

# GaussianVTON: 3D Human Virtual Try-ON via Multi-Stage Gaussian Splatting Editing with Image Prompting

Haodong Chen<sup>1</sup> Yongle Huang<sup>1</sup> Haojian Huang<sup>2</sup> Xiangsheng Ge<sup>1</sup> Dian Shao<sup>1†</sup>

<sup>1</sup>Northwestern Polytechnical University, <sup>2</sup>The University of Hong Kong

## Abstract

The increasing prominence of e-commerce has underscored the importance of Virtual Try-On (VTON). However, previous studies predominantly focus on the 2D realm and rely heavily on extensive data for training. Research on 3D VTON primarily centers on garment-body shape compatibility, a topic extensively covered in 2D VTON. Thanks to advances in 3D scene editing, a 2D diffusion model has now been adapted for 3D editing via multi-viewpoint editing. In this work, we propose GaussianVTON, an innovative 3D VTON pipeline integrating Gaussian Splatting (GS) editing with 2D VTON. To facilitate a seamless transition from 2D to 3D VTON, we propose, for the first time, the use of only images as editing prompts for 3D editing. To further address issues, e.g., face blurring, garment inaccuracy, and degraded viewpoint quality during editing, we devise a three-stage refinement strategy to gradually mitigate potential issues. Furthermore, we introduce a new editing strategy termed Edit Recall Reconstruction (ERR) to tackle the limitations of previous editing strategies in leading to complex geometric changes. Our comprehensive experiments demonstrate the superiority of GaussianVTON, offering a novel perspective on 3D VTON while also establishing a novel starting point for image-prompting 3D scene editing. Project page: <https://haroldchen19.github.io/gsvton/>

## 1. Introduction

The rise of e-commerce and online shopping has led to significant advancements in recommendation systems [17, 19, 38, 88] and visual or keyword-based search [2, 57, 67, 98]. From the perspective of potential clothing consumers, accurately identifying their needs is crucial. However, determining the suitability of the clothing before making a purchase decision is equally significant. Generally,

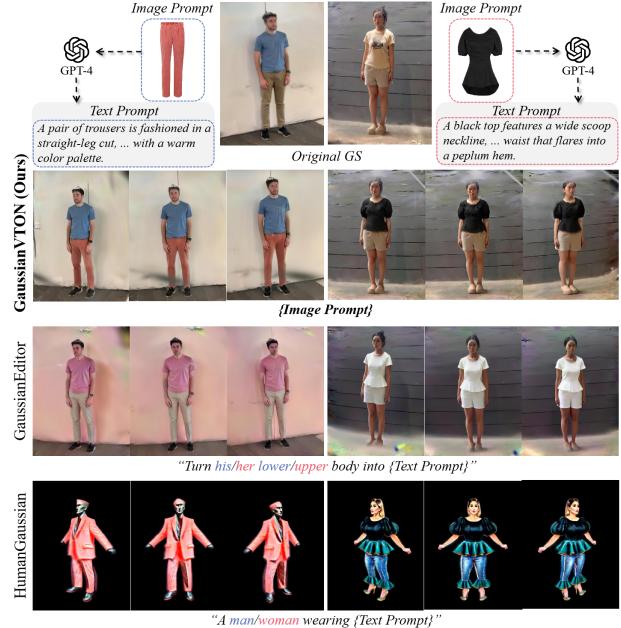


Figure 1. GaussianVTON enables efficient human-environment interaction in try-on applications by reconstructing and editing multi-view images. Our method for the first time employs image prompting to achieve more precise and customized 3D Gaussian Splatting editing. Unlike existing works that rely on text prompts for 3D editing, e.g., GaussianEditor [12], our method avoids erroneously replacing clothing and affecting other areas of the garment, as well as causing changes in other elements like background and facial features. Furthermore, compared to text-driven 3D clothed human generation or reconstruction works, e.g., HumanGaussian [56], our method is based on real human images, avoids resulting in odd body shapes, and aligns with the prompts.

consumers typically rely on product images, model photos, and buyer reviews often leads to uncertainty. Virtual Try-On (VTON) [1, 15, 23, 30, 33, 41, 68, 69, 91], facilitated by advancements in Generative Adversarial Networks (GANs) [29] and Latent Diffusion Models (LDMs) [80], al-

<sup>†</sup>Corresponding authors.

lows users to virtually try on garments using catalog images of persons and clothing items. However, existing research [70, 71, 89, 101] primarily focuses on image-based (2D) VTON and lacks robustness with custom user images [66]. While traditional 3D VTON [42, 61, 86, 87] prioritizes garment design rationality and stylistic effects over authentic human body simulation, which has extensively covered in 2D VTON. These limitations restrict the convenience that VTON currently provides to users.

Recently, the rise of implicit 3D representations, represented by Neural Radiation Fields (NeRF) [63], marks a paradigm shift in 3D scene rendering [39, 44, 53, 74, 96]. This has also spurred research aimed at user-friendly 3D representations and editing algorithms [4, 21, 35, 90], which achieve 3D editing by employing a 2D diffusion model to edit each multi-viewpoint. Unlike conventional reconstruction or generation [10, 18, 48, 51, 56] of 3D clothed humans using meshes or SMPL models [58], reconstructing 3D scenes using real multi-view human images can yield more realistic results and mitigate issues, *e.g.*, odd proportions, face blurring, etc., as shown in the bottom of Fig. 1. With its real-time rendering and explicit point cloud-like representation, the recent advent of Gaussian Splatting (GS) [45] further empowers 3D scene editing techniques, rendering them more flexible and rapid. Whether NeRF or GS-based 3D scene editing, current methods predominantly allow for editing exclusively through text prompts [12, 22, 35, 47], relying on 2D diffusion models (*e.g.*, InstructPix2Pix [9]) for iterative editing of images from each viewpoint. However, text-driven editing usually leads to discrepancies between the users' expectations and the actual output, especially in tasks like VTON that require specific editing for certain areas as demonstrated in Fig. 1.

To this end, our work is dedicated to advancing 3D scene editing, specifically by devising a novel VTON pipeline. We diverge from previous methods, which primarily enabled the editing within a scene through text prompts [22, 35, 47]. Instead, we adopt image prompts for personalized editing based on user-specified garment images, thereby achieving a more human-centric experience, given that people tend to focus more on images rather than textual descriptions when shopping online. However, the development of this work also faces several challenges. A primary issue is the inconsistency in the editing results across different viewpoints when employing a 2D diffusion model for editing multi-view images within a 3D scene [12, 47]. If edits are guided by text prompts for just clothing color (*e.g.*, "Turn his pants red.") or style (*e.g.*, "Turn him into a clown."), the diffusion model tends to output somewhat similar results across each viewpoint, though not entirely consistent [22, 35]. Nonetheless, VTON necessitates customized editing of specific regions based on specific images, posing a significant challenge to achieving

consistency across multi-view. Additionally, existing 2D VTON models fail to maintain robustness on custom user data [1, 15, 23, 33, 69]. Even within trained datasets, issues including facial blurring and garment inaccuracies arise.

In this work, we propose GaussianVTON, a novel 3D Virtual Try-On framework leveraging Gaussian Splatting editing. GaussianVTON can achieve highly realistic virtual try-ons through 3D scene editing, eliminating the need for additional pre-training data for the 2D VTON model. To ensure optimal multi-view human image editing results, we adopt the state-of-the-art 2D VTON model, LaDI-VTON [69] as the image editing model, and a swift and controllable 3D editing framework, GaussianEditor [12] as the 3D editing pipeline. To handle the mentioned various issues and challenges, we propose a three-stage refinement strategy: a) In Stage-1, we devise a simple yet effective facial optimization method aimed at precisely detecting facial keypoints and generating corresponding fusion masks. This addresses the prevalent issue of facial blurring during the editing of 2D VTON models, particularly when handling custom data. b) In Stage-2, to mitigate the inconsistencies arising from the diffusion model during multi-view editing and to improve the robustness of the 2D VTON model against custom data, we propose *hierarchical sparse editing* to automatically identify views with suboptimal editing quality from all initially edited views for further refinement. c) In Stage 3, to enhance the quality of editing quality and mitigate potential issues like degradation in viewpoint quality, we draw inspiration from prior works [47, 79] and employ DDNM [95] to optimize each viewpoint.

Additionally, to address the issues posed by the widely used [12, 22, 35, 65] editing strategy, *i.e.*, Iterative Dataset Update (IterativeDU), which often leads to complex geometric changes or the addition of objects in unspecified regions during the editing process, we propose a novel strategy termed Editing Recall Reconstruction (ERR). ERR incorporates multi-view images for rendering in a manner consistent with the initial reconstruction process.

To summarize, our contributions are as follows:

- We propose GaussianVTON, a novel 3D Virtual Try-On pipeline leveraging Gaussian Splatting editing via 2D VTON. To the best of our knowledge, GaussianVTON represents the first image prompting 3D editing framework.
- We have specifically designed a three-stage refinement strategy aimed at effectively mitigating various disturbances encountered in the process from 2D to 3D editing. Additionally, to better address potential challenges, *e.g.*, complex geometric changes, we propose a novel editing strategy termed Edit Recall Reconstruction (ERR).
- Extensive experiments validate the efficacy of GaussianVTON, offering novel insights for both 3D VTON and 3D editing. We believe that our research findings dis-

tinctly highlight the potential of employing 3D editing in VTON, aligning more closely with user preferences and thus establishing a promising starting point for future research.

## 2. Related Work

**2D Virtual Try-On.** Image-based Virtual Try-On (VTON) aims to transfer a desired garment onto the corresponding region of a target subject while preserving human pose and identity. Most existing works [1, 15, 16, 20, 27, 34, 36, 41, 49, 50] follow a two-stage generative framework. Some previous works [15, 20, 41, 91, 103, 105] employ neural networks to regress sparse clothing control points in the target image. These points are then fitted to a Thin Plate Spline (TPS) transformation [8] to achieve clothing deformation. Conversely, another research line [1, 16, 27, 34, 36, 69, 100] estimates an appearance flow map to model non-rigid deformation, where the flow map [109] describes the dense correspondence of each pixel in the target image to its corresponding position in the source image. It's noteworthy that prior works often relied on GANs [29] during the generation stage. However, LaDI-VTON [69] represents the first 2D VTON model entirely based on diffusion models [80]. Despite the rapid progress in VTON-related research, their performance on test data still involves issues, *e.g.*, clothing errors, face blurring, etc., not to mention when applied to custom data without pre-training. In this work, we propose a three-stage refinement strategy to achieve better results on custom data editing. Stage-1 addresses problems like face blurring. Stage-2 enhances the consistency of editing through *hierarchical sparse editing*, while Stage-3 mitigates the quality degradation during editing.

**3D Virtual Try-On.** In contrast to 3D human body or clothed human reconstruction [10, 18, 28, 32, 48, 51, 83, 106], which have been extensively studied, 3D VTON presents unique challenges due to the intricate deformations of clothing. Most research [5, 6, 31, 64, 76, 84, 85, 87, 94, 108] in this domain focus on estimating digital representations of 3D clothing and understanding how they deform relative to different body shapes. For instance, UL-NeF [87] introduces a novel neural field-based approach capable of handling multiple garments, while MGN [6] predicts parameterized clothing geometry and situates it on the SMPL model [58], enabling dressing for various body shapes and poses, though constrained to pre-defined digital wardrobes. M3D-VTON [108] employs edited human images for 3D human reconstructions, however, the reconstructed bodies often suffer from deformations. Additionally, Pix2Surf [64] aims to transfer clothing images to the SMPL model by learning dense correspondences between 2D clothing shapes and 3D clothing UV maps. Existing 3D VTON methods primarily focus on garment fitting but lack consistency with real human data, a gap filled by 2D

VTON. To this end, our work pioneers a novel framework for 3D VTON by employing 3D scene editing with image prompting to advance from 2D to 3D VTON.

**3D Editing.** Editing neural fields is challenging due to the complexity of their shape and appearance. EditNeRF [55] is a pioneering work that edits the shape and color of neural fields by adjusting the latent code of the neural field. Additionally, with the application of CLIP [94] across various fields [24, 40, 60, 102, 104], some works [3, 26, 92, 93] also employ CLIP to facilitate editing the scene reconstructed from source images through text prompts. Another research line [72, 77] focuses on predefined template models or skeletons to support operations, *e.g.* repositioning or rerendering within specific categories. InstructN2N (IN2N) [35] proposes a text-driven NeRF editing method termed IterativeDU. It iteratively replaces the reference images originally used for NeRF [63] reconstruction with edited images using a 2D diffusion model, InstructP2P (IP2P) [9]. By applying the reconstruction loss of these iteratively updated images to the input 3D scene, the scene gradually transforms into an edited scene. GaussianEditor [12] applies Gaussian Splatting [45] to IN2N, adopting Gaussian semantic tracking to track target Gaussian values, significantly improving editing speed and controllability. Additionally, recent works [47, 54, 75, 111] also apply SDS [78], DDS [37], and PDS [47] to optimize IterativeDU. However, these optimizations are all devised for text prompts. For tasks needing precise personalization like VTON, image prompting is more effective, since text cannot convey all that a garment can be. Thus, we introduce a novel image-prompts 3D editing framework with a new editing strategy termed Editing Recall Reconstruction (ERR).

## 3. Method

Our GaussianVTON framework is shown in Fig. 2. We start by taking in a reconstructed 3D scene along with its associated data: a series of captured images, their corresponding camera poses, and camera calibration parameters. The main idea revolves around utilizing image prompts to guide the editing process of 3D scenes to achieve virtual try-on. We first introduce our main pipeline, namely the 3D Gaussian Splatting and diffusion-based 2D VTON models in Sec. 3.1. Following that, we present our newly proposed editing strategy, Editing Recall Reconstruction (ERR), along with a novel three-stage refinement strategy in Sec. 3.2.

### 3.1. Background

**3D Gaussian Splatting.** Gaussian Splatting (GS) [45] embodies a direct depiction of a 3D scene utilizing point clouds, which are differentiable and readily projected to 2D splats, enabling rapid  $\alpha$ -blending for rendering purposes. Within this, GS characterizes geometry through a collec-

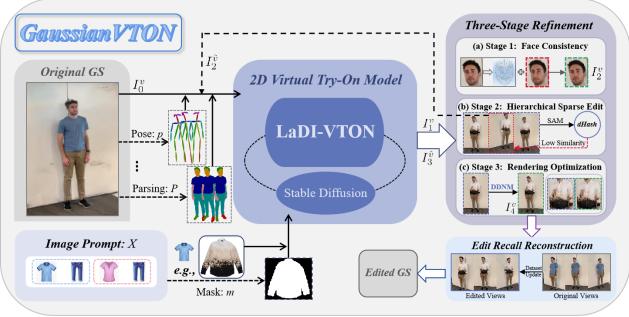


Figure 2. Overall framework of the proposed GaussianVTON.

tion of 3D Gaussians, obviating the need for normal vectors. These Gaussians are delineated by a complete 3D covariance matrix  $\Sigma$ , situated in world space [112] and centered around point  $\mu$ :

$$G(x) = e^{-\frac{1}{2}x^T \Sigma^{-1} x}. \quad (1)$$

where  $x$  denotes each Gaussian as a center point. To render 3D Gaussians, the projection method introduced in [112] is employed, which entails a viewing transformation denoted by  $W$  and the Jacobian  $J$  of the affine approximation of the projective transformation. Employing these elements, the covariance matrix  $\Sigma'$  in camera coordinates is computed as follows:

$$\Sigma' = JW\Sigma W^T J^T. \quad (2)$$

The covariance matrix  $\Sigma$  is analogous to describing the configuration of an ellipsoid, which can be decomposed into a rotation matrix  $\mathbf{R}$  and a scaling matrix  $\mathbf{S}$  for differentiable optimization (the gradient flow is detailed in [112]):

$$\Sigma = \mathbf{R}\mathbf{S}\mathbf{R}^T, \quad (3)$$

To summarize, each Gaussian point in the model is characterized by a set of attributes: its position, denoted as  $x \in \mathbb{R}^3$ , its color represented by spherical harmonic coefficients  $c \in \mathbb{R}^k$  (where  $k$  indicates the degrees of freedom), its opacity  $\alpha \in \mathbb{R}$ , a rotation quaternion  $q \in \mathbb{R}^4$ , and a scaling factor  $s \in \mathbb{R}^3$ . Specifically, the color  $C$  is given by volumetric rendering along a ray with transmittance  $T$ :

$$C = \sum_{i=1}^N T_i \alpha_i \mathbf{c}_i, \quad (4)$$

with  $\alpha_i = (1 - \exp(-\sigma_i \delta_i))$  and  $T_i = \prod_{j=1}^{i-1} (1 - \alpha_j)$ . Specifically, the color and opacity of each Gaussian are computed for every pixel according to the Gaussian representation outlined in Eq. (1). The blending mechanism for  $N$  ordered points overlapping a pixel adheres to a prescribed formula:

$$C = \sum_{i \in N} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j). \quad (5)$$

where  $c_i$  and  $\alpha_i$  signify the color and density of a given point respectively. A Gaussian distribution calculates these values using a covariance matrix  $\Sigma$ , which is subsequently adjusted by per-point opacity and spherical harmonics (SH) color coefficients that are subject to optimization.

**2D Try-On Diffusion Model.** In recent research, numerous endeavors have elevated 2D diffusion processes to 3D editing. One notable focus is multi-view rendering based on a given 3D model, enabling 2D editing through text prompts [9, 13, 79]. This involves generating a dataset of multi-view images to train and guide the 3D models. Our work aims to exploit advanced 3D scene editing techniques to explore the potential for achieving more realistic and personalized 3D VTON via 2D VTON, constituting the first exploration for image-prompting 3D editing. Hence, we have not designed a specific 2D VTON method, implying that any diffusion-based 2D VTON model can be applied in our framework.

To facilitate comprehension, we briefly introduce the 2D VTON model adopted in this work, termed LaDI-VTON [69], which builds upon a latent diffusion model [81] and extends it with an additional autoencoder module. This model harnesses learnable skip connections to augment the generation process while preserving key features. Given an in-shop garment image  $X$ , a target model image  $I$ , and auxiliary inputs including pose keypoints  $p$ , human parsing  $P$ , inpainting mask  $m$ , and dense pose  $d$ .  $X$  undergoes deformation to align with target model pose  $p$ , resulting in warped image  $X_W$ .  $I_M$  is derived by masking  $I$  with  $X_W$ . This setup is common in 2D VTON models.

LaDI-VTON is designed to generate a new image  $\tilde{I}$  by substituting a target garment in  $I$  with  $X$  provided by the user, while preserving the model's physical characteristics, pose, and identity. This approach is rooted in the Stable Diffusion inpainting pipeline [81]. The spatial input  $\gamma$  comprises the channel-wise concatenation of an encoded masked image  $\mathcal{E}(I_M)$ , a resized binary inpainting mask  $m \in \{0, 1\}^{1 \times h \times w}$ , and input from the denoising network  $z_t$ . Specifically,  $I_M$  denotes  $I$  masked according to the inpainting mask  $M \in \{0, 1\}^{1 \times H \times W}$ , and  $m$  is resized based on the spatial dimensions of the original mask  $M$ . Thus, the spatial input for the inpainting denoising network is as follows:

$$\gamma = [z_t; m; \mathcal{E}(I_M)] \in \mathbb{R}^{(4+1+4) \times h \times w}. \quad (6)$$

In summary, LaDI-VTON relies on a latent diffusion model [81], where the variables consist of latent images generated by encoding an RGB image. Likewise, to generate an RGB image from the diffusion model, it is necessary to decode the predicted  $\tilde{z}_0$  latent variables using the decoder  $\tilde{I} = \mathcal{D}(\tilde{z}_0)$ .

### 3.2. GaussianVTON

Given a reconstructed 3D GS scene along with the corresponding dataset of calibrated images  $I^v$  with view index  $v$ , and an image prompt of the target garment  $X$  for try-on, we employ fine-tuning to adjust the reconstructed scene based on editing instructions. This process results in an edited version of the GS scene, which is presented within the framework as illustrated in Fig. 2.

In this work, we employ LaDI-VTON [69] to edit each image  $I^v$  within the dataset. As mentioned in Sec. 3.1, 2D VTON requires various inputs, some of which cannot be directly obtained from everyday life. Such intricate input requirements also pose challenges when utilizing custom data. Hence, to ensure the user-friendly purpose of 3D editing, we devised a solution within GaussianVTON to automatically extract relevant input data, thus necessitating only the provision of the target model image  $I$  and the target garment image  $X$  as inputs (more details are provided in Sec. 4.1). In the upcoming part, we present our newly proposed editing strategy termed Editing Recall Reconstruction (ERR) and the three-stage refinement strategy aimed at effectively mitigating various issues encountered during the editing process, thereby facilitating better integration between 2D VTON models and 3D editing methods.

**Edit Recall Reconstruction.** Edit Recall Reconstruction (ERR) refers to the rendering of the entire dataset during editing at the same time, in a similar manner as the reconstruction process. At the onset of optimization, our image dataset comprises the original images  $I_0^v$ , where 0 denotes the editing step remains 0. These images are cached independently and employed as conditioning inputs for the diffusion model throughout all stages. Subsequently, ERR proceeds sequentially according to the viewpoints  $v$ , using the target garment  $X$  as the prompt to perform editing on all  $I_0^v$  images via LaDI-VTON. Upon completion of editing and refinement for the entire dataset, the dataset update occurs, followed by rendering for this iteration.

Differs from the editing strategy, IterativeDU [35], commonly used in previous studies [12, 21, 35], which iteratively replaces and updates dataset images. ERR waits until all images in the dataset have been edited and refined before updating the dataset, effectively alleviating inconsistent editing issues, *e.g.*, complex geometric alterations and the addition of objects in unspecified regions caused by inconsistent editing moments (further analysis in Sec. 4.4). Meanwhile, compared to recent applications, *e.g.*, SDS [78], DDS [37] and PDS [47], ERR demonstrates the ability to accommodate image editing prompts without inducing fundamental changes in the edited views compared to the input views.

**Three-Stage Refinement.** As previously discussed, both 2D editing and 3D editing suffer from various drawbacks such as multi-view inconsistency, partial blurring (*e.g.* face

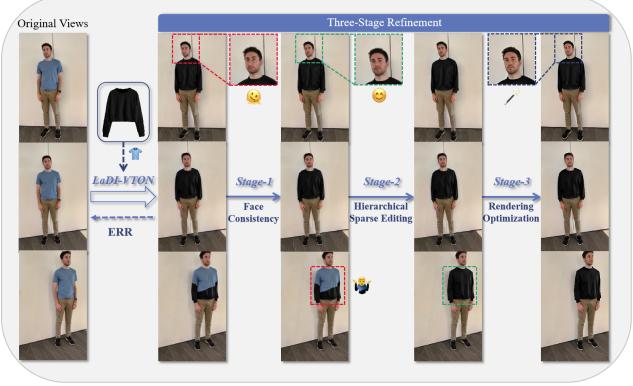


Figure 3. **Three-Stage Refinement.** The three-stage refinement strategy we devised demonstrates sequential mitigation of prominent issues encountered when utilizing 2D VTON models (*i.e.* LaDI-VTON) for image editing, including facial blurring, garment inaccuracies, and degradation in image quality.

blurring), and degradation of image quality due to diffusion models [9, 69, 81]. To mitigate these issues, GaussianVTON adopts a three-stage refinement strategy outlined as follows (specific examples are shown in Fig. 3):

*a) Stage-1: Face Consistency.* Face blurring is a pervasive yet significant issue encountered in various computer vision tasks, including VTON tasks [1, 15, 16, 20, 69, 108], pose transfer tasks [7, 14, 107, 110], and 2D/3D human reconstruction or generation tasks [10, 18, 43, 48, 51, 106]. However, in tasks like 3D scene reconstruction, facial features within the scene are not heavily affected [12, 21, 22, 35, 47]. Thus, to address this issue in GaussianVTON, which is mainly caused by VTON, we first employ the FaceMesh model from MediaPipe [59] to initialize a facial keypoint detector  $K(\cdot)$ . For each input view image  $I_0^v$ , upon detecting facial information within this view, facial keypoints are extracted and stored as the facial network  $F^v = K(I_0^v)$ . Subsequently, after the view  $I_0^v$  undergoes editing via LaDI-VTON, resulting in  $I_1^v$ , a matching process is employed between the pre- and post-edited views according to the index of view  $v$ . Finally, the extracted facial network  $F^v$  is transferred to the post-edited view  $I_1^v$ , where facial distortion is more likely to occur.

*b) Stage-2: Hierarchical Sparse Editing.* In addition to face blurring, the inconsistency in editing between multi-views is also a major challenge faced by GaussianVTON during the editing process. Previous works [35, 47] have confirmed that this issue arises primarily due to the inability of diffusion models [81] to achieve consistency. However, not only that, VTON typically relies on a large amount of training data to achieve relatively satisfactory results on relevant data, but issues like clothing errors still exist. Therefore, to mitigate the impact of the aforementioned problems, especially on custom data for GaussianVTON, we propose *hier-*

archical sparse editing.

In the context of post-editing view images after face refinement, denoted as  $I_2^v$ , we employ Lang-Segment-Anything based on Segment Anything (SAM) [46] as our segmentation model  $S(\cdot)$  to obtain masks  $M_W^v = S(I_2^v)$  corresponding to the edited garment regions (*e.g.*, upper clothing, lower clothing). Subsequently, we retrieve the corresponding images  $I_{M_W}^v$  based on these masks. Sequentially, following the order of view  $v$ , we acquire every three segmented garment images  $I_{M_W}^i$ ,  $I_{M_W}^j$ ,  $I_{M_W}^k$  associated with view indices  $v = i$ ,  $v = j$ ,  $v = k$  in order, respectively. We then employ the *dHash* algorithm to compute the similarity between these three images, and identify the view index  $\hat{v}$  of the image with the lowest similarity to the other two images:

$$\begin{aligned}\hat{v} &= \min_{i \neq j \neq k} (\max(dHash(I_{M_W}^i, I_{M_W}^j), \\ &\quad dHash(I_{M_W}^i, I_{M_W}^k), dHash(I_{M_W}^j, I_{M_W}^k))).\end{aligned}\quad (7)$$

This approach aids in detecting potential errors in the edited garment, providing an approximate assessment of garment fidelity. After traversing all segmented garment images  $I_{M_W}^v$  in the dataset, indices  $\hat{v}$  corresponding to all low-similarity images are obtained. Subsequently, the perspective image  $I_2^{\hat{v}}$ , corresponding to the post-face refinement one, is re-edited with the target garment  $X$  as the image prompt to achieve the refinement.

*c) Stage-3: Rendering Optimization.* Given the possibility of image quality deterioration or distortion during editing, we are inspired by previous research [47, 79] which opt to employ SDEdit [62] to denoise images. In this work, we adopt DDNM [95], which shows promising performance in image restoration via denoising diffusion null-space model:

$$\mathbf{y} = \mathbf{Ax} + \mathbf{n}, \quad \mathbf{n} \in \mathbb{R}^{d \times 1} \sim \mathcal{N}(\mathbf{0}, \sigma_y^2 \mathbf{I}), \quad (8)$$

where  $\mathbf{x} \in \mathbb{R}^{D \times 1}$  represents the original image,  $\mathbf{A} \in \mathbb{R}^{d \times D}$  denotes the degradation matrix, and  $\mathbf{y} \in \mathbb{R}^{d \times 1}$  represents the degraded image. Through this approach, it becomes feasible to refine the rendering  $\tilde{I}_3^{\hat{v}}$  of the image after re-editing and  $\tilde{I}_2^{v-\hat{v}}$  that has not undergone hierarchical sparse editing, into a more visually realistic representation. This ultimately yields the final perspective views  $I_4^v$  to update the dataset.

## 4. Experiments

### 4.1. Implementation Details

We adopt the highly optimized renderer implementation from Gaussian Splatting [45] for Gaussian rendering, as proposed in [12]. For the input multi-view human data, the original 3D Gaussians employed in this work are trained using the method outlined in [45]. Subsequently, we employ the 2D VTON model, LaDI-VTON [69], pre-trained

on DressCode [68], without further training on the multi-view data we use. To simplify the inputs, we only use the multi-view human images as target model images and target garment images as image prompting. GaussianVTON automatically preprocesses these data, including obtaining the pose keypoints of the target model using OpenPose [11], extracting the human parsing using [52], and obtaining the target human dense pose with Detectron2 [99]. We commit to releasing all the data and code of this work. Additionally, we mostly adhere to the hyperparameter settings outlined in GaussianEditor [12] and LaDI-VTON [69]. Moreover, previous works [12, 35] enhance the editing performance by cropping multi-view images to a limited size, *e.g.* 512×512. To ensure the overall editing effect and controllable area of custom data remains unaffected, we employ adaptive sizing editing on all input data.

### 4.2. Qualitative Comparisons

As discussed earlier, existing works on 3D scene editing [12, 35] rely solely on text prompts for editing (*e.g.*, "Turn his short sleeves black.") with InstructPix2Pix [9]. Therefore, for a fairer comparison, given a garment image, we first adopt Multimodal Large Language Model (MLLM), *i.e.* GPT-4 [73] to describe the garment, and then input the corresponding description in a specific format (*e.g.*, "Turn his upper body into {Garment Description}.") into text-driven 3D editing models [12, 35]. More details can be found in the Supplementary Material.

As depicted in Fig. 4, although we describe our desired garments as comprehensively as possible through the use of GPT-4 [73], methods [12, 35] edited based on text prompts fail to meet expectations. Not only do they inaccurately represent clothing in corresponding areas, but they also alter garments in other regions (*e.g.*, upper body). Additionally, editing results in unintended changes in the overall scenario (*e.g.*, background), facial features, and view quality (especially InstructN2N [35]). However, our GaussianVTON directly employs garment images as image prompts, which enables more personalized editing of humans in 3D scenes compared to methods utilizing text prompts. This further underscores the limitations of text-driven 3D editing when confronted with tasks like Virtual Try-On, which necessitate specific edits in particular regions. To further demonstrate the superiority of our novel 3D VTON framework, GaussianVTON, in this task, we demonstrate additional editing results in Fig. 5.

### 4.3. Quantitative Comparisons

Following [25, 35, 82], we first present a quantitative comparison of CLIP Text-Image Directional Similarity (Text-Image) and CLIP Image-Image Directional Similarity (Image-Image). Similarly, to ensure a fair comparison, we use corresponding garment descriptions instead of image



**Figure 4. Qualitative Comparison.** We ask GPT-4 [73] to generate detailed descriptions of target garments, followed by format “*Turn his upper body into ...*” as the text prompt for InstructN2N [35] and GaussianEditor [12]. We adopt GSEditor-iN2N from GaussianEditor as the comparative model due to its superior performance.

prompts to compute Text-Image. Subsequently, inspired by the generation tasks[56, 97], we further employ several evaluation metrics for image quality (*i.e.*, FID, SSIM, PSNR and LPIPS) to evaluate the edited multi-views. Adhering to the principle of human-centricity, we also conduct human evaluations on the editing results. As demonstrated in Tab. 1, GaussianVTON depicts superior performance across all evaluation metrics, which implies that our approach not only ensures relatively precise editing but also effectively manages to mitigate issues like image distortion. Notably, our editing results exhibit a substantial preference over the baseline methods in human evaluation. More details are available in the Supplementary Material.

#### 4.4. Ablation Study

As demonstrated in the left panel of Fig. 6, we further compare our newly proposed editing strategy, namely ERR, with IterativeDU (which is widely used in the text-driven 3D editing framework). It can be observed that, compared to IterativeDU, ERR effectively mitigates complex geometric changes, *e.g.*, missing or deformed regions in garments, by simultaneously updating and rendering, rather than updating the dataset sequentially from each viewpoint.

Furthermore, as illustrated in Fig. 3, we have exemplified specific instance views employing the three-stage refinement strategy. To gain a deeper insight into the impact of potential challenges encountered during the transition from 2D to 3D editing, we conduct ablation experiments on the rendering results of this strategy, depicted in the right panel of Fig. 6, which can gradually address various issues, *e.g.*, face blurring, garment inaccuracies and quality degradation.

To provide quantitative analysis, we utilize the same evaluation metrics as presented in Tab. 1 to further scrutinize our proposed strategies, as demonstrated in Tab. 2. It is noteworthy that following Stage-1 (face consistency), the metrics at the pixel level (*i.e.*, FID, SSIM, PSNR, and LPIPS) reach their optimal values for multi-viewpoints, indicating the achievement of facial optimization to maintain consistency with the original view. However, garment optimization (Stage-2) has not yet been performed, thus preserving more elements of the original attire, as depicted in the right panel of Fig. 6. Building upon both qualitative and quantitative ablation analyses, we further substantiate that both ERR and the three-stage refinement strategy underscore the effectiveness of our GaussianVTON framework. This effectiveness is attributed to its innovative optimiza-



Figure 5. **Extensive Results of GaussianVTON.** To further validate the efficacy of our framework, we also employ multi-view image data of a female, which further substantiates the superiority and capability to adopt custom data of GaussianVTON.

Table 1. **Quantitative Comparisons.** We compare against the text-driven 3D scene editing techniques *i.e.*, InstructN2N [35] and GaussianEditor-iN2N [12].

	Text-Image↑	Image-Image↑	FID↓	SSIM↑	PSNR↑	LPIPS↓	User Study↑
InstructNeRF2NeRF [35]	0.1600	0.6879	295.8	0.7321	14.02	0.3110	15.70%
GaussianEditor-iN2N [12]	0.2071	0.7558	195.4	0.8091	16.61	0.1686	10.68%
GaussianVTON (Ours)	<b>0.3293</b>	<b>0.8481</b>	<b>176.1</b>	<b>0.8171</b>	<b>18.00</b>	<b>0.1654</b>	<b>73.62%</b>

tion designs, which facilitate direct image-prompting 3D editing.

#### 4.5. Limitations

Although GaussianVTON shows decent performance, it inherits limitations of the 2D try-on diffusion model within 3D editing. Similar to most existing 2D VTON models, the 2D try-on diffusion model confronts challenges due to its reliance on extensive training data and narrow test-

ing datasets, hindering practical application. This situation poses unique challenges for developing 3D VTON systems tailored to custom user data, as explored in this work. Moreover, while GaussianVTON also centers on image-guided 3D editing, VTON cannot encompass its entirety. Future work will advance this framework, contributing to more comprehensive and personalized image-prompting 3D editing.

Table 2. **Ablation Studies.** We adopt the same metrics in Tab. 1 to further evaluate the effect of our proposed ERR and the three-stage refinement strategy.

	Text-Image↑	Image-Image↑	FID↓	SSIM↑	PSNR↑	LPIPS↓
IterativeDU	0.2876	0.8156	197.3	0.7866	17.52	0.2175
ERR	<b>0.3293</b>	<b>0.8481</b>	<b>176.1</b>	<b>0.8171</b>	<b>18.00</b>	<b>0.1654</b>
w/o Three-Stage	0.2151	0.7621	142.1	0.8407	17.95	0.1697
After Stage-1	0.2481	0.7829	<b>136.8</b>	<b>0.8443</b>	<b>18.25</b>	<b>0.1563</b>
After Stage-2	0.2713	0.8294	162.6	0.8281	18.12	0.1695
After Stage-3	<b>0.3293</b>	<b>0.8481</b>	176.1	0.8171	18.00	0.1654



Figure 6. **Ablation Studies.** This figure depicts the results of our proposed editing strategy, ERR, compared to IterativeDU from IN2N [35] (shown on the left). On the right, the rendering results of our Three-Stage Refinement process are showcased: Stage-1 maintains facial consistency, Stage-2 refines garments using hierarchical sparse editing, and Stage-3 optimizes overall image quality.

## 5. Conclusion

In this work, we propose GaussianVTON, a novel 3D Virtual Try-On (VTON) pipeline leveraging 3D Gaussian Splatting editing, which represents a significant advancement in both image-prompting 3D editing and 3D VTON. Our method enables realistic try-on experiences for users through the reconstruction and editing of real scenes. To address the challenges inherent in transitioning from 2D to 3D editing, our method employs a three-stage refinement strategy. Furthermore, we introduce a specialized editing strategy termed Edit Recall Reconstruction (ERR), which enhances rendering smoothness and prevents undesirable artifacts resulting from complex geometry alterations.

## References

- [1] Shuai Bai, Huiling Zhou, Zhihang Li, Chang Zhou, and Hongxia Yang. Single stage virtual try-on via deformable attention flows, 2022. [1](#), [2](#), [3](#), [5](#)
- [2] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto del Bimbo. Composed image retrieval using contrastive learning and task-oriented clip-based features, 2023. [1](#)
- [3] Chong Bao, Yinda Zhang, Bangbang Yang, Tianxing Fan, Zesong Yang, Hujun Bao, Guofeng Zhang, and Zhaopeng Gao. Semantic-driven image-based nerf editing with prior-guided editing field, 2023. [3](#)
- [4] Edward Bartrum, Thu Nguyen-Phuoc, Chris Xie, Zhengqin Li, Numair Khan, Armen Avetisyan, Douglas Lanman, and Lei Xiao. Replaceanything3d:text-guided 3d scene editing with compositional neural radiance fields, 2024. [2](#)
- [5] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Pbns: Physically based neural simulator for unsupervised garment pose space deformation, 2021. [3](#)
- [6] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images, 2019. [3](#)
- [7] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model, 2023. [5](#)
- [8] F.L. Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 1989. [3](#)
- [9] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. [2](#), [3](#), [4](#), [5](#), [6](#)
- [10] Yukang Cao, Kai Han, and Kwan-Yee K. Wong. Sesdf: Self-evolved signed distance field for implicit 3d clothed human reconstruction, 2023. [2](#), [3](#), [5](#)

- [11] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields, 2019. 6
- [12] 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, 2023. 1, 2, 3, 5, 6, 7, 8
- [13] Yiwen Chen, Chi Zhang, Xiaofeng Yang, Zhongang Cai, Gang Yu, Lei Yang, and Guosheng Lin. It3d: Improved text-to-3d generation with explicit view synthesis, 2023. 4
- [14] Soon Yau Cheong, Armin Mustafa, and Andrew Gilbert. Uppt: Universal diffusion model for person image generation, editing and pose transfer, 2023. 5
- [15] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization, 2021. 1, 2, 3, 5
- [16] Ayush Chopra, Rishabh Jain, Mayur Hemani, and Balaji Krishnamurthy. Zflow: Gated appearance flow-based virtual try-on with 3d priors, 2021. 3, 5
- [17] Guillem Cucurull, Perouz Taslakian, and David Vazquez. Context-aware visual compatibility prediction, 2019. 1
- [18] Lu Dai, Liqian Ma, Shenhan Qian, Hao Liu, Ziwei Liu, and Hui Xiong. Cloth2body: Generating 3d human body mesh from 2d clothing, 2023. 2, 3, 5
- [19] Lavinia De Divitiis, Federico Becattini, Claudio Baecchi, and Alberto Del Bimbo. Disentangling features for fashion recommendation. *ACM Trans. Multimedia Comput. Commun. Appl.*, 19(1s), 2023. 1
- [20] Haoye Dong, Xiaodan Liang, Bochao Wang, Hanjiang Lai, Jia Zhu, and Jian Yin. Towards multi-pose guided virtual try-on network, 2019. 3, 5
- [21] Jiahua Dong and Yu-Xiong Wang. Vica-nerf: View-consistency-aware 3d editing of neural radiance fields, 2024. 2, 5
- [22] Jiemin Fang, Junjie Wang, Xiaopeng Zhang, Lingxi Xie, and Qi Tian. Gaussianeditor: Editing 3d gaussians delicately with text instructions, 2023. 2, 5
- [23] Emanuele Fenocchi, Davide Morelli, Marcella Cornia, Lorenzo Baraldi, Fabio Cesari, and Rita Cucchiara. Dual-branch collaborative transformer for virtual try-on. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2246–2250, 2022. 1, 2
- [24] Niki Maria Foteinopoulou and Ioannis Patras. Emoclip: A vision-language method for zero-shot video facial expression recognition, 2023. 3
- [25] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators, 2021. 6
- [26] William Gao, Noam Aigerman, Thibault Groueix, Vladimir G. Kim, and Rana Hanocka. Textdeformer: Geometry manipulation using text guidance, 2023. 3
- [27] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows, 2021. 3
- [28] Andrew Gilbert, Marco Volino, John Collomosse, and Adrian Hilton. Volumetric performance capture from minimal camera viewpoints, 2018. 3
- [29] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1, 3
- [30] Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In *Proceedings of the 31st ACM International Conference on Multimedia*. ACM, 2023. 1
- [31] Peng Guan, Loretta Reiss, David A. Hirshberg, Alexander Weiss, and Michael J. Black. Drape: Dressing any person. *ACM Trans. Graph.*, 31(4), 2012. 3
- [32] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision, 2020. 3
- [33] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. Viton: An image-based virtual try-on network, 2018. 1, 2
- [34] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R. Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3
- [35] Ayaan Haque, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. Instructerf2nerf: Editing 3d scenes with instructions, 2023. 2, 3, 5, 6, 7, 8, 9
- [36] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on, 2022. 3
- [37] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score, 2023. 3, 5
- [38] Wei-Lin Hsiao and Kristen Grauman. Creating capsule wardrobes from fashion images, 2018. 1
- [39] Pengchong Hu and Zhizhong Han. Learning neural implicit through volume rendering with attentive depth fusion priors, 2024. 2
- [40] Haojian Huang, Xiaozhen Qiao, Zhuo Chen, Haodong Chen, Bingyu Li, Zhe Sun, Mulin Chen, and Xuelong Li. Crest: Cross-modal resonance through evidential deep learning for enhanced zero-shot learning. *arXiv preprint arXiv:2404.09640*, 2024. 3
- [41] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzènes. Do not mask what you do not need to mask: a parser-free virtual try-on, 2020. 1, 3
- [42] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Benet: Learning body and cloth shape from a single image, 2020. 2
- [43] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation, 2022. 5
- [44] Takuhiro Kaneko. Mimo-nerf: Fast neural rendering with multi-input multi-output neural radiance fields, 2023. 2
- [45] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023. 2, 3, 6

- [46] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 6
- [47] Juil Koo, Chanho Park, and Minhyuk Sung. Posterior distillation sampling, 2023. 2, 3, 5, 6
- [48] Hansol Lee, Junuk Cha, Yunhoe Ku, Jae Shin Yoon, and Seungryul Baek. Dynamic appearance modeling of clothed 3d human avatars using a single camera, 2023. 2, 3, 5
- [49] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions, 2022. 3
- [50] Kedan Li, Min jin Chong, Jeffrey Zhang, and Jingen Liu. Toward accurate and realistic outfits visualization with attention to details, 2021. 3
- [51] Mengtian Li, Shengxiang Yao, Zhifeng Xie, and Keyu Chen. Gaussianbody: Clothed human reconstruction via 3d gaussian splatting, 2024. 2, 3, 5
- [52] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 6
- [53] Ru Li, Jia Liu, Guanghui Liu, Shengping Zhang, Bing Zeng, and Shuaicheng Liu. Spectralnerf: Physically based spectral rendering with neural radiance field, 2023. 2
- [54] Yuhang Li, Yishun Dou, Yue Shi, Yu Lei, Xuanhong Chen, Yi Zhang, Peng Zhou, and Bingbing Ni. Focaldreamer: Text-driven 3d editing via focal-fusion assembly, 2023. 3
- [55] Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. Editing conditional radiance fields, 2021. 3
- [56] Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, and Ziwei Liu. Humangaussian: Text-driven 3d human generation with gaussian splatting, 2023. 1, 2, 7
- [57] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 1
- [58] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. *SMPL: A Skinned Multi-Person Linear Model*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023. 2, 3
- [59] Camillo Lugaressi, Jiuqiang Tang, Hadon Nash, Chris McClellan, Esha Ubweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines, 2019. 5
- [60] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval, 2022. 3
- [61] Sahib Majithia, Sandeep N. Parameswaran, Sadhbhavana Babar, Vikram Garg, Astitva Srivastava, and Avinash Sharma. Robust 3d garment digitization from monocular 2d images for 3d virtual try-on systems, 2021. 2
- [62] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations, 2022. 6
- [63] 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. 2, 3
- [64] Aymen Mir, Thiemo Alldieck, and Gerard Pons-Moll. Learning to transfer texture from clothing images to 3d humans, 2020. 3
- [65] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Marcus A. Brubaker, Jonathan Kelly, Alex Levinstein, Konstantinos G. Derpanis, and Igor Gilitschenski. Watch your steps: Local image and scene editing by text instructions, 2023. 2
- [66] Seyed Omid Mohammadi and Ahmad Kalhor. Smart fashion: a review of ai applications in virtual try-on & fashion synthesis. *Journal of Artificial Intelligence*, 3(4):284, 2021. 2
- [67] Davide Morelli, Marcella Cornia, and Rita Cucchiara. Fashionsearch++: Improving consumer-to-shop clothes retrieval with hard negatives. In *Italian Information Retrieval Workshop*, 2021. 1
- [68] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on, 2022. 1, 6
- [69] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on, 2023. 1, 2, 3, 4, 5, 6
- [70] Shuliang Ning, Duomin Wang, Yipeng Qin, Zirong Jin, Baoyuan Wang, and Xiaoguang Han. Picture: Photorealistic virtual try-on from unconstrained designs, 2023. 2
- [71] Yunfang Niu, Dong Yi, Lingxiang Wu, Zhiwei Liu, Pengxiang Cai, and Jinqiao Wang. Pfdm: Parser-free virtual try-on via diffusion model, 2024. 2
- [72] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field, 2021. 3
- [73] OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Donhan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty

Eletti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrei Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeyoung Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong

- Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Jun-tang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2023. 6, 7
- [74] Xiao Pan, Zongxin Yang, Jianxin Ma, Chang Zhou, and Yi Yang. Transhuman: A transformer-based human representation for generalizable neural human rendering, 2023. 2
  - [75] Jangho Park, Gihyun Kwon, and Jong Chul Ye. Ed-nerf: Efficient text-guided editing of 3d scene using latent space nerf, 2023. 3
  - [76] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style, 2020. 3
  - [77] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans, 2021. 3
  - [78] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion, 2022. 3, 5
  - [79] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, Yuanzhen Li, and Varun Jampani. Dreambooth3d: Subject-driven text-to-3d generation, 2023. 2, 4, 6
  - [80] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3
  - [81] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 4, 5
  - [82] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023. 6
  - [83] Nitin Saini, Eric Price, Rahul Tallamraju, Raffi Enficiaud, Roman Ludwig, Igor Martinovic, Aamir Ahmad, and Michael Black. Markerless outdoor human motion capture using multiple autonomous micro aerial vehicles. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 823–832, 2019. 3
  - [84] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization, 2020. 3
  - [85] Igor Santesteban, Miguel A. Otaduy, and Dan Casas. Learning-based animation of clothing for virtual try-on, 2019. 3
  - [86] Igor Santesteban, Nils Thuerey, Miguel A. Otaduy, and Dan Casas. Self-supervised collision handling via generative 3d garment models for virtual try-on, 2021. 2
  - [87] Igor Santesteban, Miguel Otaduy, Nils Thuerey, and Dan Casas. Ulnef: Untangled layered neural fields for mix-and-match virtual try-on. In *Advances in Neural Information Processing Systems*, pages 12110–12125. Curran Associates, Inc., 2022. 2, 3

- [88] Rohan Sarkar, Navaneeth Bodla, Mariya I. Vasileva, Yen-Liang Lin, Anurag Beniwal, Alan Lu, and Gerard Medioni. Outfittransformer: Learning outfit representations for fashion recommendation, 2022. 1
- [89] Sang-Heon Shim, Jiwoo Chung, and Jae-Pil Heo. Towards squeezing-averse virtual try-on via sequential deformation, 2023. 2
- [90] Liangchen Song, Liangliang Cao, Jiatao Gu, Yifan Jiang, Junsong Yuan, and Hao Tang. Efficient-nerf2nerf: Streamlining text-driven 3d editing with multi-view correspondence-enhanced diffusion models, 2023. 2
- [91] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network, 2018. 1, 3
- [92] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields, 2022. 3
- [93] Can Wang, Ruiyang Jiang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Nerf-art: Text-driven neural radiance fields stylization, 2022. 3
- [94] Tuanfeng Y. Wang, Duygu Ceylan, Jovan Popovic, and Niloy J. Mitra. Learning a shared shape space for multimodal garment design, 2018. 3
- [95] Yinhua Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *The Eleventh International Conference on Learning Representations*, 2023. 2, 6
- [96] Zian Wang, Tianchang Shen, Merlin Nimier-David, Nicholas Sharp, Jun Gao, Alexander Keller, Sanja Fidler, Thomas Müller, and Zan Gojcic. Adaptive shells for efficient neural radiance field rendering, 2023. 2
- [97] Sangmin Woo, Byeongjun Park, Hyojun Go, Jin-Young Kim, and Changick Kim. Harmonyview: Harmonizing consistency and diversity in one-image-to-3d, 2023. 7
- [98] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback, 2020. 1
- [99] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 6
- [100] Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning, 2023. 3
- [101] xujie zhang, Xiu Li, Michael Kampffmeyer, Xin Dong, Zhenyu Xie, Feida Zhu, Haoye Dong, and Xiaodan Liang. Warpdiffusion: Efficient diffusion model for high-fidelity virtual try-on, 2023. 2
- [102] Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. When urban region profiling meets large language models, 2023. 3
- [103] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating↔preserving image content, 2020. 3
- [104] Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K. Hong, Woonhyuk Baek, and Byungseok Roh. *CXR-CLIP: Toward Large Scale Chest X-ray Language-Image Pre-training*, page 101–111. Springer Nature Switzerland, 2023. 3
- [105] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10510–10519, 2019. 3
- [106] Tao Yu, Zerong Zheng, Yuan Zhong, Jianhui Zhao, Qionghai Dai, Gerard Pons-Moll, and Yebin Liu. Simulcap : Single-view human performance capture with cloth simulation, 2019. 3, 5
- [107] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan, 2021. 5
- [108] Fuwei Zhao, Zhenyu Xie, Michael Kampffmeyer, Haoye Dong, Songfang Han, Tianxiang Zheng, Tao Zhang, and Xiaodan Liang. M3d-vton: A monocular-to-3d virtual try-on network, 2021. 3, 5
- [109] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A. Efros. View synthesis by appearance flow, 2017. 3
- [110] Xinyue Zhou, Mingyu Yin, Xinyuan Chen, Li Sun, Changxin Gao, and Qingli Li. Cross attention based style distribution for controllable person image synthesis, 2022. 5
- [111] Jingyu Zhuang, Chen Wang, Lingjie Liu, Liang Lin, and Guanbin Li. Dreameditor: Text-driven 3d scene editing with neural fields, 2023. 3
- [112] M. Zwicker, H. Pfister, J. van Baar, and M. Gross. Ewa volume splatting. In *Proceedings Visualization, 2001. VIS '01*, pages 29–538, 2001. 4