Pattern Recognition 175 (2026) 112923

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

FocalGaussian: Improving text-driven 3D human generation with body part focus

Yifan Yang [a,b,*], Zeshuai Deng [c], Dong Liu [c], Zixiong Huang [c], Kai Zhou [c], Hailin Luo [c], Qing Du [c,d,*], Mingkui Tan [c,d]

[a] Electric Power Research Institute, CSG, 510663, Guangzhou, China
[b] Guangdong Provincial Key La-boratory of Power System Network Security, China
[c] School of Software Engineering, South China University of Technology, Guangzhou, China
[d] Key Laboratory of Big Data and Intelligent Robot (South China University of Technology), Ministry of Education, Guangzhou, China

ARTICLE INFO

Keywords:
3D Gaussian splatting
Diffusion model
Digital human

ABSTRACT

Text-driven 3D human generation significantly reduces manual labor for professionals and enables non-professionals to create 3D assets, facilitating applications across various fields, such as digital games, advertising, and films. Conventional methods usually follow the paradigm of optimizing 3D representations such as neural radiance field and 3D Gaussian Splatting by Score distillation Sampling (SDS) using a diffusion model. However, existing methods struggle to generate delicate and 3D consistent human body parts, primarily due to the ignorance of imposing stable topology control and precise local view control. Our key idea is to focus on the critical components of the human body parts to impose precise control while optimizing the 3D model. Following this, we propose FocalGaussian. Specifically, to generate delicate body parts, we propose a focal depth loss that recovers delicate human body parts by aligning the depth of local body parts in the 3D human model and SMPL-X at local and global scales. Moreover, to achieve 3D consistent local body parts, we propose a focal view-dependent SDS that emphasizes key body-part features and provides finer control over local geometry. Extensive experiments demonstrate the superiority of our FocalGaussian across a variety of prompts. Critically, our generated 3D humans accurately capture complex features of human body parts, particularly the hands. For more results please check our project page at Project page.

1. Introduction

Creating personalized 3D virtual avatars [1,2] is crucial for content generation across various applications such as virtual try-on, Games, and VR. Users desire the ability to customize characters, ranging from altering their specific identities (such as clothes, hair, *etc.*) to applying particular artistic styles. However, traditional methods for creating digital avatars often require skilled artists to undertake tedious and time-consuming work such as multi-view captures [3], texturing [4], and animating [5]. Recent advancements have demonstrated more user-friendly solutions, *i.e.*, automatically creating high-quality 3D avatars solely through natural language descriptions.

Recent advances [1,6,7] typically utilize text-to-image (T2I) generation tools and combine 3D human priors to create a 3D human from a text prompt. Among them, the CLIP-based methods [6] tend to exploit the semantic information of humans from CLIP models [8] to benefit the 3D Human Generation. Since CLIP models primarily focus on aligning the high-level semantic information between text and images, they often struggle with accurately generating fine-grained details, particularly regarding human features and clothing specified in the text prompt. Recently, considering the strong capabilities of diffusion models [9] in text-to-image synthesis, most researchers have sought to use the powerful diffusion model to facilitate 3D Human Generation [1,10]. They usually use a Score Distillation Sampling (SDS) loss [11] to distill knowledge from diffusion models to optimize 3D models such as 3D Gaussian Splatting (3DGS) [12] or Neural Radiance Fields (NeRF) [13]. However, these methods still suffer from two significant limitations.

First, existing methods tend to generate 3D Humans without delicate local parts like hands and legs [1]. These methods heavily rely on the text-to-image models, such as the diffusion model, which may generate implausible human body parts. The reason is that the T2I models usually do not explicitly model the fine-grained structure of human bodies [14]. For example, as shown in Fig. 1(a), the hand image gen-

* Corresponding authors.
E-mail addresses: youngyif1@gmail.com (Y. Yang), duqing@scut.edu.cn (Q. Du).

https://doi.org/10.1016/j.patcog.2025.112923
Received 31 August 2025; Received in revised form 9 December 2025; Accepted 14 December 2025
Available online 5 January 2026

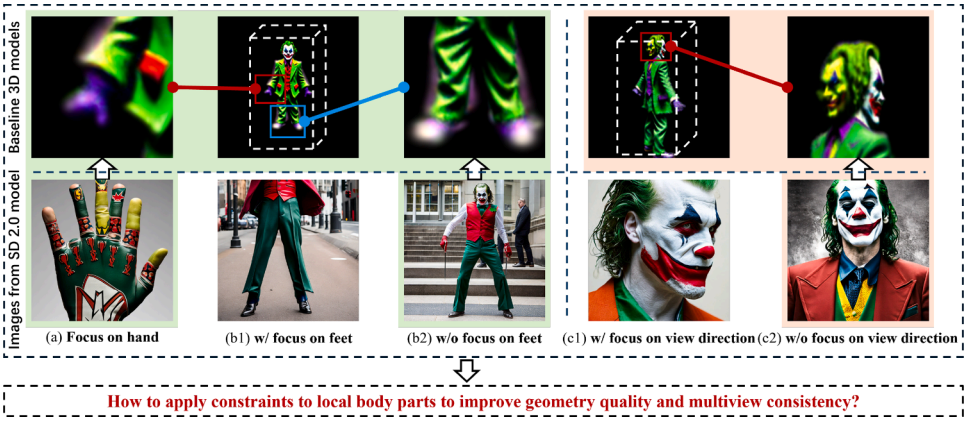

Fig. 1. Motivation of our proposed FocalGaussian. (a) T2I models struggle to synthesize delicate body parts like hands. (b) Employing prompt weighting on keywords helps match the generated content and text prompt. (c) Adding view direction while generating local body parts provides a more precise view control. T is the text prompt "a Joker" in this case.

erated by the Diffusion model has six fingers. Most existing methods tend to overlook the influence of these wrong object information on 3D human body generation, resulting in insufficient precise local limbs for 3D human generation. Although mesh-based methods [10] preserve hand structure from SMPL-X [15], they struggle to model detailed topology for hair and accessories [1]. Therefore, how to avoid the impact of inaccurate information and improve the geometry of body parts is still an open question.

Second, a text prompt without careful design may not control the T2I model to generate high-quality and text-aligned content [16]. From Fig. 1(b2), despite we focus on legs by using text prompt specifying "*Two legs*", the image produced by Stable Diffusion 2.0 [9] depicts the entire body. The content misalignment leads to blurry local body parts in a 3D human model. Moreover, the misalignment often causes a mismatch between the perspective of images generated by T2I models and the actual camera angle used for rendering. This multi-view misalignment issue may lead to 3D inconsistent body parts (see Fig. 1(c2)). Therefore, how to control the T2I model to output desired information for enhancing the consistency of 3D body parts remains an important issue.

Considering the limitations of existing approaches, it is natural to ask: *How to apply constraints to local body parts to enhance geometric quality and multi-view consistency?* To address this question, we propose FocalGaussian, which focuses on key human body parts when optimizing a 3DGS-based human model with a fixed diffusion model. Specifically, to generate reasonable human body parts that have different size levels, we leverage the geometry prior from SMPL-X [15] and design a focal depth loss to align the SMPL-X depth with the corresponding 3DGS depth at local and global levels. At the local level, we align the hand depth, while at the global level, we align the whole body depth except the hand and head. Considering the value mismatch when aligning depth maps, we introduce Pearson correlation loss [17] to regularize them. Moreover, we propose a visibility strategy to address the occlusion of human body over hands when the camera rotates around the 3D human body. Secondly, to achieve 3D consistency in local body parts, we propose a focal view-dependent SDS loss that focuses on key components and employs fine-grained view control on local body parts. Specifically, to ensure the matching between generated content and text prompt, we design focal view-dependent prompt that emphasizes keywords about the human body parts in the prompt with prompt weighting mechanism. In addition, to achieve 3D-consistent generation without introducing multiple faces, we perform a more precise view control on SDS loss by adding view-related text about local body parts and global body to the original text prompt. Extensive experiments and ablation studies show that our proposed FocalGaussian generates 3D humans with detailed, customized hand topology and legs featuring reasonable shapes and clear joint structures.

We summarize our contributions in three folds:

- A focal paradigm for text-driven 3D human generation. We find that existing methods tend to generate unreasonable or inconsistent human body parts like hands and legs, primarily due to the neglect of focusing on local geometry and precise viewpoint control. To address this, we propose a focal paradigm, which focuses on key components of the 3D human and the text prompt to regularize local geometry and view direction.
- A focal depth loss for recovering delicate human body parts. Unlike existing methods that use coarse control like a human skeleton to guide a diffusion model, we focus on aligning the depth of local body parts in our target 3D human model (*i.e.*3D Gaussian) and SMPL-X, which contains prior information for delicate human limbs. We design a visibility mechanism and introduce the Pearson correlation loss to address the depth value mismatching issue. The focal depth loss enables us to generate delicate human body parts like hands with precise topology.
- A focal view-dependent SDS for addressing the 3D inconsistency problem of local body parts. We design a focal view-dependent prompt with fine view control on diffusion model to promote 3D consistency of local body parts. Empirical results show that our FocalGaussian is able to generate 3D human models that achieve view-consistency in local body parts.

2. Related works

We discuss the related works and divide them into text-to-3D-content generation methods and text-to-3D-avatar generation.

2.1. Text-to-3D generation

Text-to-3D generation aims at generating 3D assets from a text prompt. Some text-to-3D methods directly capture the distribution of 3D data [18] or the reconstructed intermediate features [19] on specific domains. By training on large-scale 3D datasets [20], some recent methods [21] attempt to extend the model's capacity. However, it is non-trivial to curate large-scale 3D datasets, and these methods may not generalize well due to a limited diversity of object categories. More recently, some text-to-3D methods [11] optimize a 3D representation in a per-prompt manner by distilling the rich prior learned in the 2D domain. These methods make significant strides by leveraging score distillation sampling (SDS) to inherit unprecedented rendering quality from

exemplar text-to-image models [22]. Though these optimization-based 2D lifting approaches achieve open-vocabulary 3D generation, they still face challenges due to long processing times. Since the optimization time depends heavily on the 3D representation, some existing methods [7,23] adopt efficient 3D Gaussian Splatting model [12] instead of Neural Radiance Fields (NeRF) [13] for text-to-3D generation.

However, these methods tend to focus on generative various kinds of 3D objects and merely achieve limited performance on 3D human generation. In this paper, we seek to generate 3D humans with complicated topology and detailed texture.

2.2. Text-to-3D-avatar generation

Text-to-3D-Avatar generation typically incorporates human prior like SMPL-X [15] and imGHUM [24] models into general text-to-3D techniques to generate 3D avatars [25]. AvatarCLIP [6] use the CLIP guidance to learn a NeRF representation [26] based on the SMPL-X [15] prior. DreamHuman [27] leverages imGHUM [24] to learn a pose-conditioned NeRF of the human with SDS. However, the NeRF representation suffers from slow training and inference, limiting its practical utility. To address this, researchers use explicit representations like mesh and 3D Gaussian Splatting to enhance computational efficiency. For example, Text2Mesh [28] seek to explore using vertex displacement on a predefined mesh template. Nonetheless, these methods may be difficult to generate diverse shapes due to the fixed topology. To facilitate more diverse character generation, TADA [10] deforms the SMPL-X shape with displacement, jointly optimizing the shape, expression, and displacement. To further enhance geometry and texture quality, HumanGaussian [1] designs a structure-aware SDS to simultaneously optimize the human appearance and geometry to learn a 3D Gaussian representation. However, these methods often fail to capture fine details of the local human body, while our approach ensures reasonable global and local topology of the legs and hands.

3. Preliminaries

3.1. 3D Gaussian splatting

3D Gaussian Splatting (3DGS) [12] represents a static scene with a set of 3D Gaussian primitives which have direction-dependent colors. Take the ith Gaussian primitive as an example, the primitive is parameterized by the center position μ_i and covariance matrix Σ_i, which can be defined as follows:

$$G(p, \mu_i, \Sigma_i) = e^{(-\frac{1}{2}(p-\mu_i)^{\mathrm{T}}\Sigma_i^{-1}(p-\mu_i))}, \tag{1}$$

where p is an xyz position of queried point. During rendering, we project 3D Gaussian primitives onto the image plane and form 2D Gaussian. The color $\mathbf{C}_o$ of a pixel is then computed by α-blending across these 2D Gaussian:

$$\mathbf{C_o} = \sum_{i\in\mathcal{N}} c_i\sigma_i \prod_{j=1}^{i-1}(1-\sigma_j), \sigma_i = \alpha_i G(p, \mu_i, \Sigma_i), \tag{2}$$

where α_i denotes the opacity, $\mathcal{N}$ denotes the number of 2D Gaussians, and c_i is a view-dependent color obtained from spherical harmonics (SH) coefficients [29]. By arranging the pixel color $\mathbf{C}_o$, we obtain the rendered RGB Image $\mathbf{I}_{gs}$. Our approach builds upon 3DGS and designs a focal depth loss to maximize the similarity between SMPL-X hand depth and the corresponding 3D Gaussian depth.

3.2. Score distillation sampling

The diffusion model [9] has recently attracted considerable attention due to its amazing capabilities in generating high-fidelity images using text prompts. Recent attempts try to transfer the knowledge from pre-trained text-to-image diffusion models to 3D scenes synthesis (*e.g.* 3D Human) while not using additional 3D data. The most widely-used approach is Score Distillation Sampling (SDS) proposed in Dreamfusion [11]. Specifically, considering θ as a parameterized 3D scene and $g(\cdot)$ as a differentiable rendering function, we obtain the rendered image $\mathbf{x} = g(\theta)$. To acquire a reasonable 3D scene, we need to push each $\mathbf{x}$ rendered from an arbitrary viewpoint to a plausible sample derived from the guidance diffusion model. To this end, the authors of DreamFusion propose an SDS loss $\mathcal{L}_{\mathrm{SDS}}$, which uses gradient decent to optimize the 3D scene θ:

$$\nabla_\theta \mathcal{L}_{\mathrm{SDS}} = \mathbb{E}_{\epsilon,t}\left[w(t)\left(\epsilon_\phi(\mathbf{z}_t; y, t) - \epsilon\right)\frac{\partial \mathbf{x}}{\partial \theta}\right], \tag{3}$$

where $\epsilon_\phi(\mathbf{x}_t; y, t)$ is a score function from Imagen [30]. $\mathbf{z}_t$, y, and t denote the noisy latent feature of $\mathbf{x}$, text embedding, and timestep, respectively. $\epsilon \in \mathcal{N}(0, \mathbf{I})$ is a Gaussian noise and $w(t)$ is a weighting function. In our study, we devise focal view-dependent SDS loss to optimize our 3D Human model parameterized by $\mathcal{GS}$ for more detailed and 3D consistent human models.

3.3. Skinned multi-person linear model

SMPL (Skinned Multi-Person Linear Model) is a 3D parametric model of the human body, defined by shape parameters $\boldsymbol{\beta}$ and pose parameters $\boldsymbol{\theta}$. It generates a body mesh $\mathbf{V}$ via equation: $\mathbf{V} = \mathbf{T}(\boldsymbol{\beta}, \boldsymbol{\theta})$, where $\mathbf{T}$ is a function that deforms a template mesh based on $\boldsymbol{\beta}$ (shape) and $\boldsymbol{\theta}$ (pose).

SMPL-X extends SMPL by adding facial expressions, hand poses, and body variations. It introduces additional parameters $\boldsymbol{\phi}$ for the face and hands: $\mathbf{V}_{\mathrm{X}} = \mathbf{T}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\phi})$. This richer model allows for more detailed and realistic reconstructions.

SMPL and SMPL-X are effective models for 3D human reconstruction and animation. Recently, numerous works [31,32] have been proposed, focusing on generating SMPL or SMPL-X from images or videos. The objective is to estimate the parameters $\boldsymbol{\beta}$ (shape), $\boldsymbol{\theta}$ (pose), and $\boldsymbol{\phi}$ by minimizing the discrepancy between the model's 3D joints, projected 2D joints, and observed pose and shape parameters.

4. Focal 3D human Gaussian

We create high-quality 3D avatars from text prompts by optimizing a 3D Gaussian with a fixed diffusion model, which presents two challenges. 1) Since it is difficult for pre-trained diffusion models to generate hands with reasonable topology and clear fingers, it is difficult to generate human body parts such as hands and legs that have delicate topology. To address this, we focus on exploiting the geometry prior from SMPL-X, and design a focal depth loss that maximizes the similarity of SMPL-X depth and the corresponding 3D Gaussian at two size levels. 2) The local 3D inconsistency problem emerges due to the limitation of coarse view control on current Score Distillation Sampling (SDS), which makes it difficult to effectively control a 2D diffusion model to generate images from specified views. This deficiency leads to inconsistency in 3D model synthesis, giving rise to issues like the local 3D inconsistency problem. To introduce a fine-grained view condition, we propose a focal view-dependent SDS, which aims to focus on the keywords of the local human torso in the prompt. At the same time, we add view descriptions of the keywords to perform finer view control on the SDS loss. The training pipeline of our proposed FocalGaussian is presented in Fig. 2 and Algorithm 1.

Clarity of Stable Topology Control and Precise Local View Control. Stable topology control refers to preserving the structural integrity of fine-scale anatomical regions (e.g., fingers, limbs) during optimization, ensuring that local geometry maintains proper separation and shape without collapsing or merging. Formally, for a local region Ω, the geometric discrepancy $d_{\mathrm{geo}}(S_t(\Omega), S^{\mathrm{ref}}(\Omega))$ should remain bounded over t. In contrast, precise local view control requires a deterministic and smooth mapping from a target camera view $\mathbf{v}$ to the rendered appearance $\mathcal{R}(\mathbf{x}, \mathbf{v})$, avoiding unintended distortions caused by diffusion models' implicit viewpoint biases. Existing SDS-based methods lack ex-

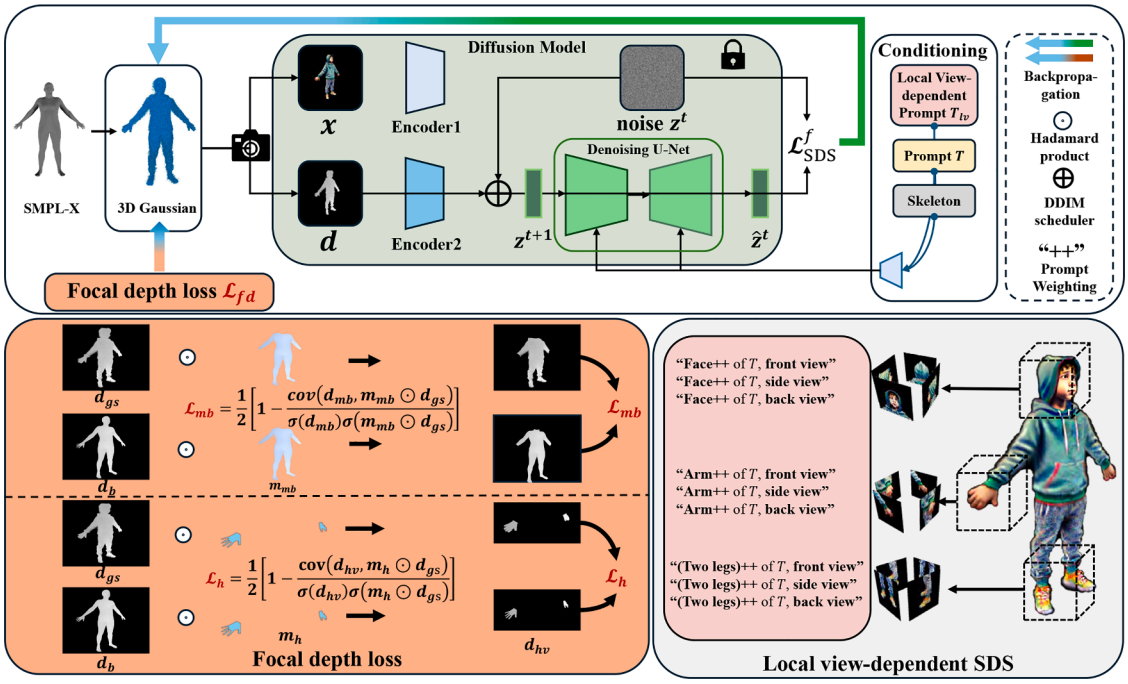

Fig. 2. Overview of our FocalGaussian. Given a text prompt $\mathcal{T}$, we create a realistic 3D avatar with delicate body parts like hands and face that matches $\mathcal{T}$. The key idea of FocalGaussian is to focus on the key components of the human body. We first initialize a 3D Gaussian Splatting (3DGS) model with a SMPL-X model. During training, to recover delicate human parts, we design focal depth loss (*c.f.* Section 4.1) to align the visible the depth of SMPL-X and 3DGS. Moreover, to address the local inconsistency problem, we design a focal view-dependent prompt Score Distillation Sampling loss (*c.f.* Section 4.2) that uses our designed prompts $\mathcal{T}_{LV}$ to up-weight the keywords of body parts and to impose finer view control on a fixed diffusion model.

plicit mechanisms to enforce these two properties, leading to unstable fine-scale topology and inconsistent local geometry across views.

Conceptual Meaning of the *Focal* Paradigm. The term *focal* refers to a unified strategy that selectively amplifies optimization signals on critical body regions and their corresponding semantic cues, rather than uniformly treating the entire human body during SDS optimization. It jointly incorporates (i) *spatial focusing*, which emphasizes geometrically delicate parts through depth alignment with SMPL-X to maintain stable topology; (ii) *view focusing*, which enforces precise viewpoint-dependent control to ensure multi-view consistency of local regions; and (iii) *semantic focusing*, which up-weights key body-part tokens in the prompt to steer the diffusion model's attention toward accurate synthesis of these regions.

4.1. Focal depth loss for recovering delicate human parts

Although existing methods may generate plausible global human body, they struggle to guide the diffusion to produce detailed human body parts. This is particularly challenging for hands due to their intricate topology and frequent self-occlusion. A branch of existing methods [33,34] use explicit 3D representation like a fixed human mesh and only synthesize the corresponding texture. These methods require a high-quality human mesh, which is labor-intensive to obtain. Moreover, the mesh-based methods struggle to model highly-detailed topology like cloth wrinkles and accessories [1]. Another branch of methods [1,27] leverages implicit 3D representation like NeRF and 3D Gaussian to model the desired 3D human. However, these methods struggle to generate delicate and reasonable human body parts like human hands.

To generate a high-quality 3D human with accurate body parts, a practical approach involves using the human body prior from SMPL-X model to establish depth and semantic regularization. However, the challenge arises because the scales of SMPL-X and 3D Gaussian differ, necessitating an alignment method. Additionally, we do not want to use the entire SMPL-X model to constrain the 3D Gaussian, as it represents the anatomy of a naked human without clothing and hair.

From Fig. 2(a), we propose aligning the depth of local human body parts in SMPL-X and the 3D Gaussian with a local Pearson correlation loss. Specifically, to ensure that these two depth maps of different scales are close, we seek the help of normalized negative Pearson correlation loss [17]. Since we aim to regularize the geometry of the human body to be delicate and reasonable, we propose a focal approach at the local

Algorithm 1: The training pipeline of FocalGaussian.

Input: A text prompt $\mathcal{T}$ of a human, a parametric human body model SMPL-X $\mathcal{M}_b$, the hand index Ind_h, the face indexInd_f, a 3D Gaussian model parameterized by $\mathcal{GS}$, a fixed diffusion model $\mathcal{SD}$.

Output: A 3D Gaussian model of the human that aligns with $\mathcal{T}$.

Initializing the 3D Gaussian model $\mathcal{GS}$ with the SMPL-X $\mathcal{M}_b$.
while *step* $n <$ *max iteration* **do**
  Sample a camera matrix $\mathbf{C}_h$ by hybrid sampling in Section 4.3. *# Take batch size = 1 as an example.*
  Get the focal view-dependent prompt $\mathcal{T}_{LV}$ that aligns with $\mathbf{C}_h$. *#e.g. suppose camera is capturing the side view of left arm,* $\mathcal{T}_{LV}$ *= "Arm++ of"* $\mathcal{T}$ *", side view".*
  Render an image $\mathbf{I}_{gs} = \mathcal{GS}(\mathbf{C}_h)$ via Eqs. (1) and (2).
  Render a depth image $\mathbf{d}_{gs}$ from $\mathcal{GS}$ using Eq. (5).
  Obtaining the focal view-dependent SDS loss gradient $\nabla_{\mathcal{GS}}\mathcal{L}^f_{\text{SDS}}$ for updating $\mathcal{GS}$ with $\mathcal{SD}$, $\mathcal{T}_{LV}$, $\mathbf{I}_{gs}$, $\mathbf{d}_{gs}$ and other condition signal using Eq. (9).
  Getting $\mathbf{d}_{hv}$ and m_h following Eq. (4), obtaining $\mathbf{d}_{mb}$ and m_{mb} using $\mathcal{M}_b$, Ind_f, Ind_h and $\mathcal{R}$.
  Obtaining focal depth loss $\mathcal{L}_{fd}$ via Eqs. (6) and (8)
  Update the 3D Gaussian $\mathcal{GS}$ using $\mathcal{L}_{fd}$ and $\nabla_{\mathcal{GS}}\mathcal{L}^f_{\text{SDS}}$.
end

and global size levels.

Aligning depth of hands. Specifically, At the local level, we regularize the depth of hands. To get the depth of hands, we first create the mesh of two hands $\mathcal{M}_h$. To this end, we extract indices Ind_h corresponding to hand vertices from a given whole body SMPL-X [15] mesh $\mathcal{M}_b$ and use Ind_h to obtain $\mathcal{M}_h$ via equation $\mathcal{M}_h = \mathcal{M}_b[\text{Ind}]$. Since there are cases that hands are occluded by body (*e.g.*when the camera is on the person's side), we do not simply render depth from $\mathcal{M}_h$ as we cannot obtain the 3D Gaussian depth of the occluded hand. We seek for a visibility mechanism to obtain only the visible hand depth. Specifically, we first leverage a depth renderer $\mathcal{R}$ to render the whole body depth $\mathbf{d}_b$ and hand depth $\mathbf{d}_h$ from $\mathcal{M}_b$ and $\mathcal{M}_h$, respectively. With $\mathcal{M}_h$, we get the hand mask $\mathbf{m}_h$ and then use the mask $\mathbf{m}_h$ to filter out $\mathbf{d}_b$ except the hands to obtain the final visible hand depth:

$$\mathbf{d}_{hv} = \mathbf{m}_h \odot \mathbf{d}_b, \quad \mathbf{m}_h(i,j) = \begin{cases} 1 & \text{if } \mathbf{d}_h(i,j) \neq 0 \\ 0 & \text{if } \mathbf{d}_h(i,j) = 0 \end{cases}, \tag{4}$$

where $\odot$ is the Hadamard product, i and j are indexes of $\mathbf{d}_h$.

Since we aim to align SMPL-X depth and 3D Gaussian depth w.r.t. hands, we also obtain a per-pixel 3D Gaussian depth of a 3D human via the equation:

$$\mathbf{d} = \sum_{i \in \mathcal{N}_b} d_i \sigma_i \prod_{j=1}^{i-1} \left(1 - \sigma_j\right), \sigma_i = \alpha_i G\left(p, \mu_i, \Sigma_i\right), \tag{5}$$

where $\mathcal{N}_b$ is the number of 3D Gaussian for rendering human body, d_i is the depth of the ith Gaussian at a queried point p. We arrange $\mathbf{d}$ for the depth map $\mathbf{d}_{gs}$. We use the Pearson correlation loss between $\mathbf{d}_{hv}$ and $\mathbf{d}_{gs}$ to align $\mathbf{d}_{hv}$ with $\mathbf{m}_h \odot \mathbf{d}_{gs}$. Here, we propose the focal depth loss regarding hands:

$$\mathcal{L}_h = \frac{1}{2}\left[1 - \frac{\text{cov}(\mathbf{d}_{hv}, \mathbf{m}_h \odot \mathbf{d}_{gs})}{\sigma(\mathbf{d}_{hv})\sigma(\mathbf{m}_h \odot \mathbf{d}_{gs})}\right]. \tag{6}$$

Aligning depth of main body. At the global level, we regularize the main body excluding the face and hands. The reason for excluding the face is the face part of SMPL-X is without hair, while we expect our generated avatar to have customized hair that aligns with the given prompt. To obtain the depth of the main body, we first get the main body mesh $\mathcal{M}_{mb}$. we achieve this by excluding the face and hands with the corresponding index Ind_f and Ind_h of SMPL-X $\mathcal{M}_b$ via equation $\mathcal{M}_{mb} = \mathcal{M}_b[\neg\text{Ind}_\text{f}, \neg\text{Ind}_\text{h}]$. We then obtain the depth d_{mb} of the main body by rendering $\mathcal{M}_{mb}$ with $\mathcal{R}$. We also get the main body mask m_{mb} by judging whether the value of d_{mb} is zero:

$$m_{mb} = \begin{cases} 1 & \text{if } \mathbf{d}_{mb}(i,j) \neq 0 \\ 0 & \text{if } \mathbf{d}_{mb}(i,j) = 0 \end{cases}. \tag{7}$$

Following Eq. (6), we align the depth of the main body by the focal depth loss regarding the main body:

$$\mathcal{L}_{mb} = \frac{1}{2}\left[1 - \frac{\text{cov}(\mathbf{d}_{mb}, \mathbf{m}_{mb} \odot \mathbf{d}_{gs})}{\sigma(\mathbf{d}_{mb})\sigma(\mathbf{m}_{mb} \odot \mathbf{d}_{gs})}\right]. \tag{8}$$

Focal depth loss. By summing $\mathcal{L}_h$ and $\mathcal{L}_{mb}$ as expressed in the equation $\mathcal{L}_{fd} = \mathcal{L}_h + \lambda_0 \mathcal{L}_{mb}$, we obtain our proposed focal depth loss $\mathcal{L}_{fd}$. $\mathcal{L}_{fd}$ aligns SMPL-X depth maps and 3DGS depth maps of different value scales using the Pearson correlation. From Fig. 5 and Section 5.3, $\mathcal{L}_{fd}$ is vital for recovering reasonable human parts.

4.2. Focal view-dependent SDS

To address the local 3D inconsistency problem, existing methods [1,27] employ an SDS loss with global view-dependent prompts, such as "$\mathcal{T}$, side view" to introduce view control when obtaining SDS loss. However, relying solely on global view-conditioned text prompts is inadequate for local and delicate human body parts such as face and hands. Differently, our proposed focal view-dependent SDS only slightly modifies an original prompt $\mathcal{T}$ (see Fig. 2(b)) without additional training, and results in highly detailed 3D human with 3D consistency.

Focal view-dependent control. Given a text prompt $\mathcal{T}$ ="A boy with a beanie wearing a hoodie and joggers", a straightforward way to achieve detailed local human parts generation is to design local prompts $\{\mathcal{T}_L^i\}_{i=1}^n$ w.r.t. local human body part, *e.g.*$\mathcal{T}_L^1$ ="Face of" $\mathcal{T}$, and $\mathcal{T}_L^2$ ="Arm of" $\mathcal{T}$. The $\{\mathcal{T}_L^i\}_{i=1}^n$ are then injected into the diffusion model to generate an SDS loss with local information regarding human body parts. However, in this manner, the camera only focuses on the front view of the human parts, lacking control over synthesizing different viewing angles, refer to Fig. 1(c2).

Differently, We start by designing a set of focal view-dependent prompts $\{\mathcal{T}_{LV}^i\}_{i=1}^n$ that consists of $\mathcal{T}$, a view prompt set $\mathcal{T}_v$ = {*"front view"*, *"back view"*, *"side view"*, *"overhead view"*}, and a local body part prompt set $\mathcal{T}_l$ ={*"Face of"*, *"Arm of"*, *"Two legs of"*}. For instance, $\mathcal{T}_{LV}^1$ = "Face of" $\mathcal{T}$ ", front view" and $\mathcal{T}_{LV}^5$ ="Arm of" $\mathcal{T}$ ", front view". We also rotate, zoom in, and zoom out the camera to align with the view information in $\mathcal{T}_{LV}$. By using $\{\mathcal{T}_{LV}^i\}_{i=1}^n$, we exploit multi-view fine-grained semantic information of human body parts regarding $\mathcal{T}$.

Keywords up-weighting. When the length of $\mathcal{T}$ exceeds a threshold and contains many keywords, the diffusion model struggles to focus on specific targets such as "Face" or "Arm" during human-body synthesis. This occurs because attention becomes dispersed across numerous semantic concepts in $\mathcal{T}$, making it difficult for the model to concentrate on the desired local parts (see Fig. 1(b2)).

To enhance the model's ability to emphasize critical concepts, we adopt the prompt weighting mechanism [35]. Specifically, we increase the magnitude of the token embeddings for target keywords before they are injected into the diffusion model's cross-attention module. In practice, this is implemented by appending "++" to the keyword, which up-weights its embedding (e.g., $\mathcal{T}_{LV}^1$ = "Face++ of" $\mathcal{T}$ "side view"; $\mathcal{T}_{LV}^2$ = "Arm++ of" $\mathcal{T}$ "front view").

Specifically, each "+" scales the corresponding token embedding prior to text encoding, thereby strengthening its influence in the cross-attention layers of the denoising U-Net. The up-weighted keyword thus receives (i) a higher-magnitude embedding leading to stronger attention scores and (ii) increased gradient flow tied to the emphasized concept. As a result, the diffusion model allocates more attention to the highlighted body part and generates its local geometry more faithfully. Combined with our focal view-dependent prompt design, this mechanism helps mitigate local 3D inconsistency by promoting multi-view coherent generation of fine human-body details.

Focal view-dependent SDS loss. Based on the above components, we design a focal view-dependent SDS loss $\mathcal{L}_{\text{SDS}}^f$ that is able to exploit find-grained semantic information regarding delicate human body parts. Specifically, we inject our proposed $\mathcal{T}_{LV}$ into the noise-free SDS [36]:

$$\nabla_{\mathcal{GS}} \mathcal{L}_{\text{SDS}}^f(\mathbf{x}, \mathcal{T}_{LV}, c, t) = w(t)(\delta_{\mathbf{D}}(\mathbf{z}_t, \mathbf{c}, y_{lv}) + \tau_1 \cdot \delta_{\mathbf{cf}}(\mathbf{z}_t, \mathbf{c}, y_{lv}))\frac{\partial \mathbf{x}}{\partial \mathcal{GS}}, \tag{9}$$

$$\delta_{\mathbf{cf}}(\mathbf{z}_t, \mathbf{c}, y_{lv}) = \epsilon_\phi(\mathbf{z}_t; y_{lv}, \mathbf{c}, t) - \epsilon_\phi(\mathbf{z}_t; \varnothing, \mathbf{c}, t), \tag{10}$$

$$\delta_{\mathbf{D}}(\mathbf{z}_t, \mathbf{c}, y_{lv}) = \epsilon_\phi(\mathbf{z}_t; \varnothing, \mathbf{c}, t) - \epsilon_\phi(\mathbf{z}_t; y = p_{\text{neg}}, \mathbf{c}, t) \cdot \mathbf{1}_{\{t \geq 200\}}, \tag{11}$$

where $\delta_{\mathbf{cf}}$ aligns the generated information with the prompt $\mathcal{T}_{LV}$, and $\delta_{\mathbf{D}}$ repels the information away from negative mode. y_{lv} is the text embedding of $\mathcal{T}_{LV}$, $\varnothing$ denotes an empty prompt "", and p_{neg} signifies the negative prompt commonly used to prevent undesired properties. τ_1 is the coefficient to balance $\delta_{\mathbf{D}}$ and $\delta_{\mathbf{cf}}$, $\mathbf{c}$ is a skeleton map from the SMPL-X model $\mathcal{M}_b$. In practice, we use both the $\mathbf{x} = \mathbf{I}_{gs}$ and $\mathbf{x} = \mathbf{d}_{gs}$ for Eq. (9).

4.3. Optimization and training details

Hybrid camera position sampling. The balance between the local body parts and the entire body is crucial when creating a 3D human. The local view enhances the details, while the global perspective ensures accurate global 3D topology and consistency. Focusing on only one perspective can lead to blurry details or global inconsistencies. To achieve the balance, we propose a hybrid camera pose sampling method that randomly samples a batch of hybrid camera matrix $\{\mathbf{C}_h\}_{i=1}^n$ from

Fig. 3. Comparisons with general Text-to-3D and 3D Human Models. Our FocalGaussian generates 3D human with realistic human body parts and 3D consistency. Note that the blurry and unrealistic results are highlighted with **mygreen**green rectangles; the local 3D inconsistency problem are highlighted with *myyellow*orange arrows. Zoom in and refer to demo videos for more details. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

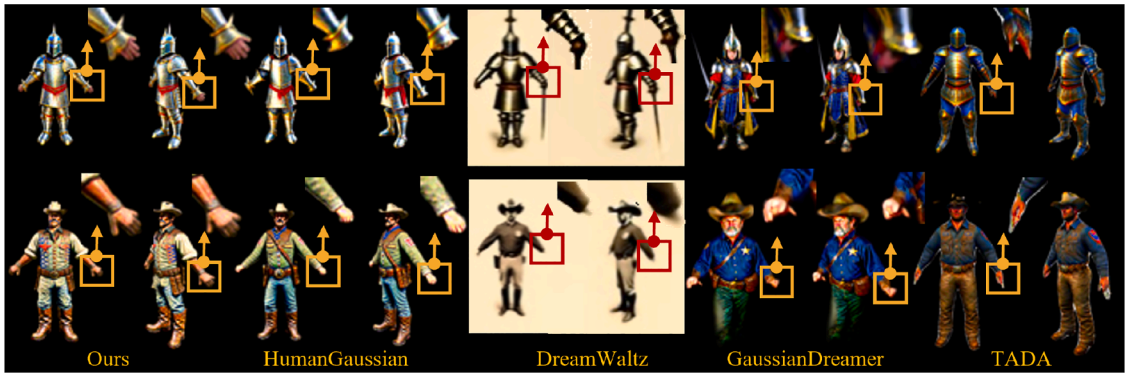

Fig. 4. Zoom-in comparison results, note the blurry results of existing methods in hand region.

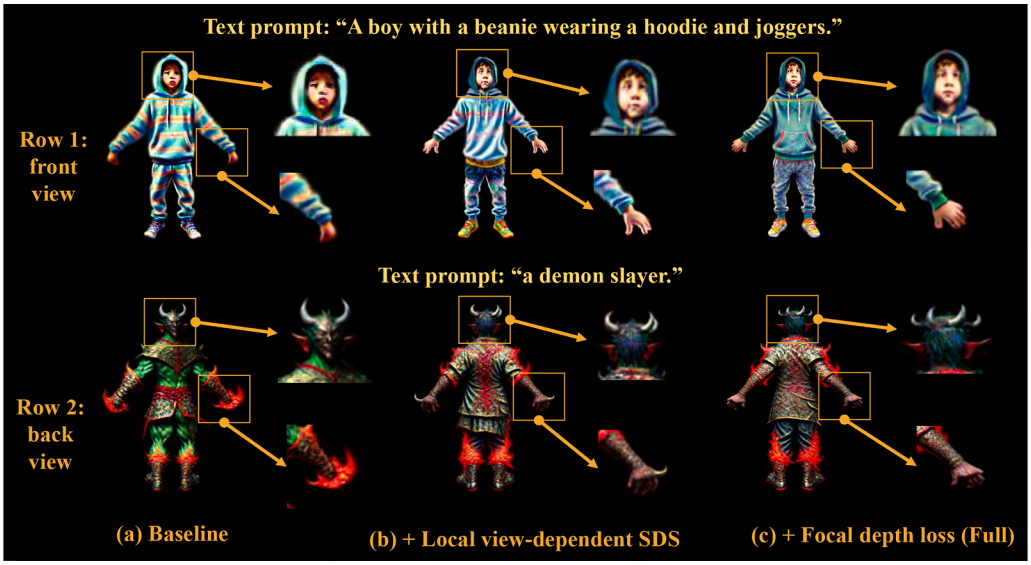

Fig. 5. Ablation studies. Our proposed focal view-dependent SDS. mitigates the local 3D inconsistency problem (Row 2: (a) vs (b)) while the focal depth loss recovers plausible hands ((b) vs (c)).

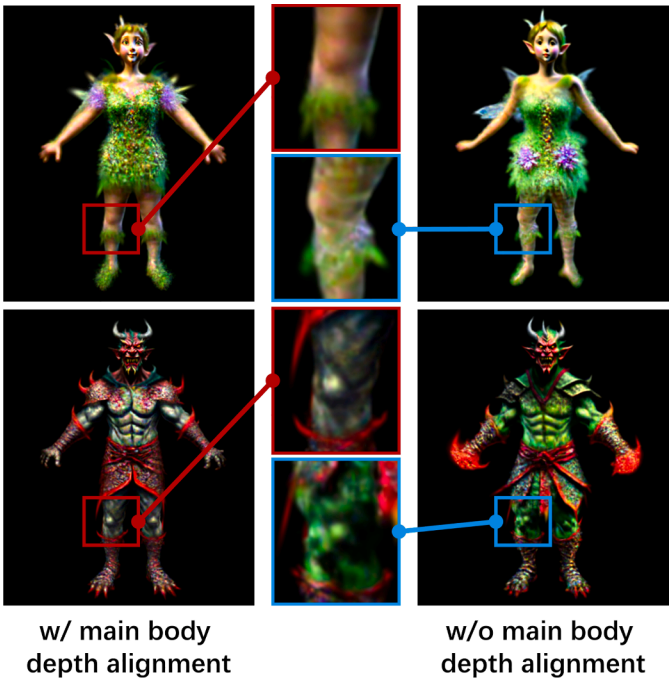

Fig. 6. Effectiveness of aligning the main body depth. Our generated avatars exhibit more reasonable shapes and clearer joint structures.

the combination of a batch of global camera matrices $\{\mathbf{C}_g\}_{i=1}^{n}$ and local camera matrices $\{\mathbf{C}_g\}_{i=1}^{n}$. We obtain the camera position based on the human structure prior from the SMPL-X for initialization. By doing this, $\{\mathbf{C}_h\}_{i=1}^{n}$ emphasizes both local and global views of the human body while training our FocalGaussian.

Local Camera Alignment. We initialize an SMPL-X model that provides prior knowledge of human body part locations. Using this positional information, we set the camera center to target a specific body part. By varying the focal length and azimuth angles within the range [-180, 180], the camera can effectively capture the desired local region.

Camera Sampling Ratio. We balance the sampling between global and local cameras with equal probability. Specifically, local cameras are designed to focus on individual body parts the arms, head, and legs – while the global camera captures the entire body. Each of the four camera types (arms, head, legs, full body) is assigned an equal sampling probability of 0.25.

Overall loss function. We optimize the 3D Gaussian with the following objective function:

$$\mathcal{L}_{\text{overall}} = \lambda_1 \cdot \mathcal{L}_{\text{fd}} + \mathcal{L}^{f}_{\text{SDS}}, \tag{12}$$

where λ_1 is a trade-off parameter for balancing the two losses, see the discussion on λ_1 in Section 5.3.

5. Experiments

5.1. Comparison methods

We compare with the state-of-the-art methods to demonstrate the effectiveness of our FocalGaussian. The first category of methods for comparison is recent general text-to-3D methods, *i.e.* DreamGaussian [7] and GaussianDreamer [23]. DreamGaussian is a generative 3D Gaussian splatting model with mesh extraction and UV-space texture refinement. GaussianDreamer also utilizes 3D Gaussian splatting to bridge 3D and 2D diffusion models. Moreover, we compare with recent text-to-3D human models, *i.e.* DreamHuman [27], TADA [10], and HumanGaussian [1]. DreamHuman's key component is a 2D diffusion model and a deformable, pose-conditioned NeRF constrained by an implicit statistical 3D human pose and shape model [24]. TADA leverages a 2D diffusion model and a parametric body model SMPL-X to achieve text-to-3D human generation. HumanGaussian uses 3D Gaussian splatting and a structure-aware SDS for text-to-3D human generation.

5.2. Implementation details

3D Gaussian Splatting settings. The 3D Gaussians are initialized with 100K instances uniformly sampled on the SMPL-X [15] mesh surface, with opacity of 0.1. The color is represented by Spherical Harmonics (SH) coefficients [37] of degree 0, as described in [12]. The entire 3D Gaussian Splatting training process consists of 3600 iterations, with densification and pruning occurring from iterations 300 to 2100 at 300-step intervals. The prune-only phase occurs from 2400 to 3300 iterations, every 300 steps, with a scaling factor threshold of 0.008. Following HumanGaussian [1], training is conducted using the Adam optimizer with beta values of [0.9, 0.99] and learning rates set to 5×10^{-5}, 10^{-3}, 10^{-2},

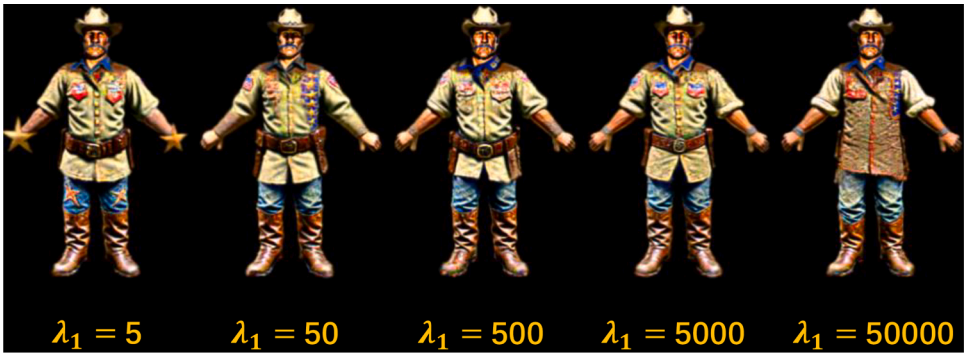

Fig. 7. More visual results using different focal depth loss ratio λ_1. The prompt is "a texas ranger". Zoom in for details on hands.

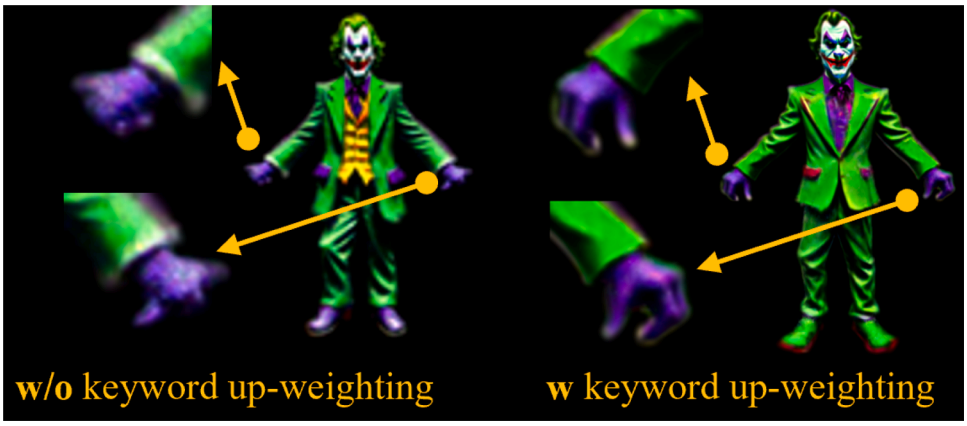

Fig. 8. Ablation studies on keyword up-weighting, both images are without focal depth loss.

1.25×10^{-2}, and 10^{-2} for the center position μ, scaling factor s, rotation quaternion q, color c, and opacity α, respectively.

Training settings. Following HumanGaussian [1], we use the dual-branch diffusion model. We set the DDIM scheduler [38]. We set the camera distance range to $[1.5, 2.0]$, the field of view range to $[40°, 70°]$, the elevation range to $[-30°, 30°]$, and the azimuth range to $[-180°, 180°]$. We use the proposed hybrid camera position sampling during the whole training process to balance between local details and global consistency. The coefficients τ_1 and λ_1 are set to 7.5 and 5000, respectively. All implementations are based on PyTorch and ThreeStudio. We use a training resolution of 1024 with a batch size of 4. The optimization process takes 2.5 h on a single NVIDIA GeForce RTX 3090 (24GB) GPU.

Table 1
User study results. User preference evaluations on generated 3D human models from three aspects: (1) *Texture Quality*; (2) *Geometry Quality*; (3) *Text Alignment*. The best results are highlighted in **bold**.

Methods	Text. Qual. (↑)	Geo. Qual.(↑)	Text Align.(↑)
DreamGaussian [7]	1.41	1.40	1.73
GaussianDreamer [23]	2.83	2.54	3.18
TADA [10]	3.53	3.34	4.15
HumanGaussian [1]	3.89	3.67	4.31
FocalGaussian (Ours)	**4.29**	**4.25**	**4.48**

5.3. Ablation studies

To analyze the effectiveness of sub-modules of our proposed FocalGaussian, we conduct ablation studies over three variants that gradually add the sub-components in Fig. 5: (a) A baseline without using focal depth loss $\mathcal{T}_{LV}$ and focal view-dependent prompt $\mathcal{L}_{fd}$. (b) A variant that adds $\mathcal{T}_{LV}$. (c) our FocalGaussian that adds $\mathcal{L}_{fd}$.

How does focal view-dependent prompt $\mathcal{T}_{LV}$ improve geometry quality and 3D consistency? As shown in Fig. 5(a) and (b), since our $\mathcal{T}_{LV}$ provides more precise view direction control and emphasize on the local body parts of human, we synthesize a more realistic 3D human with clear face, hands, and legs. However, the topology of hand is not reasonable, which has twisted fingers.

How does focal depth loss $\mathcal{L}_{fd}$ improve local details? To recover local body parts with plausible geometry, we qualitatively demonstrate the necessity of focal depth loss in Fig. 5(c). The generated hands have reasonable topology thanks to the $\mathcal{L}_{fd}$ that aligns hand depth of generated 3D Gaussian and SMPL-X. From Fig. 6, aligning the depth of the main body leads to more reasonable shapes of legs.

Impact of focal depth loss We evaluate the impact of focal depth loss by varying the weight ratio λ_1 across a set of values: $5, 50, 500, 5000, 50000$, as illustrated in Fig. 7. Our observation indicates that at lower values of $\lambda_1 = 0.5$ and $\lambda_1 = 50$, the synthesized 3D human models tend to include undesired objects and exhibit blurry body parts in the hands area, respectively. As λ_1 increases, the focal depth loss becomes more effective in driving our FocalGaussian to produce more realistic hand representations. However, excessively high values of $\lambda_1 = 50000$ overly prioritize focal depth loss, resulting in unrealistic clothing texture at the waist. Based on these findings, we select $\lambda_1 = 5000$ as the optimal balance between hand detail, texture quality and overall body consistency.

How does keyword up-weighting improve the quality of human body parts? For challenging cases like "the Joker," where the pre-trained data often include hands holding various objects like playing cards, the generated hands tend to be blurry. We report the results of disabling keyword up-weighting only in Fig. 8, the results show that without the strategy, the hands are blurry. Our keyword up-weighting alleviates this issue and generates clear hands.

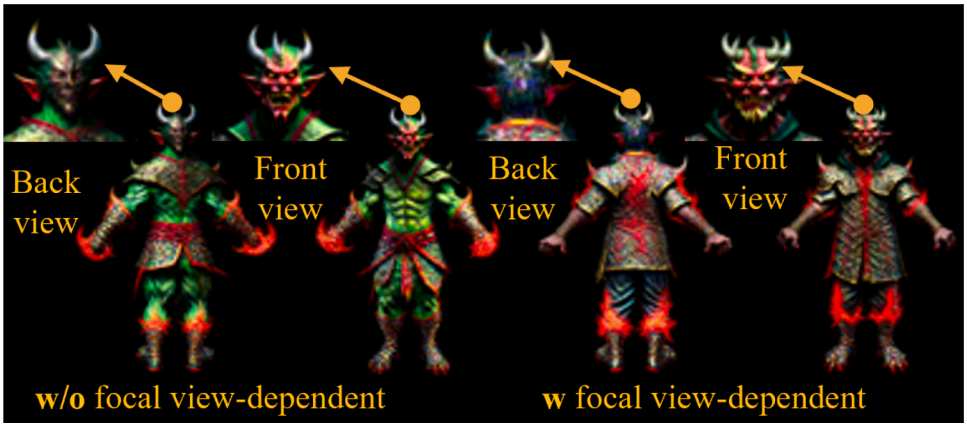

Fig. 9. Effectives of focal view-dependent prompt in addressing the 3D inconsistency problem.

Table 2
We compare our method with HumanGaussian in terms of clip score on hands and legs area. We generate 3D human with higher quality hands and legs. The best results are highlighted in **bold**, view 1 is front view and view 2 is 45 ° side view.

	Left Hand		Right Hand		Legs	
	View 1	View 2	View 1	View 2	View 1	View 2
HumanGaussian [1]	24.7	24.4	25.3	26.1	27.6	27.2
FocalGaussian (Ours)	**27.0**	**25.0**	**27.3**	**27.1**	**28.3**	**28.0**

Table 3
We compare our FocalGaussian with existing methods in terms of CLIP Score between the images and their corresponding prompts. We randomly select 35 prompts for evaluation.

Methods	CLIP Score ↑
DreamWaltz [39]	21.3
HumanGaussian [1]	29.0
DreamGaussian [7]	26.2
GaussianDreamer [23]	27.9
TADA [10]	28.3
FocalGaussian (Ours)	**29.1**

How does the focal view-dependent SDS alleviate the local 3D inconsistency issue? Empirical results in Fig. 9 show that our proposed focal view-dependent SDS helps to generate 3D human models that achieve view consistency in local body parts, especially in the head region.
Why not use focal depth loss for the head region? Our focal depth loss can be applied to specific body parts, like the main body, hands and head, using SMPL-X indices. However, the SMPL-X head model has a balding shape, which overly constrains head geometry, limiting diversity in generating hair and hat shapes. From Fig. 10, applying focal depth loss to the head results in an ellipsoidal shape, which is undesirable for diverse head generation.

5.4. Comparison experimental results

Qualitative comparison. The visual results of our method are shown in Fig. 4. To further verify the effectiveness of our method, we summarize the visual comparisons with recent state-of-the-art methods [1,7,10,23,27] in Fig. 3. We observe that although the general text-to-3D method GaussianDreamer [23] incorporates a 3D diffusion model [18] to enhance their 3D human models, GaussianDreamer and DreamGaussian still struggle to match accurate human body structure. When comparing with text-to-3D human methods, DreamHuman and TADA tend to generate blurry or low-resolution faces. HumanGaussian usually fails to generate plausible hands and legs. Differently, our FocalGaussian achieves more accurate alignment with text descriptions, more realistic appearance textures, and more coherent local body structure. Since DreamHuman [27] does not provide code, we compare with their released demo in **Appx**.
Quantitative comparison. We conducted a user study following previous methods [1,27] to evaluate the quality of the generated 3D human models. For fair comparisons, we randomly select 35 prompts from HumanGaussian and have 18 participants evaluate them. We ask the participants to rate three attributes: Texture Quality, Geometry Quality, and Text Alignment, using discrete scores from 1 to 5, where higher scores indicate better quality. We report the average score for each attribute. As shown in Table 1, the results demonstrate that our method outperforms existing methods in all three evaluation categories. We also compare the CLIP Score [1] metric of local body parts and whole body with existing methods in Tables 2 and 3. The results demonstrate that our generated 3D humans exhibit higher-quality hands, legs and full body.

5.5. More experiments

Customizing the hand pose of the generated avatar. Our focal depth loss plays a crucial role in ensuring that the reconstructed human body parts retain a structure closely resembling that of SMPL-X. Specifically, it enforces spatial consistency and shape accuracy across various body parts. As shown in Fig. 11, we further customize the model's output for the left and right hands, with the left hand reconstructed in a palm pose and the right hand in a fist pose. This customization highlights the versatility of our approach in adapting the model's output to different, specific hand configurations.
Demonstration of the geometry of our generated avatar. To present the geometric details of our generated avatars. We compare the normal of our generated 3D human with that of HumanGaussian. From Fig. 12, Our generated avatar has clearer hand geometry. The prompt is "A texas ranger".
Animation results. Our FocalGaussian is able to generate animatable avatars driven by SMPL-X. We provide animation results of our generated human driven by SMPL-X in the Fig. A.14 of **Appx** and the last video in the https://anonymous.4open.science/w/text23d/project page.
Comparisons of training time. From Table 4, our method is trained on a single RTX 3090 GPU for approximately 43.2 min, which is comparable to or faster than the baselines. Specifically, HumanGaussian reports a training time of about 1 h 4 min on the same GPU, while TADA re-

[1] https://github.com/ taited/clip-score, version 0.1.1 (2023).

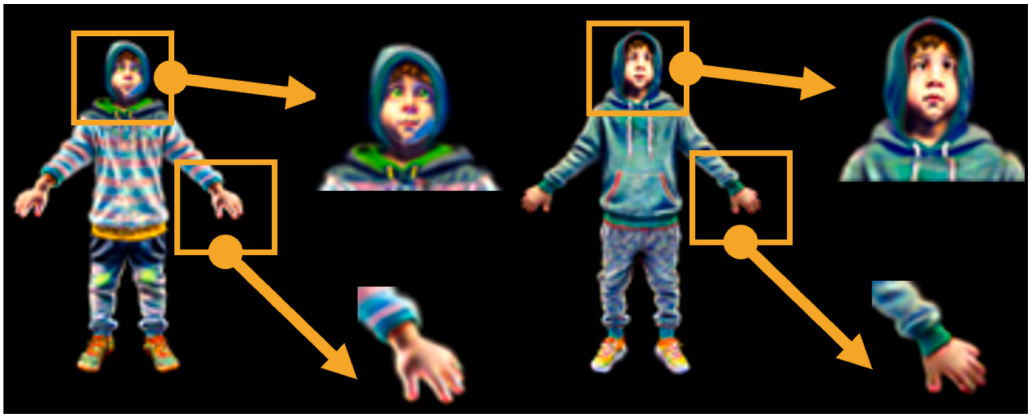

Fig. 10. Leveraging focal depth Loss for head only (left) and hand only (right).

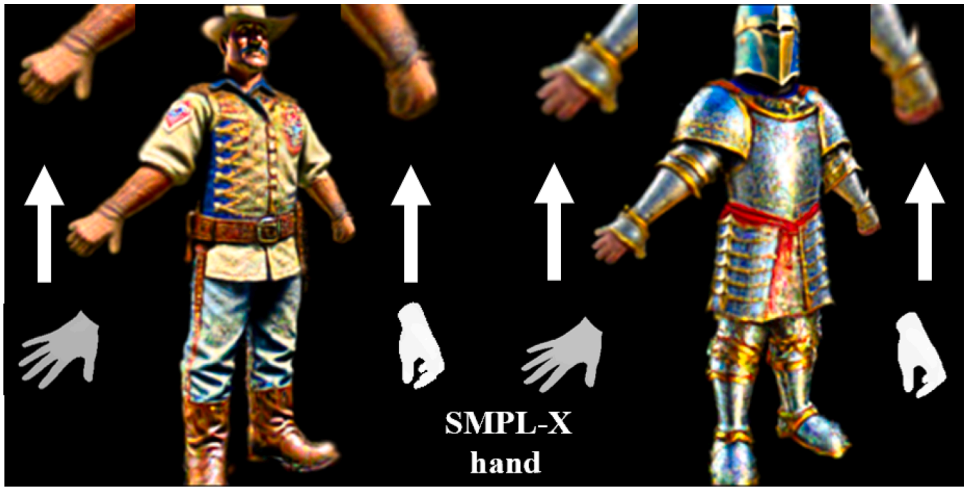

Fig. 11. Transfer hand pose in accordance with the changes of SMPL-X.

Fig. 12. Comparisons with HumanGaussian [1] on Normal map.

quires 3 h 56 min due to its additional diffusion-based refinement stage. Therefore, our method achieves better quality without increasing the computational cost.

Failure cases. We observe that FocalGaussian may struggle with fine-grained descriptions involving complex hand-object interactions (e.g., “a person wearing fingerless gloves holding a pen”) or rare poses (e.g., "A boy with a beanie wearing a hoodie and joggers with crossed legs"), where the diffusion-based supervision provides insufficient depth or appearance cues. From Fig. 13, these limitations primarily stem from the difficulty of generating accurate diffusion depth under such rare configurations. We have incorporated representative examples and corresponding analysis in the main paper to provide a more balanced and comprehensive evaluation of our method.

Table 4
Comparison of training time among different methods on a single RTX 3090 GPU.

Method	GPU	Training Time
HumanGaussian	RTX 3090	1 h 4 min
TADA	RTX 3090	3 h 56 min
Ours	RTX 3090	**43.2 min**

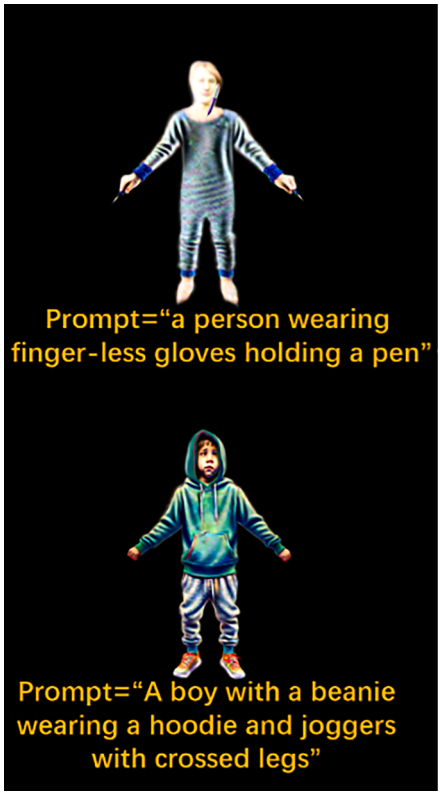

Fig. 13. Failure cases on challenging or conflicting prompts.

6. Discussion and conclusion

Conclusion. Creating high-quality 3D humans from text prompts provides an intuitive and efficient way to generate digital assets for various applications such as gaming, virtual reality (VR), and digital content creation. However, existing text-to-3D methods often struggle to synthesize fine-grained human body parts (e.g., hands and facial details) and suffer from local 3D inconsistency, leading to perceptual artifacts and structural distortions. To overcome these limitations, we first investigate the intrinsic weaknesses of text-to-image diffusion models that underlie such artifacts, particularly their bias in representing delicate body regions and inconsistent spatial depth. Based on these insights, we propose FocalGaussian, a novel framework that enhances both local detail quality and global 3D consistency. Specifically, FocalGaussian incorporates: (1) a focal depth loss designed to accurately reconstruct fine-scale human body parts, and (2) a focal view-dependent prompt with an improved SDS loss that explicitly constrains local 3D geometry to ensure coherent multi-view consistency. Extensive experiments demonstrate that FocalGaussian significantly improves the fidelity and 3D consistency of generated humans compared to state-of-the-art approaches, validating the effectiveness of our design.

Broader impact. The generated 3D human generator could be misused to create disinformation. For example, fabricated 3D humans might be exploited to spread fake news or rumors.

Limitations and future work. Because our method is based on 3D Gaussian splatting, it requires more disk storage than NeRF- or mesh-based approaches. Moreover, it lacks explicit lighting control, so direct adjustments to scene illumination or relighting of generated 3D humans are not currently supported. Our pipeline also does not yet allow refeeding text prompts for region-specific edits (e.g., adding watermarks or changing clothing on an already generated model). We consider such interactive, localized editing a promising and feasible direction and plan to pursue it in future work.

Declaration of generative AI and AI-assisted technologies in the writing process. During the preparation of this work, the authors used ChatGPT in order to polish and correct grammar. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Yifan Yang: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Zeshuai Deng:** Writing – review & editing, Methodology, Investigation, Formal analysis; **Dong Liu:** Writing – review & editing, Visualization, Validation, Methodology, Investigation, Data curation; **Zixiong Huang:** Writing – review & editing, Visualization, Methodology, Investigation, Formal analysis; **Kai Zhou:** Writing – review & editing, Visualization, Validation; **Hailin Luo:** Writing – review & editing, Visualization, Data curation; **Qing Du:** Writing – review & editing, Supervision, Resources, Methodology; **Mingkui Tan:** Writing – review & editing, Validation, Supervision, Resources, Funding acquisition.

Data availability

Data will be made available on request.

Declaration of competing interest

I declare that all the authors have no financial or personal relationships with individuals or organizations that could inappropriately influence or bias the content of this manuscript, titled "FocalGaussian: Improving Text-Driven 3D Human Generation with Body Part Focus" My primary commitment is to uphold the highest standards of academic integrity and impartiality.

Yifan Yang is in Electric Power Research Institute, CSG, China and Guangdong Provincial Key Laboratory of Power System Network Security. Zeshuai Deng, Dong Liu, Zixiong Huang, Kai Zhou, Hailin Luo, Qing Du, Mingkui Tan are in South China University of Technology, they have relationship with pazhou laboratory.

I hereby affirm that the information provided in this Conflict of Interest Statement is accurate and complete to the best of my knowledge.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.patcog.2025.112923

References

[1] X. Liu, X. Zhan, J. Tang, Y. Shan, G. Zeng, D. Lin, X. Liu, Z. Liu, HumanGaussian: text-driven 3D human generation with gaussian splatting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 6646–6657.
[2] Y. Yang, D. Liu, S. Zhang, Z. Deng, Z. Huang, M. Tan, Hilo: detailed and robust 3D clothed human reconstruction with high-and low-frequency information of parametric models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 10671–10681.
[3] J. Luo, J. Tang, T. Tjahjadi, X. Xiao, Robust arbitrary view gait recognition based on parametric 3D human body reconstruction and virtual posture synthesis, Pattern Recognit. 60 (2016) 361–377.
[4] X. Cao, P. Quan, Y. Mao, R. Cao, L. Su, K. Li, TRRS-DM: two-stage resampling and residual shifting for high-fidelity texture inpainting of terracotta warriors utilizing diffusion models, Pattern Recognit. 167 (2025) 111753.
[5] K. Chen, S. Seneviratne, W. Wang, D. Hu, S. Saha, M.T. Hasan, S. Rasnayaka, T. Malepathirana, M. Gong, S. Halgamuge, Anifacediff: animating stylized avatars via parametric conditioned diffusion models, Pattern Recognit. 170 (2026) 112017.
[6] F. Hong, M. Zhang, L. Pan, Z. Cai, L. Yang, Z. Liu, AvatarCLIP: zero-shot text-driven generation and animation of 3D avatars, ACM Trans. Graphics 41 (4) (2022) 1–19.

[7] J. Tang, J. Ren, H. Zhou, Z. Liu, G. Zeng, DreamGaussian: generative gaussian splatting for efficient 3D content creation, in: B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, Y. Sun (Eds.), International Conference on Representation Learning, 2024, 2024, pp. 33879–33896.
[8] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning (ICML), PMLR, 2021, pp. 8748–8763.
[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.
[10] T. Liao, H. Yi, Y. Xiu, J. Tang, Y. Huang, J. Thies, M.J. Black, Tada! text to animatable digital avatars, in: 2024 International Conference on 3D Vision (3DV), IEEE, 2024, pp. 1508–1519.
[11] B. Poole, A. Jain, J.T. Barron, B. Mildenhall, DreamFusion: text-to-3D using 2D diffusion, in: International Conference on Learning Representations (ICLR), 2023.
[12] B. Kerbl, G. Kopanas, T. Leimkühler, G. Drettakis, 3D Gaussian splatting for real-time radiance field rendering, ACM Trans. Graphics 42 (4) (2023) 1–14.
[13] Y. Yang, S. Zhang, Z. Huang, Y. Zhang, M. Tan, Cross-ray neural radiance fields for novel-view synthesis from unconstrained image collections, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15901–15911.
[14] Z. Weng, L. Bravo-Sánchez, S. Yeung-Levy, Diffusion-HPC: synthetic data generation for human mesh recovery in challenging domains, in: 2024 International Conference on 3D Vision (3DV), 2024, pp. 257–267.
[15] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A.A.A. Osman, D. Tzionas, M.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 (CVPR), 2019, pp. 10975–10985.
[16] Y. Hao, Z. Chi, L. Dong, F. Wei, Optimizing prompts for text-to-image generation, in: Advances in Neural Information Processing Systems (NeurIPS), 36, 2024, pp. 66923–66939.
[17] P. Sedgwick, Pearson's correlation coefficient, BMJ 345 (2012), p. e4483.
[18] H. Jun, A. Nichol, Shap-E: Generating conditional 3D implicit functions, arXiv:2305.02463 (2023).
[19] H. Chen, J. Gu, A. Chen, W. Tian, Z. Tu, L. Liu, H. Su, Single-stage diffusion NeRF: a unified approach to 3D generation and reconstruction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.
[20] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, A. Farhadi, Objaverse: a universe of annotated 3D objects, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13142–13153.
[21] Y. Hong, K. Zhang, J. Gu, S. Bi, Y. Zhou, D. Liu, F. Liu, K. Sunkavalli, T. Bui, H. Tan, LRM: large reconstruction model for single image to 3D, in: The Twelfth International Conference on Learning Representations (ICLR), 2024.
[22] Z. Wang, C. Lu, Y. Wang, F. Bao, C. LI, H. Su, J. Zhu, ProlificDreamer: high-fidelity and diverse text-to-3D generation with variational score distillation, in: Advances in Neural Information Processing Systems (NeurIPS), 36, 2023, pp. 8406–8441.
[23] T. Yi, J. Fang, J. Wang, G. Wu, L. Xie, X. Zhang, W. Liu, Q. Tian, X. Wang, GaussianDreamer: fast generation from text to 3D Gaussians by bridging 2D and 3D diffusion models, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6796–6807.
[24] T. Alldieck, H. Xu, C. Sminchisescu, ImGHUM: implicit generative models of 3D human shape and articulated pose, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5461–5470.
[25] Y. Cao, Y.-P. Cao, K. Han, Y. Shan, K.-Y.K. Wong, DreamAvatar: text-and-shape guided 3D human avatar generation via diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 958–968.
[26] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, W. Wang, NeuS: learning neural implicit surfaces by volume rendering for multi-view reconstruction, in: Advances in Neural Information Processing Systems (NeurIPS), 34, 2021, pp. 27171–27183.
[27] N. Kolotouros, T. Alldieck, A. Zanfir, E.G. Bazavan, M. Fieraru, C. Sminchisescu, DreamHuman: animatable 3D avatars from text, in: Advances in Neural Information Processing Systems (NeurIPS), 2023.
[28] O. Michel, R. Bar-On, R. Liu, S. Benaim, R. Hanocka, Text2Mesh: text-driven neural stylization for meshes, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 13492–13502.
[29] R. Green, Spherical harmonic lighting: the gritty details, in: Archives of the Game Developers Conference, 56, 2003, p. 4.
[30] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E.L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al., Photorealistic text-to-image diffusion models with deep language understanding, Adv. Neural Inf. Process. Syst. 35 (2022) 36479–36494.
[31] Y. Liu, Z. Zhang, STGFormer: spatio-temporal graphformer for 3D human pose estimation in video, Pattern Recognit. 171 (2026) 112239.
[32] X. Zhang, Y. Chen, H. Lai, H. Zhang, Weakly supervised 3D human pose estimation based on PnP projection model, Pattern Recognit. 163 (2025) 111464.
[33] J. Yu, H. Zhu, L. Jiang, C.C. Loy, W. Cai, W. Wu, PaintHuman: towards high-fidelity text-to-3D human texturing via denoised score distillation, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 38, 2024, pp. 6800–6807.
[34] E. Richardson, G. Metzer, Y. Alaluf, R. Giryes, D. Cohen-Or, TEXTure: text-guided texturing of 3D shapes, ACM SIGGRAPH 2023 Conf. Proc. (2023), pp.1-11.
[35] Diffusers, Prompt weighting, 2024. https://huggingface.co/docs/diffusers/main/en/using-diffusers/weighted_prompts#prompt-weighting, Last accessed on May 20, 2024.
[36] O. Katzir, O. Patashnik, D. Cohen-Or, D. Lischinski, Noise-free score distillation, in: International Conference on Learning Representations (ICLR), 2023.
[37] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, A. Kanazawa, Plenoxels: radiance fields without neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 5501–5510.
[38] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, in: International Conference on Learning Representations (ICLR), 2021.
[39] Y. Huang, J. Wang, A. Zeng, H. Cao, X. Qi, Y. Shi, Z.-J. Zha, L. Zhang, DreamWaltz: make a scene with complex 3D animatable avatars, in: Advances in Neural Information Processing Systems (NeurIPS), 36, 2024.