr}}$ and $\mathbf{e}_{\mathrm{fb}}$. Subsequently, the lower boundary of the edited skirt is projected back to determine the bottom plane. Its center is adjusted along the body longitudinal axis to match the projected skirt hem. This lower ellipse is scaled from the upper ellipse using the pixel width ratio between the edited skirt hem and the original upper boundary. Finally, the geometry between the two ellipses is smoothly interpolated into K rings with centers $\mathbf{c}_k$ and semi-axis lengths a_k and b_k. By uniformly sampling each ring, its j-th vertex $\mathbf{q}_{k,j}$ is defined as

$$\begin{aligned} \mathbf{q}_{k,j} &= \mathbf{c}_k + a_k \cos\theta_j \, \mathbf{e}_{\mathrm{lr}} + b_k \sin\theta_j \, \mathbf{e}_{\mathrm{fb}}, \\ \theta_j &= \frac{2\pi j}{J}. \end{aligned} \tag{5}$$

The ring structure naturally induces a regular UV parameterization. For K longitudinal rings and J angular intervals, the UV coordinate of each proxy vertex is defined as

$$\mathbf{u}_{k,j} = \left(\frac{j}{J}, 1 - \frac{k}{K-1} \right). \tag{6}$$

The circumferential and longitudinal directions are thereby mapped to the horizontal and vertical axes of a rectangular UV atlas, respectively. Duplicating the seam vertices at the circumferential start $u = 0$ and end $u = 1$ preserves texture continuity around the skirt. The skirt proxy design and regular parameterization avoid the texture separation and stretching caused by the original leg UV layout.

3.2 UV Atlas Transport with Side-View Feedback

The garment-adaptive carrier provides a garment-aligned surface and stable UV parameterization, but a single reference view leaves the atlas incomplete in extreme side views. To this end, UV Atlas Transport with Side-View Feedback combines Reference-View UV Atlas Transport and Boundary-Locked Side-View Completion with Atlas Feedback to anchor the edited appearance in a shared atlas and recover unobserved regions. The harmonized side-view appearances are then fed back into the same texture domain, enabling consistent propagation to the remaining views and efficient, coherent supervision for 3DGS updating.

Reference-View UV Atlas Transport. The adapted carrier is rasterized to establish correspondences between visible mesh faces and reference-view pixels. Each mesh face has a corresponding triangle in the carrier UV parameterization, defined by the UV coordinates of its three vertices. Therefore, the barycentric coordinates of a garment pixel $\mathbf{x}$ on its associated face $f(\mathbf{x})$ can be transferred to the corresponding UV triangle, establishing the mapping from the reference image to the UV atlas. Its UV coordinate is computed as

$$\mathbf{u}(\mathbf{x}) = \sum_{m=1}^{3} \beta_m(\mathbf{x}) \mathbf{u}_{f(\mathbf{x}),m}, \tag{7}$$

where $\beta_m(\mathbf{x})$ and $\mathbf{u}_{f(\mathbf{x}),m}$ denote the barycentric weight and UV coordinate of the m-th face vertex, respectively. The edited colors are accumulated at the corresponding texels to construct the reference UV atlas. Only correspondences between the frontmost visible mesh regions of the edited garment and the reference image pixels are retained, thereby excluding occluded surfaces and regions outside the edited garment from atlas construction.

Once the reference appearance is baked into the UV atlas, it can be propagated to other views through the same UV to pixel correspondence. However, other views may expose garment regions that are invisible in the reference view, resulting in incomplete garment surfaces after propagation. Therefore, DCAT propagates the atlas only to the extreme left and right views, where these missing regions are subsequently completed under boundary locking.

Boundary-Locked Side-View Completion with Atlas Feedback. For each extreme side view, the edited reference view, text instruction, and UV-propagated side-view image condition a diffusion

model (FLUX.2-klein-4B Black Forest Labs (2025)) to synthesize the garment regions unobserved in the reference view. The propagated side-view image is encoded into the latent space, while its propagated garment region is converted into a soft locking mask L_s. During each denoising step t, the propagated latent is re-noised to the current timestep as $\mathbf{z}^t_{s,\text{prop}}$ and blended with the latent $\mathbf{z}^t_s$:

$$\mathbf{z}^t_s \leftarrow L_s \odot \mathbf{z}^t_{s,\text{prop}} + (1 - L_s) \odot \mathbf{z}^t_s, \tag{8}$$

where $s \in \{L, R\}$ denotes the left or right extreme side view. This boundary-aware latent blending preserves the propagated garment appearance in the locked region while allowing the diffusion model to synthesize the missing regions. After decoding, the propagated region is restored, and garment segmentation extracts the newly generated regions.

The generated side appearance $\mathbf{I}^{\text{gen}}_s$ may still exhibit color and illumination discrepancies from the propagated garment. For each connected component of the newly completed region F_s, nearest boundary correspondences are established with the adjacent propagated region, and channel-wise log gains are estimated in linear RGB space. Unreliable correspondences are rejected using robust statistics based on the median and median absolute deviation (MAD), after which the sparse boundary gains are smoothly propagated throughout the completed region via screened-Laplacian interpolation Morel et al. (2014). The resulting spatially varying gain field $\mathbf{g}_s$ corrects $\mathbf{I}^{\text{gen}}_s$ at each pixel $\mathbf{x} \in F_s$, yielding the harmonized side appearance:

$$\widehat{\mathbf{I}}_s(\mathbf{x}) = \Gamma\big(\exp(\mathbf{g}_s(\mathbf{x})) \odot \Gamma^{-1}(\mathbf{I}^{\text{gen}}_s(\mathbf{x}))\big)\,, \tag{9}$$

where Γ^{-1} and Γ convert colors from sRGB to linear RGB and back, respectively. The smooth gain field removes large-scale color and illumination shifts while retaining the local texture synthesized by FLUX.2-klein-4B.

The harmonized side appearances are finally mapped back to the shared UV domain using the same image-to-atlas correspondence as in Eq. 7. Only texels associated with the newly completed regions are inserted, leaving the reliable reference-view texels unchanged. The complementary feedback from the two extreme side views enriches the previously unobserved atlas regions, after which the completed atlas is rendered directly to all remaining views. The resulting edited views supervise the update of the original 3DGS using an ℓ_1 reconstruction loss and an LPIPS perceptual loss:

$$\mathcal{L}_{\text{3DGS}} = \left\|\mathbf{I}^{\text{rend}}_v - \mathbf{I}^{e}_v\right\|_1 + \text{LPIPS}\left(\mathbf{I}^{\text{rend}}_v, \mathbf{I}^{e}_v\right), \tag{10}$$

where $\mathbf{I}^{\text{rend}}_v$ and $\mathbf{I}^{e}_v$ denote the 3DGS rendering and its edited supervision image at view v, respectively.

4 Experiments

4.1 Experimental Setup

Implementation Details. DCAT represents human scenes using 3D Gaussian Splatting (3DGS). InstructPix2Pix Brooks et al. (2023) and FLUX.1-Kontext Labs et al. (2025) are employed as the 2D diffusion-based image editors. For each scene, the number of selected views is set to 20, and the 3DGS representation is optimized for 1,200 iterations. Further details on the parameter settings are provided in the Appendix. Following the recommendations of IN2N Haque et al. (2023), both LPIPS Zhang et al. (2018) and L1 losses are utilized to update the 3DGS human scenes. All experiments are conducted on a workstation equipped with two NVIDIA GeForce RTX 4090 GPU (48 GB VRAM).

Compared Methods. DCAT is first compared with three state-of-the-art text-driven 3D Gaussian scene editing methods: GaussianEditor Chen et al. (2024c), DGE Chen et al. (2024b), and EditSplat Lee et al. (2025). These methods employ InstructPix2Pix as the 2D image editor, while DGE and EditSplat particularly emphasize cross-view consistency, making them strong baselines. DCAT is then compared with Omni-3DEdit Chen et al. (2026b), which adopts the newer FLUX.1 editor and similarly performs single-view editing followed by cross-view propagation. Finally, DCAT is compared with the state-of-the-art 3DGS virtual try-on method GS-VTON Cao et al. (2026). Although GS-VTON takes a garment image rather than a text instruction as input, it performs well in human garment editing and thus serves as a valuable baseline.

Evaluation Metrics. Semantic alignment is evaluated using CLIP similarity and CLIP directional similarity Radford et al. (2021). Editing quality and cross-view consistency are assessed using

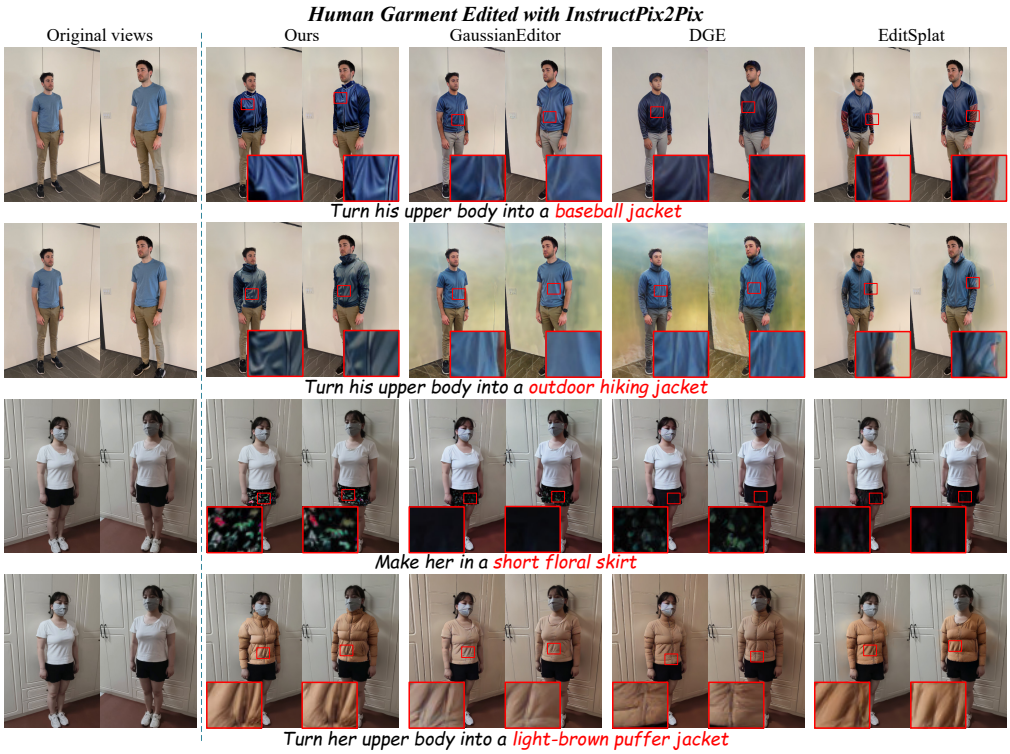


Figure 3: Qualitative comparison with InstructPix2Pix-based 3D editing methods. Red boxes show enlarged regions.

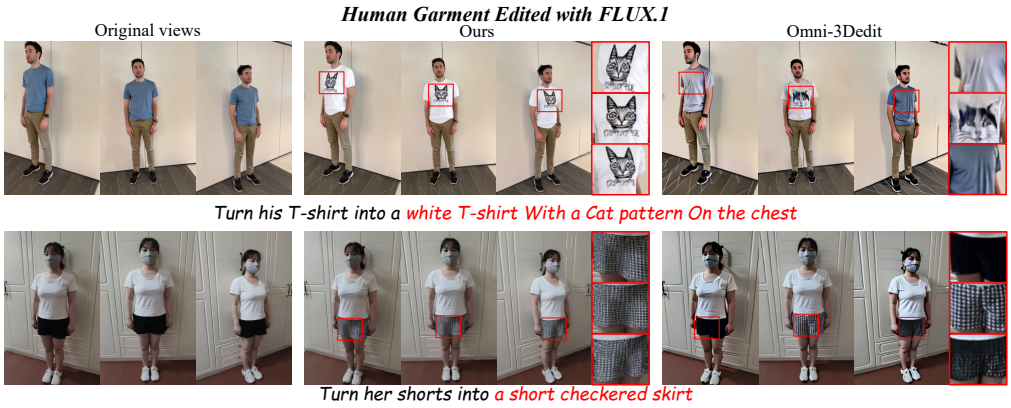


Figure 4: Qualitative comparison with the FLUX.1-based Omni-3DEdit. Red boxes show enlarged regions.

PSNR, SSIM, and LPIPS. Since no ground truth is available, we use the fully edited and propagated images as ground truth. Specifically, 80% of the images are selected as the training set for 3DGS reconstruction, and the reconstructed 3DGS is rendered on the test set to compute PSNR, SSIM, and LPIPS. In addition, the average runtime is calculated over all instructions after completing the full editing process for each instruction.

4.2 Qualitative Comparisons

Comparison with IP2P-based 3D editing methods. DCAT is first compared with InstructPix2Pix-based 3D editing methods, GaussianEditor, DGE, and EditSplat, as shown in Figure 3. The zoomed

Table 1: Quantitative comparison with state-of-the-art methods.

Method	CLIP_{sim} ↑	CLIP_{dir} ↑	PSNR↑	SSIM↑	LPIPS↓	Avg. Time (s)↓
GaussianEditor-IP2P (CVPR 2024)	0.2643	0.1149	21.3446	0.7961	0.2029	436s
DGE-IP2P (ECCV 2024)	0.2612	0.1966	25.9940	0.8578	0.1019	232s
EditSplat-IP2P (CVPR 2025)	0.2637	0.2070	21.9123	0.7808	0.1784	1575s
DCAT-IP2P (ours)	**0.2664**	**0.2220**	**31.6121**	**0.9580**	**0.0380**	**83s**
Omni-3Dedit-FLUX.1 (CVPR 2026)	0.2896	0.0549	23.2844	0.8693	0.1084	282s
DCAT-FLUX.1 (ours)	**0.3085**	**0.1980**	**32.8073**	**0.9599**	**0.0303**	**92s**

regions reveal cross-view garment inconsistencies, including blurred textures and inaccurate patterns. For example, for "Make her in a short floral skirt", all three methods fail to recover the floral patterns. Moreover, GaussianEditor and DGE introduce undesired changes in non-target regions, such as background artifacts and facial distortions, due to inconsistent multi-view supervision. Although EditSplat improves regional accuracy, it still produces distorted shoulders and blurry sleeves. In contrast, DCAT performs semantic editing only on a single reference view and propagates the appearance through a garment-adaptive carrier, thereby preserving garment details, identity information, and non-target regions across viewpoints.

Comparison with FLUX.1-based 3D editing methods. DCAT is further compared with Omni-3DEdit, which adopts the FLUX.1 and follows a single-view editing and propagation paradigm, as shown in Figure 4. Omni-3DEdit benefits from the strong generation capability of FLUX.1 and produces visually appealing garments with complex patterns in the reference view. However, other views exhibit inconsistent garment patterns and degraded local details after propagation. This is mainly because its learned propagation strategy lacks explicit cross-view texture correspondences under large viewpoint variations. In contrast, DCAT introduces a geometry-aware dual carrier and shared UV atlas to establish reliable correspondences, faithfully propagating garment structures and fine-grained textures to other views.

Comparison with 3DGS virtual try-on methods. Finally, DCAT is compared with GS-VTON, as shown in Figure 5. Since GS-VTON requires a garment image, the reference views edited by DCAT are provided as garment references for fair comparison. GS-VTON can transfer the reference garment while maintaining the original background. However, it remains sensitive to input quality and illumination, producing color and texture deviations and cross-view inconsistencies, as shown in Figure 5. In contrast, DCAT enables text-driven garment editing without requiring an external garment image and achieves more consistent multi-view appearance through geometry-aware atlas transport.

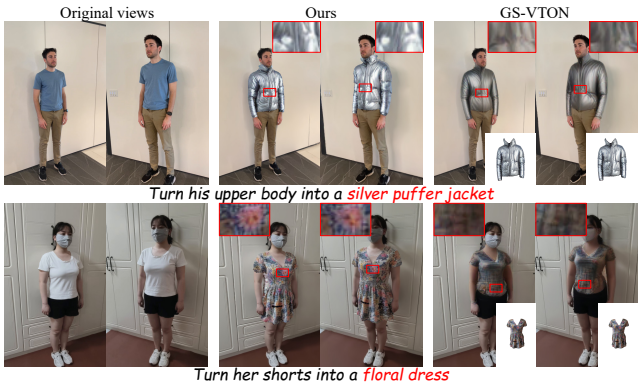


Figure 5: Qualitative comparison with GS-VTON. For a fair comparison, the garment appearance from the DCAT-edited reference view is provided to GS-VTON as the exemplar. Insets show garment references and local details.

More qualitative comparisons across methods and text instructions are provided in the Appendix.

4.3 Quantitative Comparisons

Table 1 compares DCAT with three IP2P-based methods over 15 instructions and one FLUX.1-based method over four instructions on two human scenes. Under IP2P, DCAT achieves the best performance across all reported metrics, including CLIP similarity, CLIP directional similarity, PSNR, SSIM, LPIPS, and runtime. More importantly, DCAT takes only about 83 seconds on average to

complete a full 3DGS human garment edit, making it approximately 2.8 times faster than DGE, the second most efficient method. These results quantitatively demonstrate that DCAT avoids repeated diffusion inference and the inconsistencies caused by independently editing multiple views. By editing only a single reference view and propagating the resulting UV atlas across viewpoints through the dual-geometry carrier, DCAT substantially reduces computational cost while maintaining highly consistent garment appearance across views.

4.4 Ablation Study

To evaluate the Silhouette-Aligned SMPL-X Garment Shell, Topology-Aware Skirt Proxy, and Boundary-Locked Side-View Completion with Atlas Feedback, we conduct ablation studies on each component.

First, we remove the SMPL-X shell adjustment and use the originally fitted SMPL-X as the transport carrier. As shown in the first row of Figure 6, its body-tight surface cannot match the edited garment silhouette, causing collar and sleeve boundaries to disappear and the garment to cling to the body. This confirms the necessity of the Silhouette-Aligned SMPL-X Garment Shell.

Second, without the Skirt Proxy, skirt textures are distorted by the concave topology between the SMPL-X legs, as shown in the second row of Figure 6. The proposed proxy better represents the loose skirt structure and preserves texture details, validating its effectiveness.

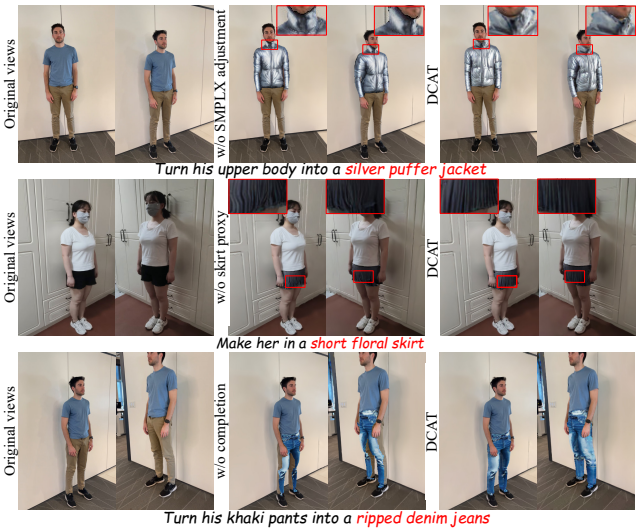


Figure 6: Ablation study of the main components of DCAT.

Finally, the third row of Figure 6 demonstrates the effectiveness of Boundary-Locked Side-View Completion with Atlas Feedback. Without it, large side-view regions retain the original garment due to reference-view occlusion. The proposed mechanism completes these unseen regions and produces a more complete garment appearance.

5 Conclusion

In this paper, we propose DCAT, an efficient and consistent single-reference framework for text-driven 3D human garment editing. DCAT employs a silhouette-aligned SMPL-X garment shell and a topology-aware skirt proxy to accommodate diverse garment structures, while UV atlas transport with side-view feedback propagates the edited appearance across views. Extensive experiments demonstrate that DCAT improves editing efficiency, cross-view consistency, and visual quality compared with existing methods. A current limitation is that appearance propagation relies on mesh triangles, which may introduce unstable or noisy pixels near garment boundaries. Future work will explore more robust boundary-aware transport and refinement strategies to further improve propagation quality.

AI use statement

Generative AI tools were used to assist with language polishing, including improving readability, shortening sentences, and refining LaTeX formatting. They were not used to design the proposed methodology or experiments, generate or modify experimental data, implement the proposed method, or interpret experimental results. All AI-assisted text and formatting were reviewed and revised by the authors. The authors take full responsibility for the final content of this work.

Ethics statement

The experiments use one publicly available human scene and one human scene recorded specifically for this study. The recorded participant provided informed consent for the recording and for the use of the resulting data for research and publication. The proposed method is intended for research on controllable 3D human garment editing and does not aim to infer sensitive personal attributes. Since human appearance editing techniques may potentially be misused to create deceptive or unauthorized content, we encourage their use only with appropriate consent and in accordance with applicable privacy and intellectual-property requirements.

Reproducibility statement

We provide detailed descriptions of the proposed components and optimization objectives in the Method section. The experimental setup, including the employed image editors, optimization protocol, evaluation metrics, and computational environment, is specified in the Experiments section. The Appendix further documents the dataset construction and preprocessing procedure, complete hyperparameter settings, sensitivity analyses, and additional experimental results. These details are provided to facilitate reproduction of the reported experiments and evaluation of the proposed method.

References

Black Forest Labs. FLUX.2: Frontier visual intelligence. https://bfl.ai/blog/flux-2, 2025. Accessed: 2026-07-27.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Yukang Cao, Masoud Hadi, Liang Pan, and Ziwei Liu. GS-VTON: Controllable 3D virtual try-on with gaussian splatting. *International Journal of Computer Vision*, 134(5):215, 2026.

Hyunsoo Cha, Byungjun Kim, and Hanbyul Joo. Pegasus: Personalized generative 3D avatars with composable attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1072–1081, 2024.

Haodong Chen, Yongle Huang, Haojian Huang, Xiangsheng Ge, and Dian Shao. GaussianVTON: 3D human virtual try-on via multi-stage gaussian splatting editing with image prompting, 2024a.

Liyi Chen, Ruihuang Li, Guowen Zhang, Pengfei Wang, and Lei Zhang. Fast multi-view consistent 3D editing with video priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2948–2956, 2026a.

Liyi Chen, Pengfei Wang, Guowen Zhang, Zhiyuan Ma, and Lei Zhang. Omni-3DEdit: Generalized versatile 3D editing in one-pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12640–12650, 2026b.

Minghao Chen, Iro Laina, and Andrea Vedaldi. DGE: Direct gaussian 3D editing by consistent multi-view editing. In *European Conference on Computer Vision*, pp. 74–92, 2024b.

Yiwen Chen, Zilong Chen, Chi Zhang, Feng Wang, Xiaofeng Yang, Yikai Wang, Zhongang Cai, Lei Yang, Huaping Liu, and Guosheng Lin. GaussianEditor: Swift and controllable 3D editing with gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21476–21485, 2024c.

Yue Chen, Xuan Wang, Xingyu Chen, Qi Zhang, Xiaoyu Li, Yu Guo, Jue Wang, and Fei Wang. Uv volumes for real-time rendering of editable free-view human performance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16621–16631, 2023.

Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-NeRF2NeRF: Editing 3D scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19740–19750, 2023.

Shoukang Hu, Tao Hu, and Ziwei Liu. GauHuman: Articulated gaussian splatting from monocular human videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20418–20431, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139–1, 2023.

Umar Khalid, Hasan Iqbal, Azib Farooq, Jing Hua, and Chen Chen. 3Dego: 3D editing on the go! In *European Conference on Computer Vision*, pp. 73–89, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space, 2025.

Dong In Lee, Hyeongcheol Park, Jiyoung Seo, Eunbyung Park, Hyunje Park, Ha Dam Baek, Sangheon Shin, Sangmin Kim, and Sangpil Kim. EditSplat: Multi-view fusion and attention-guided optimization for view-consistent 3D scene editing with 3D gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11135–11145, 2025.

Ren Li, Corentin Dumery, Benoît Guillard, and Pascal Fua. Garment recovery with shape and deformation priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1586–1595, 2024.

Ruihuang Li, Liyi Chen, Zhengqiang Zhang, Varun Jampani, Vishal M Patel, and Lei Zhang. Sync-Noise: Geometrically consistent noise prediction for instruction-based 3D editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4905–4913, 2025.

Hanxi Liu, Yifang Men, and Zhouhui Lian. Creating your editable 3D photorealistic avatar with tetrahedron-constrained gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15976–15986, 2025a.

Yifan Liu, Shengjun Zhang, Chensheng Dai, Yang Chen, Hao Liu, Chen Li, and Yueqi Duan. Learning efficient and generalizable human representation with human gaussian model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11797–11806, 2025b.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Jean-Michel Morel, Ana-Belen Petro, and Catalina Sbert. Screened poisson equation for image contrast enhancement. *Image Processing On Line*, 4:16–29, 2014.

Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Diffeditor: Boosting accuracy and flexibility on diffusion-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8488–8497, 2024.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10975–10985, 2019.

Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3DGS-Avatar: Animatable avatars via deformable 3D gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5020–5030, 2024.

Lingteng Qiu, Shenhao Zhu, Qi Zuo, Xiaodong Gu, Yuan Dong, Junfei Zhang, Chao Xu, Zhe Li, Weihao Yuan, Liefeng Bo, et al. AniGS: Animatable gaussian avatar from a single image with inconsistent gaussian reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21148–21158, 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 *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8871–8879, 2024.

Felix Taubner, Ruihang Zhang, Mathieu Tuli, and David B Lindell. Cap4D: Creating animatable 4D portrait avatars with morphable multi-view diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5318–5330, 2025.

Junjie Wang, Jiemin Fang, Xiaopeng Zhang, Lingxi Xie, and Qi Tian. Gaussianeditor: Editing 3D gaussians delicately with text instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20902–20911, 2024a.

Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3D mesh models from single rgb images. In *European Conference on Computer Vision*, pp. 52–67, 2018.

Yuxuan Wang, Xuanyu Yi, Zike Wu, Na Zhao, Long Chen, and Hanwang Zhang. View-consistent 3D editing with gaussian splatting. In *European Conference on Computer Vision*, pp. 404–420, 2024b.

Jing Wu, Jia-Wang Bian, Xinghui Li, Guangrun Wang, Ian Reid, Philip Torr, and Victor Adrian Prisacariu. GaussCtrl: Multi-view consistent text-driven 3D gaussian splatting editing. In *European Conference on Computer Vision*, pp. 55–71, 2024.

Dianbing Xi, Guoyuan An, Jingsen Zhu, Zhijian Liu, Yuan Liu, Ruiyuan Zhang, Jiayuan Lu, Yuchi Huo, and Rui Wang. PFAvatar: Pose-fusion 3D personalized avatar reconstruction from real-world outfit-of-the-day photos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10924–10932, 2026.

Junjin Xiao, Qing Zhang, Yonewei Nie, Lei Zhu, and Wei-Shi Zheng. RoGSplat: Learning robust generalizable human gaussian splatting from sparse multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5980–5990, 2025.

FeiFan Xu, Tianyi Chen, Fan Yang, Yunfei Zhang, and Si Wu. 3DHumanEdit: Multi-modal body part-aware conditioning information integration for 3D human manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8833–8841, 2025a.

Wangze Xu, Yifan Zhan, Zhihang Zhong, and Xiao Sun. Sequential gaussian avatars with hierarchical motion context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13592–13603, 2025b.

Yingjing Xu, Jie Kong, Jiazhi Wang, Xiao Pan, Bo Lin, and Qiang Liu. Insightedit: Towards better instruction following for image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2694–2703, 2025c.

Huichao Zhang, Bowen Chen, Hao Yang, Liao Qu, Xu Wang, Li Chen, Chao Long, Feida Zhu, Kang Du, and Min Zheng. AvatarVerse: High-quality & stable 3D avatar creation from text and pose, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. Idol: Instant photorealistic 3D human creation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26308–26319, 2025.

APPENDIX

A DATASET AND EXPERIMENTAL DETAILS

Our experiments are evaluated on two reconstructed human scenes, including one male subject and one female subject. The male scene is taken from the publicly available data released with Instruct-NeRF2NeRF. For the female scene, we record a monocular video and process the extracted frames using the Nerfstudio pipeline to recover camera parameters and generate the corresponding multi-view images. For both scenes, the images from different viewpoints are first center-cropped to 320×512 and then used to reconstruct the human scenes with 3D Gaussian Splatting (3DGS). Following the official 3DGS implementation, these images are automatically resized to 512×512 during both the initial scene reconstruction and the subsequent 3DGS updating stage of DCAT.

B HYPERPARAMETER SETTINGS AND SENSITIVITY ANALYSIS

B.1 HYPERPARAMETER SETTINGS

Unless otherwise specified, the hyperparameters associated with 3DGS updating are configured following the official implementations of GaussianEditor and DGE. The reconstructed 3DGS representation is optimized for 1,200 iterations using 20 selected views, consistent with the experimental settings described in the main paper. For the silhouette-aligned SMPL-X garment shell, the vertex offsets are optimized for 100 iterations using the Adam optimizer with a learning rate of 0.08. The weights of the silhouette loss, row boundary loss, and Laplacian regularization are set to $\lambda_{\text{sil}} = 1.0$, $\lambda_{\text{row}} = 1.0$, and $\lambda_{\text{lap}} = 1.0$, respectively. For the topology-aware skirt proxy, the numbers of circumferential intervals and longitudinal rings are fixed to $J = 120$ and $K = 50$, respectively. The UV atlas resolution R_{UV} is set to 2048×2048. These hyperparameters are kept unchanged across all experiments.

Table 2: Main hyperparameter settings of DCAT.

Parameter	Meaning	Value
λ_{sil}	Silhouette loss weight	1.0
λ_{row}	Row boundary loss weight	1.0
λ_{lap}	Laplacian regularization weight	1.0
N_{shell}	Number of garment shell optimization iterations	100
η_{shell}	Garment shell learning rate	0.08
J	Number of circumferential intervals	120
K	Number of longitudinal rings	50
R_{UV}	UV atlas resolution	2048×2048

B.2 SENSITIVITY ANALYSIS

We further investigate the influence of the skirt-proxy resolution, controlled by the number of longitudinal rings K and circumferential intervals J. When $K = 25$, the proxy is insufficiently sampled along the longitudinal direction. The resulting coarse geometry produces irregular skirt boundaries after appearance propagation, introducing noisy pixels into adjacent non-target regions. Increasing K to 50 substantially improves the geometric continuity of the proxy. Under this setting, both $J = 120$ and $J = 160$ yield stable propagation results with well-preserved skirt boundaries. Since their visual differences are minor, a smaller J is preferable because it reduces the number of mesh faces and the associated rasterization and UV-processing costs. When $K = 75$, setting $J = 40$ provides insufficient circumferential sampling and leads to visible boundary noise, whereas $J = 80$ produces more reliable results. Although $(K, J) = (75, 80)$ and $(50, 120)$ result in comparable mesh complexity, the latter preserves finer circumferential details and produces cleaner garment textures. Further increasing J to 120 or 160 under $K = 75$ introduces noticeable holes in the propagated skirt appearance. This occurs because the excessively fine tessellation produces very small projected triangles relative to the fixed UV baking resolution, making stable texel coverage more difficult and leaving some regions without valid appearance samples. The hole artifacts become even more pronounced when K is increased to 100. Based on this trade-off among boundary quality, texture completeness, and computational efficiency, we set $K = 50$ and $J = 120$ in all experiments, as shown in Figure 7.

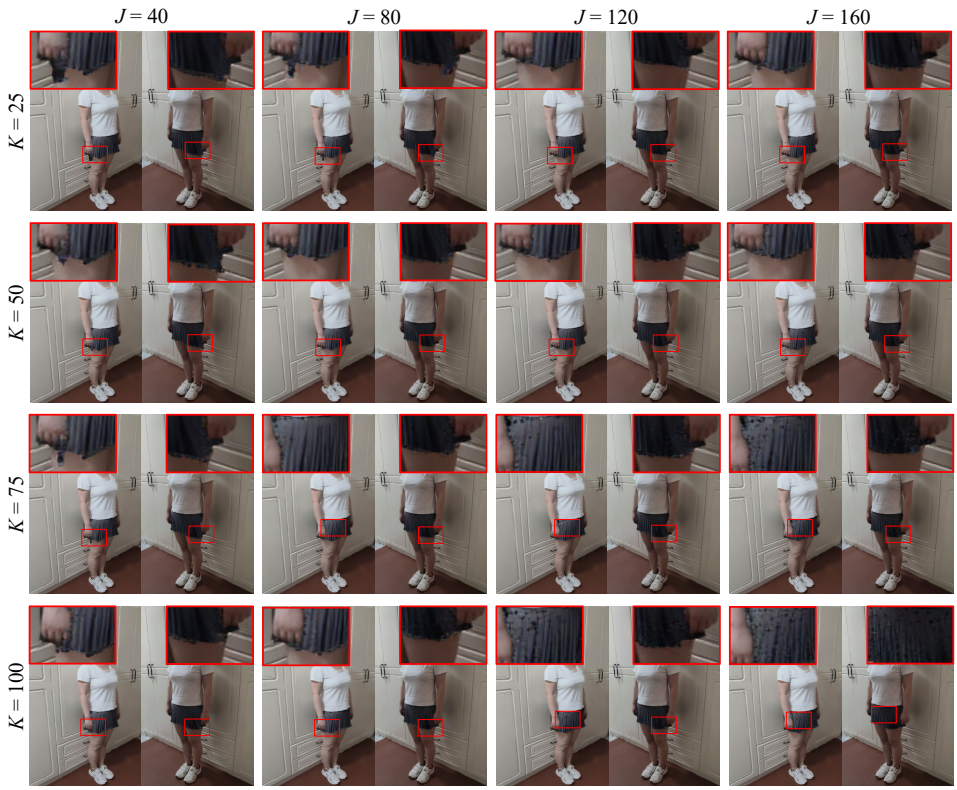


Figure 7: Qualitative sensitivity analysis of the skirt-proxy resolution with different numbers of longitudinal rings K and circumferential intervals J. Red boxes show enlarged boundary and texture regions. Coarse resolutions introduce boundary noise, while overly dense tessellation causes incomplete UV coverage and visible holes. We therefore adopt $K = 50$ and $J = 120$ as the default setting.

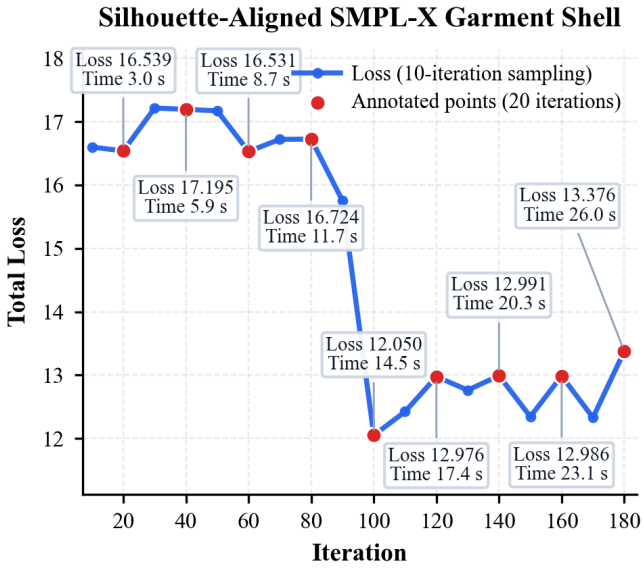


Figure 8: Convergence analysis of the silhouette-aligned SMPL-X garment shell optimization.

We analyze the influence of the optimization iteration number N_{shell} on the silhouette-aligned SMPL-X garment shell. As shown in the loss curve, the total loss exhibits an overall decreasing trend and reaches a relatively low value at approximately 100 iterations. Beyond this point, the loss

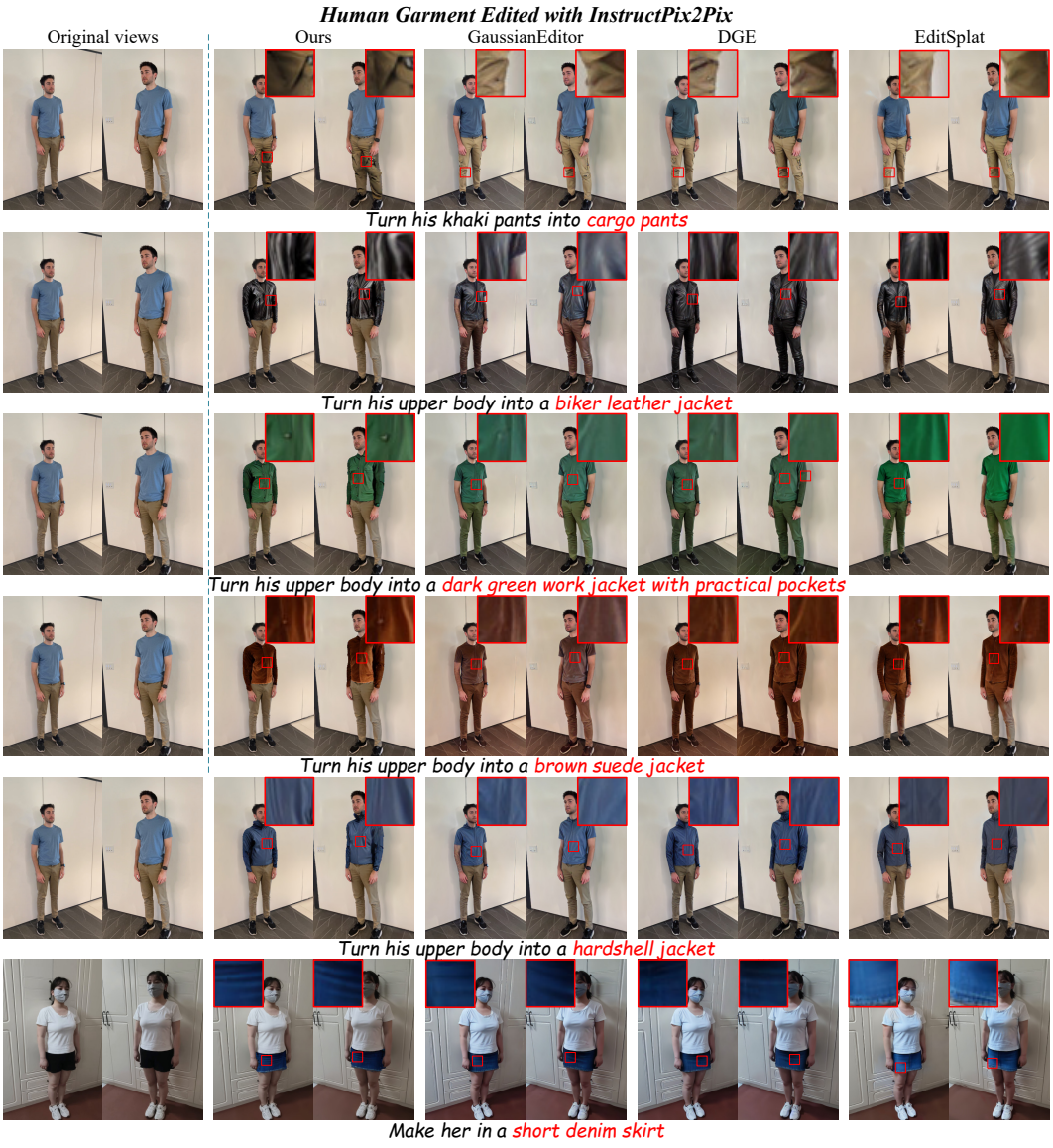


Figure 9: Additional qualitative comparisons with InstructPix2Pix-based 3D editing methods under diverse garment-editing instructions. Red boxes show enlarged garment regions, highlighting differences in cross-view consistency, texture clarity, and local artifacts.

mainly fluctuates within a stable range without showing consistent further improvement, while the computational time continues to increase. Considering the trade-off between alignment accuracy and optimization efficiency, we set $N_{\text{shell}} = 100$ in all experiments, as shown in Figure 8.

C Additional Experimental Results

C.1 Additional Comparisons with InstructPix2Pix Methods

Figure 9 presents additional qualitative comparisons with the InstructPix2Pix-based methods GaussianEditor, DGE, and EditSplat under diverse garment-editing instructions. Although these methods can produce semantically plausible edits in some viewpoints, independently editing multiple views often results in inconsistent garment appearances. In particular, local structures may vary across viewpoints, while fine-grained textures are frequently blurred or accompanied by visible ar-

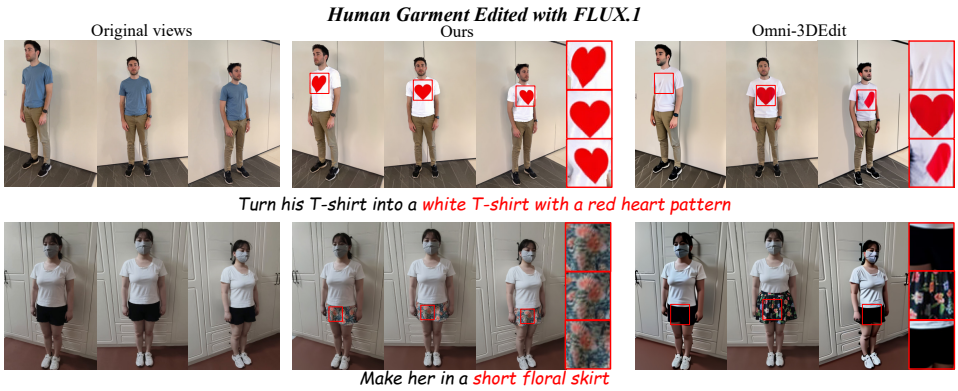


Figure 10: Additional qualitative comparisons with the FLUX.1-based Omni-3DEdit. Red boxes show enlarged garment regions. DCAT preserves garment patterns and local details more consistently under large viewpoint variations.

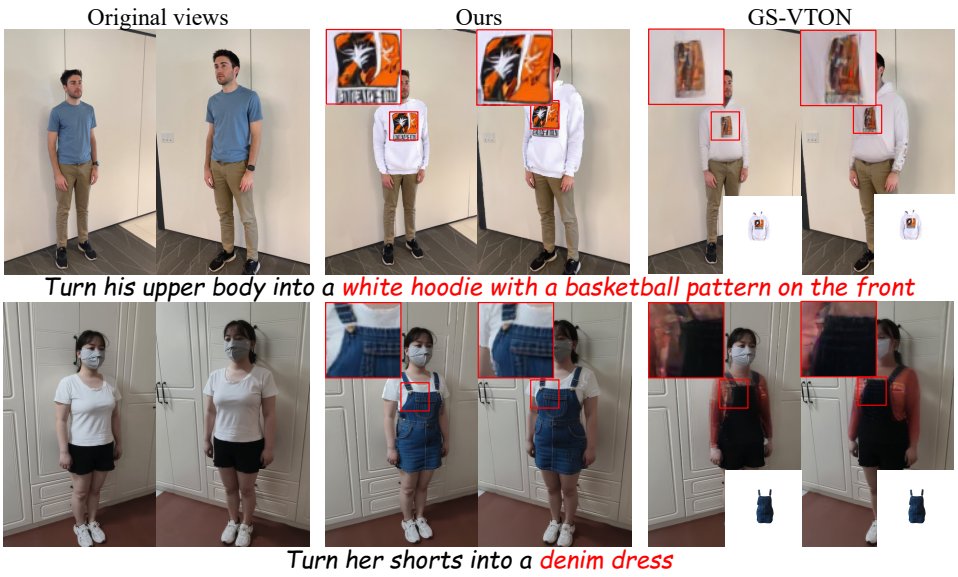


Figure 11: Additional qualitative comparisons with GS-VTON. For a fair comparison, the garment appearance extracted from the DCAT-edited reference view is provided to GS-VTON as the garment exemplar. Insets show garment references and enlarged local details.

tifacts. Some methods also introduce unintended modifications to non-target regions. In contrast, DCAT performs semantic editing only on the reference view and transports the resulting appearance through a shared UV atlas. This strategy preserves the edited garment structure and texture more consistently while reducing artifacts in adjacent body and background regions.

C.2 Additional Comparisons with FLUX.1 Method

Figure 10 provides further comparisons with Omni-3DEdit, which adopts FLUX.1 for single-view editing and subsequently propagates the edited appearance to other viewpoints. Benefiting from the strong image-generation capability of FLUX.1, Omni-3DEdit can produce visually appealing reference-view edits with complex colors and patterns. However, when the viewpoint changes substantially, its propagation process may fail to preserve the spatial arrangement and local details of the edited garment, resulting in inconsistent patterns, degraded textures, or incomplete garment appearance across views. By contrast, DCAT employs the silhouette-aligned SMPL-X garment shell and the topology-aware skirt proxy to establish explicit geometry-aware correspondences. The shared

UV atlas therefore enables the edited structures and fine-grained textures to be transported more consistently under large viewpoint variations.

C.3 Additional Comparisons with 3DGS Virtual Try-On

Figure 11 shows additional comparisons with GS-VTON. Since GS-VTON requires an external garment image rather than a text instruction, we extract the garment appearance from the DCAT-edited reference view and use it as the garment exemplar for a fair comparison. GS-VTON can preserve the overall scene background and transfer the coarse garment category in some cases. Nevertheless, its performance is sensitive to the quality, illumination, pose, and visible content of the provided garment reference. When the exemplar contains complex patterns or only partially visible garment regions, GS-VTON may fail to reproduce the intended details, leading to blurred patterns, inaccurate garment structures, and inconsistent appearances across viewpoints. In contrast, DCAT directly generates garments according to open-ended text instructions and propagates the edited appearance through the garment-adaptive dual-geometry carrier, producing clearer details and more consistent multi-view results without requiring an external garment image.