X-Oscar: A Progressive Framework for High-quality Text-guided 3D Animatable Avatar Generation

Yiwei Ma*[1] Zhekai Lin*[1] Jiayi Ji[1] Yijun Fan[1] Xiaoshuai Sun[1] Rongrong Ji[1]

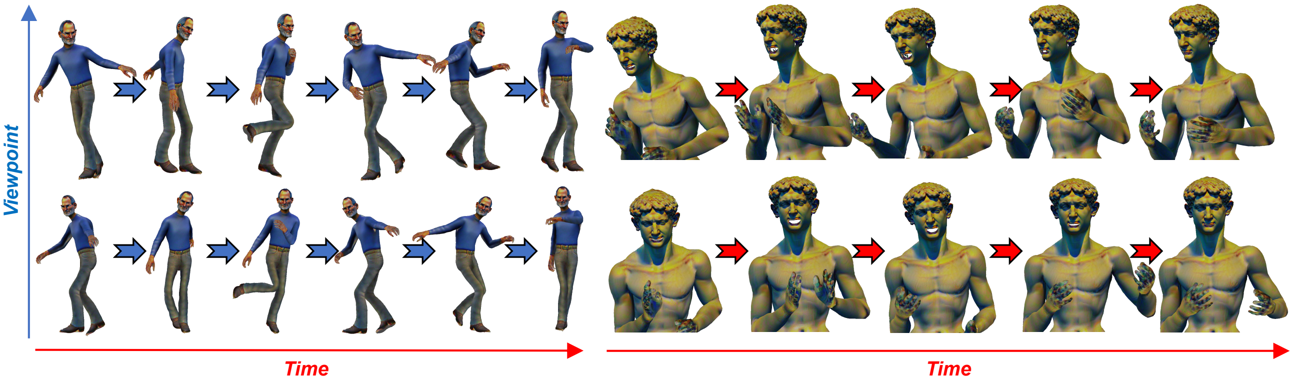

Figure 1: Samples generated by X-Oscar along temporal and viewpoint dimensions. Left Prompt: "Steven Paul Jobs". Right Prompt: "David of Michelangelo".

Abstract

Recent advancements in automatic 3D avatar generation guided by text have made significant progress. However, existing methods have limitations such as oversaturation and low-quality output. To address these challenges, we propose X-Oscar, a progressive framework for generating high-quality animatable avatars from text prompts. It follows a sequential "Geometry→Texture→Animation" paradigm, simplifying optimization through step-by-step generation. To tackle oversaturation, we introduce Adaptive Variational Parameter (AVP), representing avatars as an adaptive distribution during training. Additionally, we present Avatar-aware Score Distillation Sampling (ASDS), a novel technique that incorporates avatar-aware noise into rendered images for improved generation quality during optimization. Extensive evaluations confirm the superiority of X-Oscar over existing text-to-3D and text-to-avatar approaches. Our project page is **https://xmu-xiaoma666.github.io/Projects/X-Oscar/**.

1. Introduction

The creation of high-quality avatars holds paramount importance in a wide range of applications, including cartoon production (Li et al., 2022b; Zhang et al., 2022), virtual try-on (Santesteban et al., 2021; 2022), immersive telepresence (Li et al., 2020a;b; Xiu et al., 2023), and video game design (Zheng et al., 2021; Zhu et al., 2020). Conventional methods for avatar creation are notorious for being time-consuming and labor-intensive, often demanding thousands of hours of manual work, specialized design tools, and expertise in aesthetics and 3D modeling. In this research, we propose an innovative solution that revolutionizes the generation of high-quality 3D avatars with intricate geometry, refined appearance, and realistic animation, solely based on a text prompt. Our approach eliminates the need for manual sculpting, professional software, or extensive artistic skills, thus democratizing avatar creation and making it accessible to a broader audience.

The emergence of deep learning has brought forth a new era in 3D human body reconstruction, showcasing promising methods for automatic reconstruction from photos (Liao et al., 2023b; Han et al., 2023; Men et al., 2024; Zhang

*Equal contribution [1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, School of Informatics, Xiamen University, 361005, P.R. China. Correspondence to: Xiaoshuai Sun <xssun@xmu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

et al., 2023d) and videos (Weng et al., 2022; Jiang et al., 2022). However, these approaches primarily focus on reconstructing human bodies from visual cues, limiting their applicability to real-world scenarios and posing challenges when it comes to incorporating creativity, editing, and control. Recent advancements in large-scale vision-language models (VLM) (Radford et al., 2021; Li et al., 2022a; 2023a; Xu et al., 2023a; Ma et al., 2023b) and diffusion models (Ho et al., 2020; Sohl-Dickstein et al., 2015; Welling & Teh, 2011; Kulikov et al., 2023) have opened up exciting possibilities for generating 3D objects and avatars from text prompts. These methods effectively combine pretrained VLMs and diffusion models with 3D representations such as DeepSDF (Park et al., 2019), NeRF (Mildenhall et al., 2021), DMTET (Shen et al., 2021), and 3D Gaussian Splatting (Kerbl et al., 2023). Despite these promising developments, current approaches still face several limitations. Some methods (Ma et al., 2023c; Chen et al., 2023a; Wang et al., 2023b) focus solely on generating static everyday objects, lacking animation ability. Other methods that aim to generate avatars based on human prior knowledge often suffer from poor geometry and appearance quality (Liao et al., 2023a; Hong et al., 2022; Zhang et al., 2023b) or are incompatible with conventional computer graphics workflows (Liu et al., 2023; Huang et al., 2023b; Cao et al., 2023).

This paper presents X-Oscar, an innovative and advanced framework that leverages text prompts to generate high-quality animatable 3D avatars. Specifically, X-Oscar builds upon the SMPL-X body model (Pavlakos et al., 2019a) as prior knowledge and employs a strategic optimization sequence of “Geometry → Texture → Animation”. To overcome the common challenge of oversaturation during avatar generation, we propose Adaptive Variational Parameter (AVP), a novel technique that utilizes a trainable adaptive distribution to represent the geometry and appearance of the avatars. By optimizing the distribution as a whole instead of focusing on specific parameters, X-Oscar effectively mitigates oversaturation, resulting in visually appealing avatars. Furthermore, we introduce Avatar-aware Score Distillation Sampling (ASDS), an innovative module that incorporates geometry-aware and appearance-aware noise into the rendered image during the optimization process. This strategic approach significantly enhances the visual attributes of the avatars and improves their geometry and appearance quality. Extensive experimentation demonstrates the superiority of X-Oscar over existing methods, showcasing improvements in both geometry and appearance quality. Moreover, the avatars generated by X-Oscar are fully animatable, unlocking exciting possibilities for applications in gaming, animation, and virtual reality.

To summarize, our main contributions are three-fold:

- We present X-Oscar, an innovative and progressive framework that enables the creation of delicate animatable 3D avatars from text prompts.
- To overcome the persistent challenge of oversaturation, we propose Adaptive Variational Parameter (AVP), which represents avatars as adaptive distributions instead of specific parameters.
- We introduce Avatar-aware Score Distillation Sampling (ASDS), an advanced module that incorporates geometry-aware and appearance-aware noise into the rendered image during the optimization process, resulting in high-quality outputs.

2. Related Work

Text-to-3D Generation. The emergence of vision-language models (VLMs) (Radford et al., 2021; Ma et al., 2022) and diffusion models has brought about a revolutionary impact on text-to-3D content generation. Pioneering studies like CLIP-forge (Sanghi et al., 2022), DreamFields (Jain et al., 2022), CLIP-Mesh (Mohammad Khalid et al., 2022), and XMesh (Ma et al., 2023c) have showcased the potential of utilizing CLIP scores (Radford et al., 2021) to align 3D representations with textual prompts, enabling the generation of 3D assets based on textual descriptions. Subsequently, DreamFusion (Poole et al., 2022) introduced Score Distillation Sampling (SDS), a groundbreaking technique that leverages pretrained diffusion models (Saharia et al., 2022) to supervise text-to-3D generation. This approach has significantly elevated the quality of generated 3D content. Building on these foundations, researchers have explored various strategies to further enhance text-to-3D generation. These strategies encompass coarse-to-fine optimization (Lin et al., 2023), conditional control (Li et al., 2023c; Chen et al., 2023b), bridging the gap between 2D and 3D (Ma et al., 2023a), introducing variational score distillation (Wang et al., 2023b), and utilizing 3D Gaussian Splatting (Chen et al., 2023c; Li et al., 2023b; Yi et al., 2023; Tang et al., 2023). Nevertheless, despite these advancements, existing methodologies primarily concentrate on generating common static objects. When applied to avatar generation, they face challenges such as poor quality and the inability to animate the generated avatars. In contrast, our proposed framework, X-Oscar, specifically aims to generate high-quality 3D animatable avatars from text prompts. X-Oscar caters to the unique requirements of avatar generation, including intricate geometry, realistic textures, and fluid animations, to produce visually appealing avatars suitable for animation.

Text-to-Avatar Generation. The domain of text-to-avatar generation (Kolotouros et al., 2024; Zhang et al., 2024; Huang et al., 2023a; Xu et al., 2023b; Zhou et al., 2024) has emerged as a prominent and vital research area to cater to the demands of animated avatar creation. This field incorporates human priors such as SMPL (Loper et al., 2015),

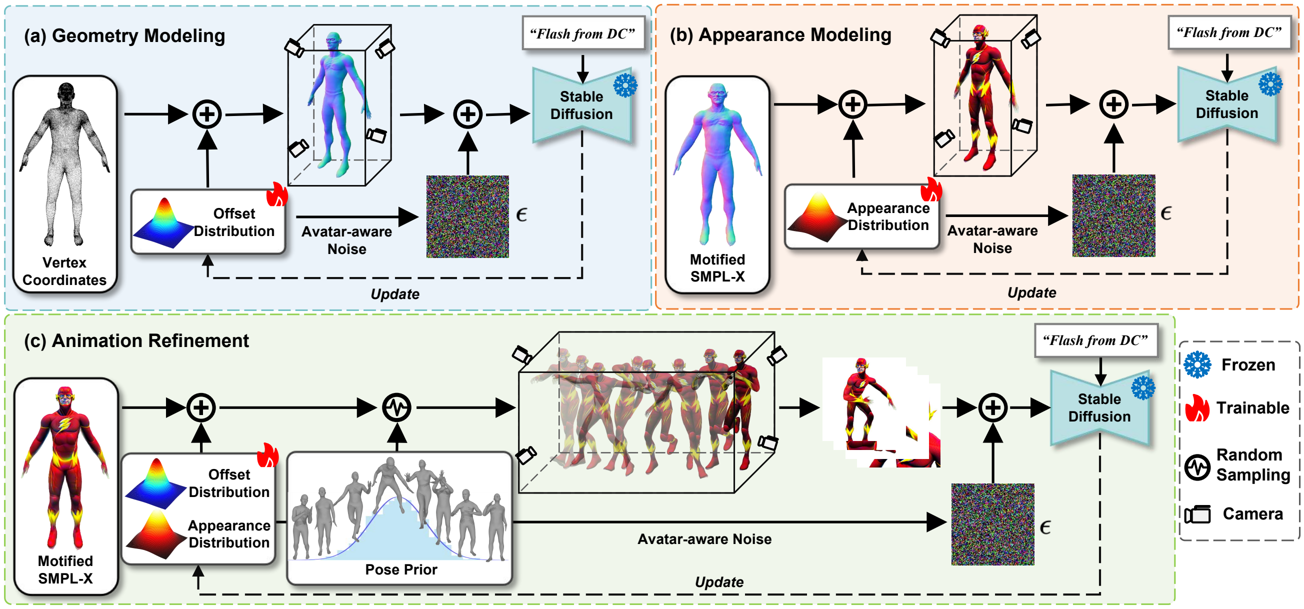

Figure 2: Overview of the proposed X-Oscar, which consists of three generation stages: (a) geometry modeling, (b) appearance modeling, and (c) animation refinement.

SMPL-X (Pavlakos et al., 2019b), and imGHUM (Alldieck et al., 2021) models. AvatarCLIP (Hong et al., 2022) utilizes SMPL and Neus (Wang et al., 2021) models to generate 3D avatars guided by the supervision of CLIP scores. Dreamwaltz (Huang et al., 2023b) introduces NeRF (Mildenhall et al., 2021) to generate 3D avatars based on 3D-consistent occlusion-aware SDS and 3D-aware skeleton conditioning. AvatarBooth (Zeng et al., 2023) leverages dual fine-tuned diffusion models to achieve customizable 3D human avatar generation. AvatarVerse (Zhang et al., 2023a) utilizes ControlNet (Zhang et al., 2023c) and DensePose (Güler et al., 2018) to enhance view consistency. TADA (Liao et al., 2023a) employs a displacement layer and a texture map to predict the geometry and appearance of avatars. HumanNorm (Huang et al., 2023a) proposes a normal diffusion model for improved geometry. HumanGaussian (Liu et al., 2023) uses 3D Gaussian Splatting as human representation for text-to-avatar generation. Despite these advancements, existing methods often produce low-quality and over-saturated results. To overcome these limitations, we introduce a progressive framework that incorporates two key modules, namely Adaptive Variational Parameter and Avatar-aware Score Distillation Sampling. Our framework effectively generates high-fidelity avatars that are visually appealing and realistic.

3. Preliminaries

Score Distillation Sampling (SDS) (Poole et al., 2022), also known as Score Jacobian Chaining (SJC) (Wang et al., 2023a), is a powerful optimization method that adapts pretrained text-to-image diffusion models for text-to-3D generation. Given a pretrained diffusion model $p_\phi(z_t|y,t)$, where ϕ represents the model's parameters, y is the input text prompt, and z_t denotes the noised image at timestep t, SDS aims to optimize a 3D representation to align with the text prompt. The forward diffusion process in SDS is formulated as $q(z_t|g(\theta,c),y,t)$, where θ represents the trainable parameters of the 3D representation, c denotes the camera, and $g(\cdot)$ is the rendering function. The objective of SDS can be expressed as follows:

$$\min \mathcal{L}_{\mathrm{SDS}}(\theta) = \mathbb{E}_{(t,c)}\left[\sqrt{\frac{1-\gamma_t}{\gamma_t}}\omega(t)\mathcal{D}_{\mathrm{KL}}(q(z_t|g(\theta,c),y,t)\parallel p_\phi(z_t|y,t))\right], \tag{1}$$

where $\omega(t)$ is a weighting function dependent on the timestep t, $z_t = \sqrt{\gamma_t}g(\theta,c) + \sqrt{1-\gamma_t}\epsilon$ is the noised image, and $\mathcal{D}_{\mathrm{KL}}(\cdot)$ represents the Kullback-Leibler Divergence (Kullback & Leibler, 1951). To approximate the gradient of the SDS objective, the following equation is leveraged:

$$\nabla_\theta \mathcal{L}_{\mathrm{SDS}}(\theta) \triangleq \mathbb{E}_{t,\epsilon,c}\left[\omega(t)(\underbrace{\hat{\epsilon}_\phi(z_t;y,t)}_{\text{predicted noise}} - \underbrace{\epsilon}_{\text{Guassian noise}})\frac{\partial g(\theta,c)}{\partial\theta}\right], \tag{2}$$

where $\epsilon \sim \mathcal{N}(0,I)$ represents sampled noise from a normal distribution, and $\hat{\epsilon}_\phi(z_t;y,t)$ denotes the predicted noise of the pretrained diffusion model at timestep t.

SMPL-X (Pavlakos et al., 2019b) is a widely adopted parametric 3D human body model in the fields of computer graphics and animation. It offers a comprehensive representation of the human body, consisting of $10,475$ vertices and 54 joints, facilitating detailed and realistic character rendering. By specifying shape $\mathfrak{s}$, pose $\mathfrak{p}$, and expression $\mathfrak{e}$ parameters, the SMPL-X model generates a human body using the following equation:

$$\mathrm{T}(\mathfrak{s},\mathfrak{p},\mathfrak{e}) = \mathcal{T} + B_s(\mathfrak{s}) + B_p(\mathfrak{p}) + B_e(\mathfrak{e}), \tag{3}$$

where $\mathcal{T}$ denotes a standard human template, while $B_s(\cdot), B_p(\cdot), B_e(\cdot)$ represent shape, expression, and pose blend shapes, respectively. These blend shapes deform the template to generate a wide range of body shapes, poses, and expressions. To transition the human body from a standard pose to a target pose, linear blend skinning (LBS) is employed:

$$\mathrm{M}(\mathfrak{s},\mathfrak{p},\mathfrak{e}) = \mathcal{W}_{LBS}(\mathrm{T}(\mathfrak{s},\mathfrak{p},\mathfrak{e}), J(\mathfrak{s}), \mathfrak{p}, W), \tag{4}$$

where $\mathcal{W}_{LBS}(\cdot)$ represents the LBS function, $J(\mathfrak{s})$ corresponds to the skeleton joints, and W represents the skinning weight. The LBS function calculates the final vertex positions by interpolating between the deformed template vertices based on the assigned skinning weights. This process ensures a smooth and natural deformation of the body mesh.

4. Approach

The overview of X-Oscar is depicted in Fig. 2, and the workflow is illustrated in Fig. 3. In the upcoming sections, we present a comprehensive description of the X-Oscar framework: In Sec. 4.1, we delve into the progressive modeling pipeline of X-Oscar. This pipeline breaks down the complex task of avatar generation into three manageable subtasks, with each subtask focusing on a specific aspect of avatar creation. In Sec. 4.2, we introduce Adaptive Variational Parameter (AVP). This component employs a trainable adaptive distribution to represent the avatar, addressing the issue of oversaturation that is commonly encountered in avatar generation. In Sec. 4.3, we present Avatar-aware Score Distillation Sampling (ASDS). This module incorporates geometry-aware and appearance-aware noise into the denoising process, enabling the pretrained diffusion model to perceive the current state of the generated avatar, resulting in the production of high-quality outputs.

4.1. Progressive Modeling

Geomotry Modeling. During this phase, our objective is to optimize the geometry of the avatars, represented by the SMPL-X model, to align with the input text prompt y. Formally, we aim to optimize the trainable vertex offsets $\psi_v \in \mathbb{R}^{N\times 3}$, initialized as a matrix of zeros, to align the modified vertex coordinates $\nu' = \nu + \psi_v$ with the text prompt y, where ν represents the vertex coordinates of the template avatar body, and N is the number of vertices of the SMPL-X model. To achieve this, we utilize a differentiable rendering pipeline. By taking the original mesh $\mathcal{M}$ of SMPL-X and the predicted vertex offsets ψ_v as inputs, we render a normal image $\mathcal{N}$ of the modified mesh using a differentiable renderer (Laine et al., 2020):

$$\mathcal{N} = g(\mathcal{M}, \psi_v, c), \tag{5}$$

where $g(\cdot)$ denotes the rendering function, and c represents a randomly sampled camera parameter. In each iteration, we introduce Gaussian noise ϵ to the normal map $\mathcal{N}$ and apply a pretrained Stable Diffusion (SD) model (Rombach et al., 2022) to denoise it. The gradient of the trainable vertex offsets ψ_v during denoising is then calculated as follows:

$$\nabla_{\psi_v}\mathcal{L}_{\text{geo}}(\psi_v,\mathcal{N}) = \mathbb{E}_{t,\epsilon}\left[w(t)\left(\hat{\epsilon}_\phi(z_t^{\mathcal{N}};y,t)-\epsilon\right)\frac{\partial\mathcal{N}}{\partial\psi_v}\right], \tag{6}$$

where $\hat{\epsilon}_\phi(z_t^{\mathcal{N}};y,t)$ represents the predicted noise by SD based on the timestep t, input text embedding y, and the noisy normal image $z_t^{\mathcal{N}}$.

Appearance Modeling. After completing the geometry modeling phase, we obtain a mesh that aligns with the prompt in terms of shape, with vertex coordinates $\nu' = \nu + \psi_v$. In this stage, our objective is to optimize an albedo map $\psi_a \in \mathbb{R}^{h\times w\times 3}$ to represent the appearance of the resulting avatar, where h and w represent the height and width of the albedo map. To achieve this, we start by rendering a colored image $\mathcal{I}$ from a randomly sampled camera parameter c based on the vertex offsets ψ_v and the albedo map ψ_a using a differentiable renderer (Laine et al., 2020):

$$\mathcal{I} = g(\mathcal{M}, \psi_v, \psi_a, c). \tag{7}$$

To optimize the albedo map ψ_a, we employ a loss function similar to Eq. (6) used in the geometry modeling phase:

$$\nabla_{\psi_a}\mathcal{L}_{\text{app}}(\psi_a,\mathcal{I}) = \mathbb{E}_{t,\epsilon}\left[w(t)\left(\hat{\epsilon}_\phi(z_t^{\mathcal{I}};y,t)-\epsilon\right)\frac{\partial\mathcal{I}}{\partial\psi_a}\right], \tag{8}$$

where $\hat{\epsilon}_\phi(z_t^{\mathcal{I}};y,t)$ represents the predicted noise by the SD model. This loss function encourages the rendered image $\mathcal{I}$ to align with the text prompt y by minimizing the discrepancy between the predicted noise $\hat{\epsilon}_\phi$ and the added Gaussian noise ϵ. By optimizing the albedo map ψ_a using this loss function, we can generate appearances for the avatars that are consistent with the provided text prompts.

Animation Refinement. Given that both the geometry modeling and appearance modeling stages optimize the avatar in a canonical pose, it is inevitable that certain parts of the avatar may be obstructed, leading to lower-quality results in those areas. To overcome this challenge, we introduce an animation refinement stage where we adjust the pose of the avatar and simultaneously optimize both the geometry and appearance. Specifically, we sample viable pose parameters $\mathfrak{p}$ from a pre-trained model such as VPoser (Pavlakos et al., 2019a). For each sampled pose, we render the normal image $\mathcal{N}_p$ and colored image $\mathcal{I}_p$ of the animated avatar using a differentiable renderer (Laine et al., 2020):

$$\mathcal{N}_p = g(\mathcal{M}, \psi_v, c, \mathfrak{p}), \quad \mathcal{I}_p = g(\mathcal{M}, \psi_v, \psi_a, c, \mathfrak{p}), \tag{9}$$

where pose parameters $\mathfrak{p}$ and camera parameters c vary in each iteration. To optimize the geometry and appearance of the avatar in the animated pose, we define an animation loss $\mathcal{L}_{\text{ani}}$ as follows:

$$\mathcal{L}_{\text{ani}}(\psi_v, \psi_a, \mathcal{N}_p, \mathcal{I}_p) = \mathcal{L}_{\text{geo}}(\psi_v, \mathcal{N}_p) + \mathcal{L}_{\text{app}}(\psi_v, \psi_a, \mathcal{I}_p), \tag{10}$$

where $\mathcal{L}_{\text{geo}}$ and $\mathcal{L}_{\text{app}}$ are the geometry loss and appearance loss, respectively. The gradients of the animation loss for the vertex offsets ψ_v and the albedo maps ψ_a are calculated as follows:

$$\begin{aligned} &\nabla_{\psi_v} \mathcal{L}_{\text{ani}}(\psi_v, \mathcal{N}_p, \mathcal{I}_p) \\ =&\mathbb{E}_{t,\epsilon} \left[w(t) \left(\hat{\epsilon}_\phi(z_t^{\mathcal{N}_p}; y, t) - \epsilon \right) \frac{\partial \mathcal{N}_p}{\partial \psi_v} + w(t) \left(\hat{\epsilon}_\phi(z_t^{\mathcal{I}_p}; y, t) - \epsilon \right) \frac{\partial \mathcal{I}_p}{\partial \psi_v} \right], \end{aligned} \tag{11}$$

$$\nabla_{\psi_a} \mathcal{L}_{\text{ani}}(\psi_a, \mathcal{I}_p) = \mathbb{E}_{(t,\epsilon)} \left[w(t) \left(\hat{\epsilon}_\phi(z_t^{\mathcal{I}_p}; y, t) - \epsilon \right) \frac{\partial \mathcal{I}_p}{\partial \psi_a} \right], \tag{12}$$

The notations used here are similar to those defined in Eq. (2). By minimizing the animation loss using these gradients, we refine the geometry and appearance of the avatar in various poses, resulting in improved quality in the final output.

4.2. Adaptive Variational Parameter

As formulated in Eq. (1) and Eq. (2), SDS aims to optimize a precise 3D representation to align all images rendered from arbitrary viewpoints with the input prompt evaluated by 2D diffusion models. However, there exists a fundamental contradiction between achieving an accurate 3D representation and the inherent multi-view inconsistency associated with 2D diffusion models. Specifically, it is often unreasonable to expect high similarity scores of a 2D diffusion model between all multi-view images of a specific 3D representation and text prompts. Consequently, when SDS is employed to enforce similarity between each perspective of a specific 3D representation and the text prompt, it can lead to the undesirable issue of oversaturation. To address this concern, we propose formulating the 3D representation as a distribution of vertex offsets, denoted as *offset distribution*, and a distribution of albedo maps, referred to as *appearance distribution*. Specifically, we perturb ψ_v and ψ_a of the 3D human representation with Gaussian noises to improve the robustness of the model and alleviate the oversaturation problem. This perturbation process can be expressed as:

$$\psi'_v \sim \psi_v + \lambda_v \mathcal{N}(0, I), \quad \psi'_a \sim \psi_a + \lambda_a \mathcal{N}(0, I), \tag{13}$$

where λ_v and λ_a serve as weights to control the magnitude of the perturbations. The mean of the offset distribution and appearance distribution can be learned by optimizing ψ_v and ψ_a, while their standard deviations are determined by λ_v and λ_a. Thus, choosing appropriate values for λ_v and λ_a is crucial and challenging. If these values are too small, the model may not fully benefit from learning the distributions. In extreme cases, when $\lambda_v = \lambda_a = 0$, the model essentially learns specific parameters instead of distributions. Conversely, when λ_v and λ_a are excessively large, the learning process becomes challenging due to highly unstable perturbations. In extreme cases, when $\lambda_v = \lambda_a = +\infty$, the generated results become independent of the underlying ψ_v and ψ_a.

To overcome the above challenges and facilitate a learning process that progresses from easy to difficult without manual weight assignment, we propose Adaptive Variational Parameter (AVP) for 3D representation. Specifically, we leverage the standard deviations of ψ_v and ψ_a as weights for perturbations, which can be formulated as follows:

$$\psi'_v \sim \psi_v + \sigma(\psi_v)\mathcal{N}(0, I) = \mathcal{N}\left(\psi_v, \sigma(\psi_v)^2\right), \tag{14}$$

$$\psi'_a \sim \psi_a + \sigma(\psi_a)\mathcal{N}(0, I) = \mathcal{N}\left(\psi_a, \sigma(\psi_a)^2\right), \tag{15}$$

where $\sigma(\cdot)$ represents the standard deviation. This adaptive approach has several advantages. *Firstly, it enables the model to learn progressively from easy to difficult scenarios.* Initially, ψ_v and ψ_a are initialized as matrices of all zeros and all 0.5, respectively, resulting in a standard deviation of 0. Consequently, during the early stages of training, the model focuses on optimizing the means of ψ'_v and ψ'_a to reasonable values. As training progresses, the standard deviations gradually increase, promoting the model's ability to maintain high similarity between the 3D representation and the text even in the presence of noise interference. *Secondly, this approach is fully automatic.* The model learns to adapt the perturbation weights based on the current state of the 3D representation, eliminating the need for manual intervention or hyperparameter tuning. During the inference phase, we utilize the mean values of ψ'_v and ψ'_a to represent the avatar.

4.3. Avatar-aware Score Distillation Sampling

In previous work on SDS (Poole et al., 2022), a Gaussian noise related to timestep t was introduced to the rendered image, and a pretrained diffusion model was utilized to denoise the noisy image for optimizing the 3D representation. The process of adding noise can be formulated as follows:

$$\begin{aligned} z_t =& \sqrt{\alpha_t} z_{t-1} + \sqrt{1 - \alpha_t} \epsilon_{t-1} \\ =& \sqrt{\alpha_t \alpha_{t-1}} z_{t-2} + \sqrt{1 - \alpha_t \alpha_{t-1}} \bar{\epsilon}_{t-2} \\ =& \cdots \\ =& \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \bar{\epsilon}_0, \end{aligned} \tag{16}$$

where z_t represents the noised image at timestep t, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, and $\epsilon_i, \bar{\epsilon}_i \sim \mathcal{N}(0, I)$. Since $t \sim \mathcal{U}(0.02, 0.98)$ is

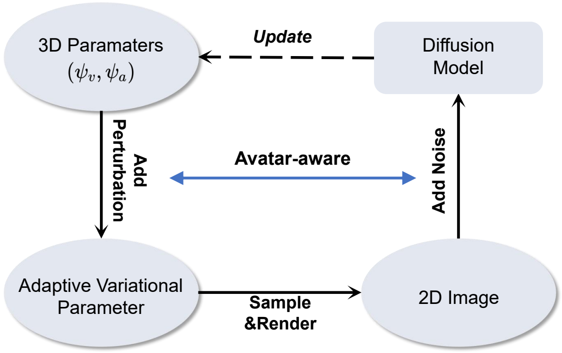

Figure 3: The workflow of the proposed X-Oscar. First, we incorporate the adaptive perturbation into the 3D parameters, forming the avatar distribution. Next, we sample a set of parameters from the avatar distribution and render a 2D image. Finally, we apply avatar-aware noise to the rendered image for denoising to optimize 3D parameters.

randomly sampled, the noise added to the rendered image is independent of the avatar's current state. To establish a correlation between the denoising process and the avatar's current state, and to facilitate a learning process from easy to difficult, we propose Avatar-aware Score Distillation Sampling (ASDS). Specifically, the noised image with avatar-aware noise can be formulated as follows:

$$\begin{aligned} z_t &= \sqrt{\bar{\alpha}} z_0 + \sqrt{1-\bar{\alpha}_t}(\lambda_n \epsilon_n + \lambda_v \sigma(\psi_v)\epsilon_v + \lambda_a \sigma(\psi_a)\epsilon_a) \\ &= \sqrt{\bar{\alpha}} z_0 + \sqrt{1-\bar{\alpha}_t}\sqrt{(\lambda_n)^2 + (\lambda_v \sigma(\psi_v))^2 + (\lambda_a \sigma(\psi_a))^2}\epsilon \\ &= \sqrt{\bar{\alpha}} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_\theta, \end{aligned} \tag{17}$$

where ϵ_n, ϵ_v, ϵ_a, and ϵ are i.i.d. Gaussian random variables with zero mean and unit variance, i.e., $\epsilon_n, \epsilon_v, \epsilon_a, \epsilon \sim \mathcal{N}(0, I)$, and $\epsilon_\theta \sim \mathcal{N}(0, (\lambda_n)^2 + (\lambda_v \sigma(\psi_v))^2 + (\lambda_a \sigma(\psi_a))^2)$. At the initial stage, when $\sigma(\psi_v) = \sigma(\psi_a) = 0$, the initial variance of the noise is relatively small, resulting in an easier denoising process for diffusion models. As the training progresses, $\sigma(\psi_v)$ and $\sigma(\psi_a)$ gradually increase, leading to an increase in the noise variance. Consequently, this increases the difficulty of denoising. By incorporating avatar-aware noise, the model can undergo a learning process from easy to difficult. The gradient of ASDS is then formulated as follows:

$$\begin{aligned} &\nabla_\theta \mathcal{L}_{\text{ASDS}}(\theta) \triangleq \\ &\mathbb{E}_{(t,\epsilon,c)}\left[\omega(t)\big(\underbrace{\hat{\epsilon}_\phi(z_t; y, t)}_{\text{precited noise}} - \underbrace{\epsilon_\theta}_{\text{avatar-aware noise}}\big)\frac{\partial g(\theta, c)}{\partial \theta}\right], \end{aligned} \tag{18}$$

where $z_t = \sqrt{\bar{\alpha}} g(\theta, c) + \sqrt{1-\bar{\alpha}}\epsilon_\theta$ represents the noised image, and ϵ_θ is an avatar-aware noise that encourages the paradigm of learning from easy to difficult.

5. Experiments

5.1. Implementation Details

Our experiments are conducted using a single Nvidia RTX 3090 GPU with 24GB of memory and the PyTorch library (Paszke et al., 2019). The diffusion model employed in our implementation is the Stable Diffusion provided by HuggingFace Diffusers (von Platen et al., 2022). During the training phase, we set the resolution of the rendered images to 800 × 800 pixels. The resolution of the albedo map is 2048 × 2048 pixels. The geometry modeling, appearance modeling, and animation refinement stages consist of 5000, 10000, and 5000 iterations, respectively. We set the learning rates for the vertex offset ψ_v and albedo map ψ_a to 1e-4 and 5e-3, respectively. Furthermore, we set the values of λ_n, λ_v, and λ_a to 0.8, 0.1, and 0.1, respectively. To enhance facial details, we employ a strategy where there is a 0.2 probability of rendering facial images for optimization during the training process, and a 0.8 probability of rendering full-body images for optimization.

5.2. Comparison

Qualitative Comparison with Text-to-Avatar Methods. We present a comparative analysis of our methodology against five state-of-the-art (SOTA) baselines: TADA (Liao et al., 2023a), DreamWaltz (Huang et al., 2023b), HumanGaussian (Liu et al., 2023), AvatarCLIP (Hong et al., 2022), and AvatarCraft (Jiang et al., 2023), as illustrated in Fig. 4. We observe certain limitations in the geometry and texture of avatars generated by TADA, which we emphasize by enclosing them within a red box. Furthermore, the outcomes produced by the other baselines exhibit issues such as blurriness and inconsistencies with the provided text. In contrast, our proposed X-Oscar consistently generates high-quality avatars with intricate details. Moreover, in addition to static avatars, X-Oscar is also capable of generating animatable avatars, as demonstrated in Fig. 1.

Qualitative Comparison with Text-to-3D Methods. We also conduct a comparative analysis of X-Oscar with SOTA text-to-3D methods, namely DreamFusion (Poole et al., 2022), Magic3D (Lin et al., 2023), Fantasia3D (Chen et al., 2023a), and ProlificDreamer (Wang et al., 2023b). As shown in Fig. 5, we observe evident limitations in the avatars generated by text-to-3D methods, including poor geometry and noisy texture. Furthermore, owing to the absence of human prior knowledge, the avatars generated by text-to-3D methods lack flexibility and pose challenges in terms of animation. In contrast, our proposed method excels in generating high-quality, animatable avatars.

Quantitative Comparison. To assess X-Oscar quantitatively, we conduct user studies comparing its performance with SOTA text-to-3D content and text-to-avatar methods

Table 1: Quantitative comparison of SOTA Methods: The top-performing and second-best results are highlighted in **bolded** and underlined, respectively. As AvatarCLIP employs the CLIP score as its training supervision signal, it is inappropriate to gauge its performance using the CLIP score. Therefore, we set the CLIP score of AvatarCLIP to gray.

Method	User Study			CLIP Score			OpenCLIP Score		
	Geo. Qua.	Tex. Qua.	Tex. Con.	ViT-B/32	ViT-B/16	ViT-L/14	ViT-B/32	ViT-B/16	ViT-L/14
DreamFusion	2.66	4.18	3.29	29.29	29.29	25.30	31.57	28.22	30.17
Magic3D	4.21	3.12	1.61	28.52	30.92	27.02	31.14	28.21	30.21
Fantasia3D	2.14	2.42	2.53	30.34	30.42	26.12	29.68	28.46	**31.46**
ProlificDreamer	2.11	3.72	6.29	30.30	30.28	25.00	30.81	28.59	30.75
AvatarCLIP	3.28	2.64	2.09	34.49	32.45	28.20	32.77	31.20	31.98
AvatarCraft	4.39	4.55	3.37	27.59	29.70	25.23	26.19	24.60	25.55
DreamWaltz	6.38	6.09	6.99	30.86	31.20	27.32	30.65	29.09	29.83
HumanGuassian	6.03	4.51	6.08	28.46	29.18	26.26	26.37	26.82	29.09
TADA	5.03	6.95	7.62	31.09	30.48	27.72	30.67	30.05	30.17
X-Oscar	**8.85**	**8.91**	**9.22**	**31.70**	**31.97**	**28.10**	**30.91**	**30.28**	30.42

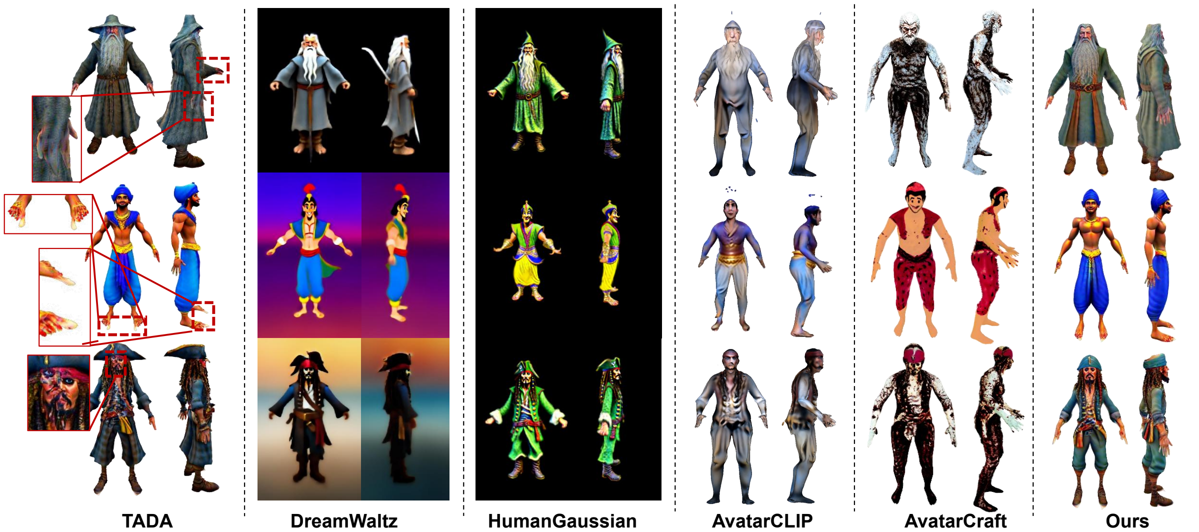

Figure 4: Qualitative comparisons with SOTA text-to-avatar methods. The prompts (top → down) are "Gandalf from The Lord of the Rings", "Aladdin in Aladdin", and "Captain Jack Sparrow from Pirates of the Caribbean".

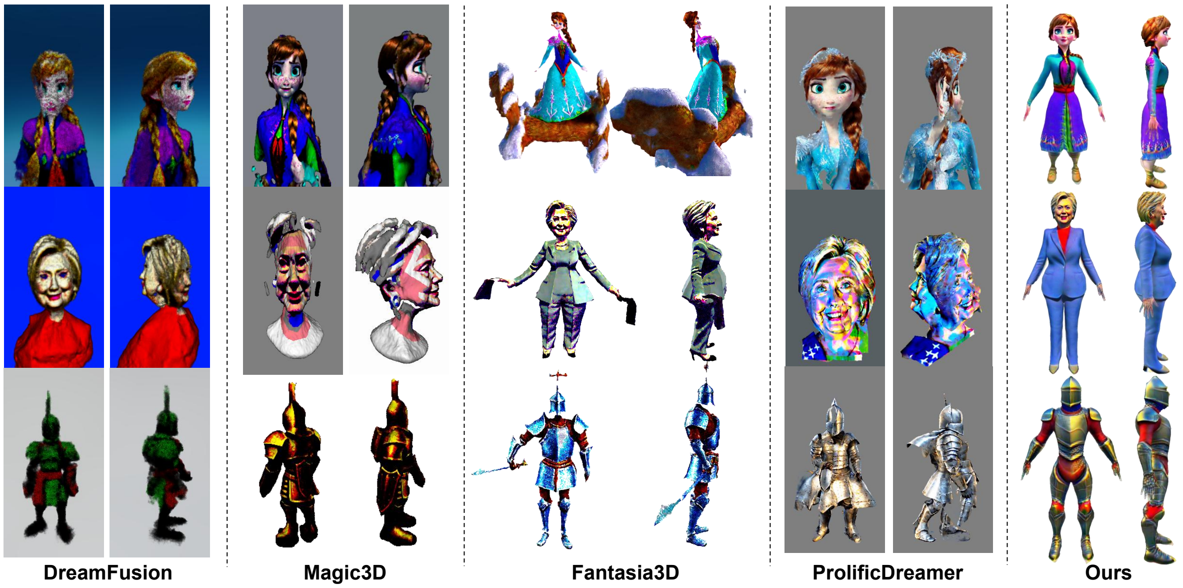

Figure 5: Qualitative comparisons with SOTA text-to-3D methods. The prompts (top → down) are "Anna in Frozen", "Hilary Clinton", and "Knight".

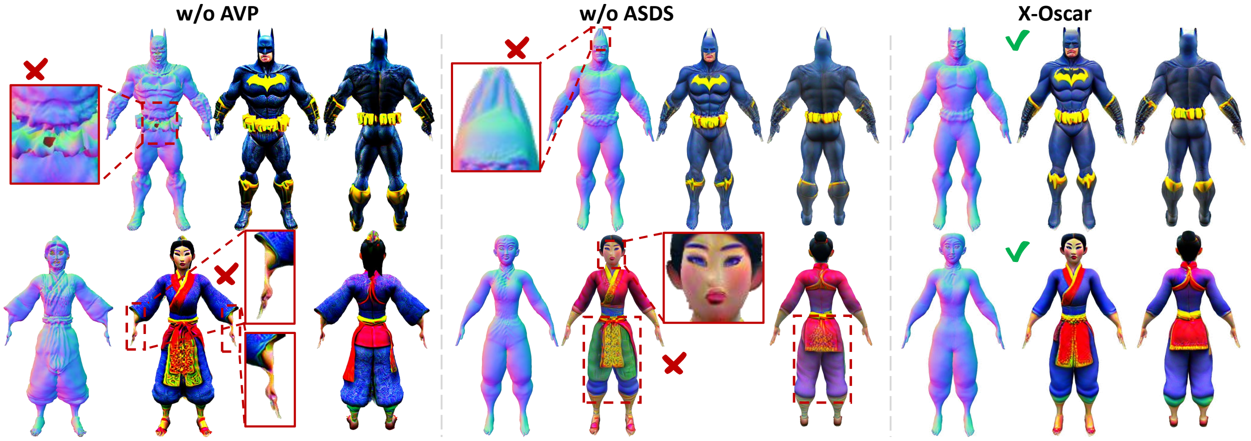

Figure 6: Ablation study on the Adaptive Variational Parameter and Avatar-aware Score Distillation Sampling. The prompts (top → down) are "Batman", and "Mulan".

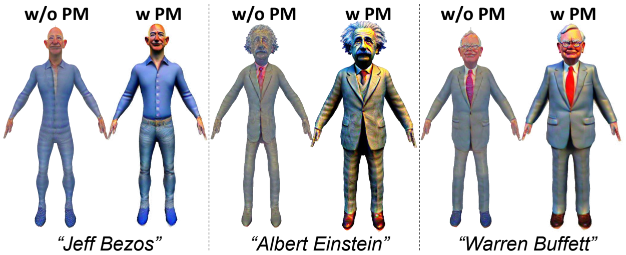

Figure 7: Ablation study on progressive modeling. "PM" is short for "progressive modeling". "w/o PM" means that geometry, appearance, and animation are optimized together.

using the same prompts. We randomly selected 40 prompts generated by ChatGPT for avatar creation, and the user studies involved 52 participants who provided subjective evaluations. Participants rated the generated avatars based on three specific aspects: texture quality (Geo. Qua.), geometry quality (Tex. Qua.), and text consistency (Tex. Con.). Scores range from 1 to 10, with higher scores indicating better quality. As shown in Tab. 1, our method consistently outperforms all other methods across all evaluated aspects. Additionally, we calculate similarity scores between the generated results and text prompts using CLIP (Radford et al., 2021) and OpenCLIP (Cherti et al., 2023) with different backbones. Our method consistently achieves either the best or second-best results, demonstrating its ability to generate 3D avatars that are semantically consistent with the provided text prompts.

5.3. Ablation Studies

Progressive Modeling. To evaluate the effectiveness of the progressive modeling paradigm in X-Oscar, we performed additional experiments by coupling the three training stages together. The results shown in Fig. 7 reveal a significant enhancement in the quality of geometry and appearance in the generated avatars when using the progressive modeling paradigm. For example, consider the prompt "Albert Einstein". Without employing the progressive modeling approach, the generated avatar is limited to a rudimentary shape and color, lacking the intricate details necessary for recognizing Albert Einstein. However, when employing the progressive modeling paradigm, we observe a remarkable improvement in the generated avatars.

Adaptive Variational Parameter. To provide robust evidence of the impact of AVP, we conducted comprehensive ablation studies by using specific parameters instead of distributions to represent avatars. As depicted in Fig. 6, our observations strongly indicate that the omission of AVP in X-Oscar can lead to an excessive optimization of geometry and appearance, as an effort to align the generated outputs with the text. This subsequently leads to the problem of oversaturation. Geometry oversaturation leads to topological overlay problems in the generated meshes, while appearance oversaturation results in avatars with exaggerated color contrast. By integrating AVP, we successfully tackle these issues, significantly improving the realism of both the geometry and appearance in the generated avatars.

Avatar-aware Score Distillation Sampling. To investigate the impact of ASDS, we conducted additional experiments by adding random Gaussian noise instead of avatar-aware noise to the rendered image for optimization. As demonstrated in Fig. 6, the absence of ASDS directly results in a noticeable decline in the overall quality of both the geometry and appearance of the generated avatars. For instance, without ASDS, two ears on Batman's head exhibit a geometric merging phenomenon. In the case of Mulan, the facial details become blurred and the colors on the front and back of the pants are inconsistent.

Table 2: Training time of different text-to-avatar methods. The performance is tested on a Nvidia GeForce RTX 3090 GPU.

Method	3D Representation	Guidance Model	Training Time
HumanGuassian (Liu et al., 2023)	3D Gaussian Splatting	Stable Diffusion	69 minutes
DreamWaltz (Huang et al., 2023b)	NeRF	Stable Diffusion	105 minutes
AvatarCLIP (Hong et al., 2022)	Neus	CLIP	156 minutes
X-Oscar	Mesh	Stable Diffusion	172 minutes
TADA (Liao et al., 2023a)	Mesh	Stable Diffusion	175 minutes
AvatarCraft (Jiang et al., 2023)	Mesh	Stable Diffusion	298 minutes

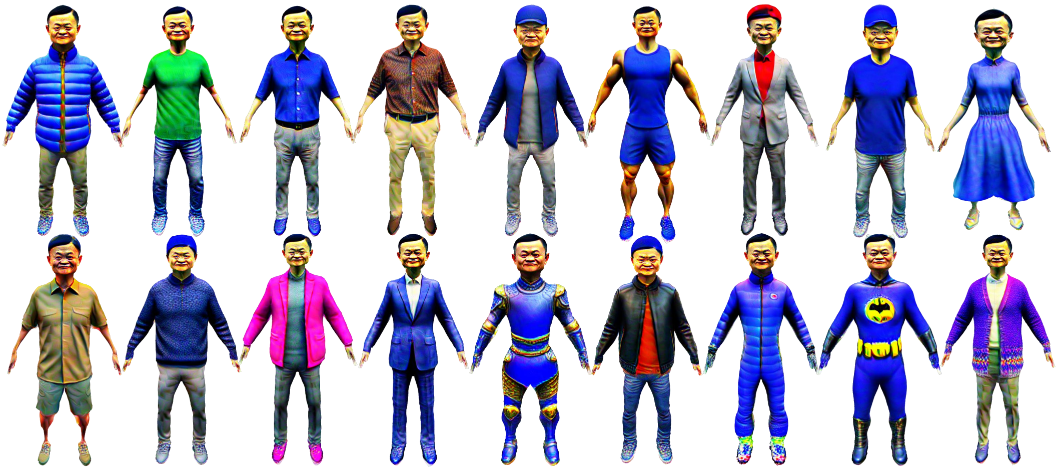

Figure 8: Virtual Try-on results of X-Oscar on "Jack Ma".

6. Training Time

The training times of various text-to-avatar methods are reported in Tab. 2. A meticulous analysis of the table reveals several significant findings that not only shed light on the efficiency and effectiveness of different approaches but also strengthen the overall argument. Notably, the HumanGaussian method exhibits the shortest training time among the considered approaches. This can be attributed to the inherent properties of 3D Gaussian Splitting, which enables faster training due to its efficient computational nature. Another method, DreamWaltz, demonstrates the second shortest training time by employing NeRF as the 3D representation. This combination of techniques allows for relatively fast training, showcasing the effectiveness of NeRF in accelerating the avatar generation process. When utilizing Neus as the 3D representation, AvatarCLIP showcases relatively short training times. However, it is crucial to acknowledge that the expedited training comes with a trade-off in terms of output quality. Avatars generated using CLIP as the guidance model may not consistently meet the desired standards of quality. In the case of Mesh representation combined with Stable Diffusion as the guiding model, our proposed method demonstrates the shortest training time. This can be attributed to our proposed progressive framework, which optimizes only a portion of the parameters during the geometric and appearance modeling stages. By leveraging this progressive framework, we effectively reduce the overall training time while maintaining the ability to generate high-quality avatars.

7. Virtual Try-On

In this section, we have demonstrated that the proposed X-Oscar can be effectively utilized for virtual try-on. As shown in Fig. 8, it successfully employs fashion outfit prompts to generate immersive virtual experiences. Our experimental results confirm that X-Oscar accurately applies the provided fashion styles to virtual scenarios, achieving realistic try-on effects. This functionality offers users a convenient way to preview and select clothing styles that suit them through virtual try-on, thereby enhancing the overall shopping experience.

8. Conclusion

This paper introduces X-Oscar, an advanced framework for generating high-quality, text-guided 3D animatable avatars. The framework incorporates three innovative designs to enhance avatar generation. Firstly, we present a progressive modeling paradigm with clear and simple optimization objectives for each training stage. Additionally, we propose Adaptive Variational Parameter (AVP), which optimizes the distribution of avatars, addressing oversaturation. Furthermore, we introduce Avatar-aware Score Distillation Sampling (ASDS), leveraging avatar-aware denoising to enhance overall avatar quality. Extensive experiments demonstrate the effectiveness of the proposed framework and modules.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgments

This work was supported by National Key R&D Program of China (No.2023YFB4502804), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. 62072389), the National Natural Science Fund for Young Scholars of China (No. 62302411), China Postdoctoral Science Foundation (No. 2023M732948), the Natural Science Foundation of Fujian Province of China (No.2021J01002, No.2022J06001), and partially sponsored by CCF-NetEase ThunderFire Innovation Research Funding (NO. CCF-Netease 202301).

References

Alldieck, T., Xu, H., and Sminchisescu, C. imghum: Implicit generative models of 3d human shape and articulated pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5461–5470, 2021.

Cao, Y., Cao, Y.-P., Han, K., Shan, Y., and Wong, K.-Y. K. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. *arXiv preprint arXiv:2304.00916*, 2023.

Chen, R., Chen, Y., Jiao, N., and Jia, K. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023a.

Chen, Y., Pan, Y., Li, Y., Yao, T., and Mei, T. Control3d: Towards controllable text-to-3d generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1148–1156, 2023b.

Chen, Z., Wang, F., and Liu, H. Text-to-3d using gaussian splatting. *arXiv preprint arXiv:2309.16585*, 2023c.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Güler, R. A., Neverova, N., and Kokkinos, I. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7297–7306, 2018.

Han, S.-H., Park, M.-G., Yoon, J. H., Kang, J.-M., Park, Y.-J., and Jeon, H.-G. High-fidelity 3d human digitization from single 2k resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12869–12879, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hong, F., Zhang, M., Pan, L., Cai, Z., Yang, L., and Liu, Z. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022. doi: 10.1145/3528223.3530094.

Huang, X., Shao, R., Zhang, Q., Zhang, H., Feng, Y., Liu, Y., and Wang, Q. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. *arXiv preprint arXiv:2310.01406*, 2023a.

Huang, Y., Wang, J., Zeng, A., Cao, H., Qi, X., Shi, Y., Zha, Z.-J., and Zhang, L. Dreamwaltz: Make a scene with complex 3d animatable avatars. *Advances in Neural Information Processing Systems*, 2023b.

Jain, A., Mildenhall, B., Barron, J. T., Abbeel, P., and Poole, B. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 867–876, 2022.

Jiang, R., Wang, C., Zhang, J., Chai, M., He, M., Chen, D., and Liao, J. Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control. 2023.

Jiang, W., Yi, K. M., Samei, G., Tuzel, O., and Ranjan, A. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision*, pp. 402–418. Springer, 2022.

Kerbl, B., Kopanas, G., Leimkühler, T., and Drettakis, G. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Kolotouros, N., Alldieck, T., Zanfir, A., Bazavan, E., Fieraru, M., and Sminchisescu, C. Dreamhuman: Animatable 3d avatars from text. *Advances in Neural Information Processing Systems*, 36, 2024.

Kulikov, V., Yadin, S., Kleiner, M., and Michaeli, T. Sinddm: A single image denoising diffusion model. In *International Conference on Machine Learning (ICML)*, pp. 17920–17930. PMLR, 2023.

Kullback, S. and Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86, 1951.

Laine, S., Hellsten, J., Karras, T., Seol, Y., Lehtinen, J., and Aila, T. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning (ICML)*, 2022a.

Li, J., Li, D., Savarese, S., and Hoi, S. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning (ICML)*, 2023a.

Li, R., Olszewski, K., Xiu, Y., Saito, S., Huang, Z., and Li, H. Volumetric human teleportation. In *ACM SIGGRAPH 2020 Real-Time Live!*, pp. 1–1. 2020a.

Li, R., Xiu, Y., Saito, S., Huang, Z., Olszewski, K., and Li, H. Monocular real-time volumetric performance capture. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pp. 49–67. Springer, 2020b.

Li, X., Wang, H., and Tseng, K.-K. Gaussiandiffusion: 3d gaussian splatting for denoising diffusion probabilistic models with structured noise. *arXiv preprint arXiv:2311.11221*, 2023b.

Li, Y., Lao, L., Cui, Z., Shan, S., and Yang, J. Graph jigsaw learning for cartoon face recognition. *IEEE Transactions on Image Processing*, 31:3961–3972, 2022b.

Li, Z., Chen, Y., Zhao, L., and Liu, P. Mvcontrol: Adding conditional control to multi-view diffusion for controllable text-to-3d generation. *arXiv preprint arXiv:2311.14494*, 2023c.

Liao, T., Yi, H., Xiu, Y., Tang, J., Huang, Y., Thies, J., and Black, M. J. Tada! text to animatable digital avatars. 2023a.

Liao, T., Zhang, X., Xiu, Y., Yi, H., Liu, X., Qi, G.-J., Zhang, Y., Wang, X., Zhu, X., and Lei, Z. High-fidelity clothed avatar reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8662–8672, 2023b.

Lin, C.-H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.-Y., and Lin, T.-Y. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 300–309, 2023.

Liu, X., Zhan, X., Tang, J., Shan, Y., Zeng, G., Lin, D., Liu, X., and Liu, Z. Humangaussian: Text-driven 3d human generation with gaussian splatting. *arXiv preprint arXiv:2311.17061*, 2023.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J. Smpl: A skinned multi-person linear model. 34(6), oct 2015. ISSN 0730-0301. doi: 10.1145/2816795.2818013. URL https://doi.org/10.1145/2816795.2818013.

Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., and Ji, R. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 638–647, 2022.

Ma, Y., Fan, Y., Ji, J., Wang, H., Sun, X., Jiang, G., Shu, A., and Ji, R. X-dreamer: Creating high-quality 3d content by bridging the domain gap between text-to-2d and text-to-3d generation. *arXiv preprint arXiv:2312.00085*, 2023a.

Ma, Y., Ji, J., Sun, X., Zhou, Y., and Ji, R. Towards local visual modeling for image captioning. *Pattern Recognition*, 138:109420, 2023b.

Ma, Y., Zhang, X., Sun, X., Ji, J., Wang, H., Jiang, G., Zhuang, W., and Ji, R. X-mesh: Towards fast and accurate text-driven 3d stylization via dynamic textual guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2749–2760, 2023c.

Men, Y., Lei, B., Yao, Y., Cui, M., Lian, Z., and Xie, X. En3d: An enhanced generative model for sculpting 3d humans from 2d synthetic data. *arXiv preprint arXiv:2401.01173*, 2024.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Mohammad Khalid, N., Xie, T., Belilovsky, E., and Popa, T. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 conference papers*, pp. 1–8, 2022.

Park, J. J., Florence, P., Straub, J., Newcombe, R., and Lovegrove, S. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 165–174, 2019.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A., Tzionas, D., and Black, M. J. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10975–10985, 2019a.

Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A. A., Tzionas, D., and Black, M. J. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019b.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. Dreamfusion: Text-to-3d using 2d diffusion. *International Conference on Learning Representation (ICLR)*, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

Sanghi, A., Chu, H., Lambourne, J. G., Wang, Y., Cheng, C.-Y., Fumero, M., and Malekshan, K. R. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18603–18613, 2022.

Santesteban, I., Thuerey, N., Otaduy, M. A., and Casas, D. Self-supervised collision handling via generative 3d garment models for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11763–11773, 2021.

Santesteban, I., Otaduy, M., Thuerey, N., and Casas, D. Ulnef: Untangled layered neural fields for mix-and-match virtual try-on. *Advances in Neural Information Processing Systems*, 35:12110–12125, 2022.

Shen, T., Gao, J., Yin, K., Liu, M.-Y., and Fidler, S. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34:6087–6101, 2021.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning (ICML)*, pp. 2256–2265. PMLR, 2015.

Tang, J., Ren, J., Zhou, H., Liu, Z., and Zeng, G. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023.

von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., and Wolf, T. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Wang, H., Du, X., Li, J., Yeh, R. A., and Shakhnarovich, G. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12619–12629, 2023a.

Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., and Wang, W. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., and Zhu, J. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. 2023b.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML)*, pp. 681–688, 2011.

Weng, C.-Y., Curless, B., Srinivasan, P. P., Barron, J. T., and Kemelmacher-Shlizerman, I. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pp. 16210–16220, 2022.

Xiu, Y., Yang, J., Cao, X., Tzionas, D., and Black, M. J. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 512–523, 2023.

Xu, H., Ye, Q., Yan, M., Shi, Y., Ye, J., Xu, Y., Li, C., Bi, B., Qian, Q., Wang, W., Xu, G., Zhang, J., Huang, S., Huang, F., and Zhou, J. mplug-2: A modularized multi-modal foundation model across text, image and video. In *International conference on machine learning (ICML)*, 2023a.

Xu, Y., Yang, Z., and Yang, Y. Seeavatar: Photorealistic text-to-3d avatar generation with constrained geometry and appearance. *arXiv preprint arXiv:2312.08889*, 2023b.

Yi, T., Fang, J., Wu, G., Xie, L., Zhang, X., Liu, W., Tian, Q., and Wang, X. Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. *arXiv preprint arXiv:2310.08529*, 2023.

Zeng, Y., Lu, Y., Ji, X., Yao, Y., Zhu, H., and Cao, X. Avatarbooth: High-quality and customizable 3d human avatar generation. *arXiv preprint arXiv:2306.09864*, 2023.

Zhang, H., Chen, B., Yang, H., Qu, L., Wang, X., Chen, L., Long, C., Zhu, F., Du, K., and Zheng, M. Avatarverse: High-quality & stable 3d avatar creation from text and pose. *arXiv preprint arXiv:2308.03610*, 2023a.

Zhang, H., Chen, B., Yang, H., Qu, L., Wang, X., Chen, L., Long, C., Zhu, F., Du, D., and Zheng, M. Avatarverse: High-quality & stable 3d avatar creation from text and pose. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7124–7132, 2024.

Zhang, J., Zhang, X., Zhang, H., Liew, J. H., Zhang, C., Yang, Y., and Feng, J. Avatarstudio: High-fidelity and animatable 3d avatar creation from text. *arXiv preprint arXiv:2311.17917*, 2023b.

Zhang, L., Wong, T.-T., and Liu, Y. Sprite-from-sprite: Cartoon animation decomposition with self-supervised sprite estimation. *ACM Transactions on Graphics (TOG)*, 41(6):1–12, 2022.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023c.

Zhang, Z., Yang, Z., and Yang, Y. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. *arXiv preprint arXiv:2312.06704*, 2023d.

Zheng, Y., Shao, R., Zhang, Y., Yu, T., Zheng, Z., Dai, Q., and Liu, Y. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6239–6249, 2021.

Zhou, Z., Ma, F., Fan, H., and Yang, Y. Headstudio: Text to animatable head avatars with 3d gaussian splatting. *arXiv preprint arXiv:2402.06149*, 2024.

Zhu, L., Rematas, K., Curless, B., Seitz, S. M., and Kemelmacher-Shlizerman, I. Reconstructing nba players. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 177–194. Springer, 2020.