HeadStudio: Text to Animatable Head Avatars with 3D Gaussian Splatting

ZHENGLIN ZHOU, ReLER Lab, Zhejiang University, China
FAN MA, ReLER Lab, Zhejiang University, China
HEHE FAN, ReLER Lab, Zhejiang University, China
YI YANG ✉, ReLER Lab, Zhejiang University, China

Fig. 1. Text-based animatable avatars generation by **HeadStudio**. With only one end-to-end training stage of 2 hours on 1 NVIDIA A6000 GPU, HeadStudio is able to generate animatable, high-fidelity and real-time rendering (≥ 40 fps) head avatars using text inputs.

Creating digital avatars from textual prompts has long been a desirable yet challenging task. Despite the promising outcomes obtained through 2D diffusion priors in recent works, current methods face challenges in achieving high-quality and animated avatars effectively. In this paper, we present **HeadStudio**, a novel framework that utilizes 3D Gaussian splatting to generate realistic and animated avatars from text prompts. Our method drives 3D Gaussians semantically to create a flexible and achievable appearance through the intermediate FLAME representation. Specifically, we incorporate the FLAME into both 3D representation and score distillation: 1) FLAME-based 3D Gaussian splatting, driving 3D Gaussian points by rigging each point to a FLAME mesh. 2) FLAME-based score distillation sampling, utilizing FLAME-based fine-grained control signal to guide score distillation from the text prompt. Extensive experiments demonstrate the efficacy of HeadStudio in generating animatable avatars from textual prompts, exhibiting visually appealing appearances. The avatars are capable of rendering high-quality real-time (≥ 40 fps) novel views at a resolution of 1024. They can be smoothly controlled by real-world speech and video. We hope that HeadStudio can advance digital avatar creation and that the present method can widely be applied across various domains. The code will be publicly available.

1 INTRODUCTION

Digital head avatar is a virtual representation of a person or character, which has a wide range of applications, such as online conferences, game character creation, and virtual social presence. Head avatar generation has improved significantly in recent years with the development of deep learning. At first, the image-based methods [Chan et al. 2022; Zielonka et al. 2023] are proposed to reconstruct the photo-realistic head avatar of a person, given one or more views. Recently, generative models (*e.g.* diffusion [Rombach et al. 2022; Zhang et al. 2023b]) have made unprecedented advancements in high-quality text-to-image synthesis. As a result, the research focus has been on text-based head avatar generation methods, which have shown superiority over image-based methods in convenience and generalization. These methods [Han et al. 2023; Poole et al. 2022] create avatars by distilling knowledge from a diffusion model into a learnable 3D representation, ensuring that any view of the representation satisfies the provided textual description.

However, current text-based methods cannot combine high-fidelity and animation effectively. For example, HeadSculpt [Han et al. 2023] leverages DMTET [Shen et al. 2021] for high-resolution optimization and excels in creating highly detailed head avatars but is unable to animate them. TADA [Liao et al. 2023] employs SMPL-X [Pavlakos et al. 2019] to generate animatable digital characters but sacrifices appearance quality. There is always a trade-off between static quality and dynamic animation. Producing high-resolution animated head avatars still presents a challenge for current methods.

In this paper, we propose a novel text-based generation framework, named **HeadStudio**, by fully exploiting 3D Gaussian splatting (3DGS) [Kerbl et al. 2023], which achieves superior rendering

✉: corresponding author.
Mails: zhenglinzhou@zju.edu.cn; flowerfan524@gmail.com; hehefan@zju.edu.cn; yangyics@zju.edu.cn.

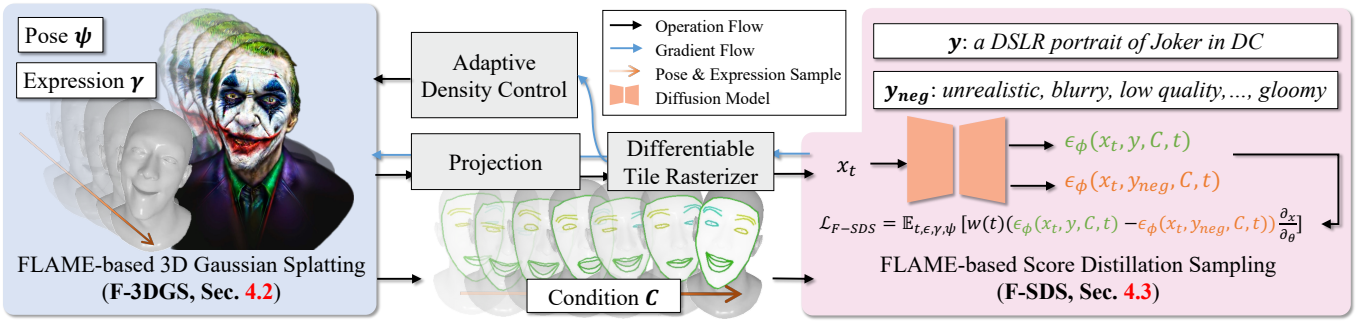

Fig. 2. Framework of HeadStudio, which integrates FLAME into 3D Gaussian splatting and score distillation sampling. 1) **FLAME-based 3D Gaussian Splatting (F-3DGS)**: each 3D point is rigged to a FLAME mesh, and then rotated, scaled, and translated by the mesh deformation. 2) **FLAME-based Score Distillation Sampling (F-SDS)**: utilizing FLAME-based fine-grained control signals to guide score distillation. Furthermore, we also introduce additional enhancements, including uniform super-resolution and mesh regularization in F-3DGS, training with animation and denoised score distillation in F-SDS.

quality and real-time performance for novel-view synthesis. Applying 3DGS directly to generate dynamic avatars from text still presents a complex challenge. The difficulty lies in two aspects: 1) deform 3D Gaussian points with facial expression control; 2) distill knowledge with facial expression guidance. To address these issues, we incorporate FLAME [Li et al. 2017], a statistical head model, as an intermediate representation. The dynamic head generation is thus accomplished by aligning the 3D Gaussian points with the FLAME representation.

To achieve this, we introduce the FLAME-based 3D Gaussian splatting (**F-3DGS**), which deforms 3D Gaussian points by rigging each 3D Gaussian point to a FLAME mesh. Additionally, we present FLAME-based score distillation score (**F-SDS**), which utilizes MediaPipe [Lugaresi et al. 2019] facial landmark map, a FLAME-based fine-grained control signal, to guide score distillation. In addition, FLAME-based regularizations are designed for both 3D representation and score distillation, such as uniform super-resolution, mesh regularization, and training with animations, to create animatable and high-fidelity head avatars.

Extensive experiments have shown that HeadStudio is highly effective and superior to state-of-the-art methods in generating dynamic avatars from text. [Han et al. 2023; Liao et al. 2023; Metzer et al. 2022; Poole et al. 2022; Wang et al. 2023; Zhang et al. 2023a]. Moreover, our methods can be easily extended to driving generated 3D avatars via both speech-based [Yi et al. 2023b] and video-based [Feng et al. 2021] methods. Overall, our contributions can be summarized as follows.

- To the best of our knowledge, we make the first attempt to incorporate 3D Gaussian splatting into the text-based dynamic head avatar generation.
- We propose HeadStudio, which employs FLAME to enhance 3D representation and score distillation for creating fidelity and animatable head avatars.
- HeadStudio is simple, efficient and effective. With only one end-to-end training stage of 2 hours on 1 NVIDIA A6000 GPU, HeadStudio is able to generate 40 fps high-fidelity head avatars.

2 RELATED WORK

Text-to-2D generation. Recently, with the development of vision-language models [Radford et al. 2021] and diffusion models [Ho et al. 2020; Sohl-Dickstein et al. 2015], great advancements have been made in text-to-image generation (T2I) [Zhang et al. 2023c]. In particular, GLIDE [Nichol et al. 2021] introduces classifier-free guidance in T2I, facilitating the utilization of free-form prompts. Additionally, Imagen [Ho et al. 2022] adopts a pretrained and frozen large language model [Brown et al. 2020; Devlin et al. 2018] as the text encoder, further improving the image fidelity and image-text alignment. Stable Diffusion [Rombach et al. 2022] is a particularly notable framework that trains the diffusion models on latent space, leading to reduced complexity and detail preservation. Meanwhile, some works are dedicated to spatial control [Voynov et al. 2023; Zhang et al. 2023b], concept control [Gal et al. 2022; Ruiz et al. 2022], and adopting knowledge-based retrieval for out-of-distribution generation [Blattmann et al. 2022; Chen et al. 2023b], *etc.* With the emergence of text-to-2D models, more fine-grained applications have been developed, including video generation [Ho et al. 2022], story visualization [Rahman et al. 2023], and text-guided image editing [Brooks et al. 2023].

Text-to-3D generation. The success of the 2D generation is incredible. However, directly transferring the image diffusion models to 3D is challenging, due to the difficulty of 3D data collection. Recently, Neural Radiance Fields (NeRF) [Barron et al. 2022; Mildenhall et al. 2020] opened a new insight for the 3D-aware generation, where only 2D multi-view images are needed in 3D scene reconstruction. Combining prior knowledge from text-to-2D models, several methods, such as DreamField [Jain et al. 2022], DreamFusion [Poole et al. 2022], and SJC [Wang et al. 2022a], have been proposed to generate 3D objects guided by text prompt [Li et al. 2023]. Moreover, the recent advancement of text-to-3D models also inspired multiple applications, including text-guided scenes generation [Cohen-Bar et al. 2023; Höllein et al. 2023], text-guided avatar generation [Cao et al. 2023; Jiang et al. 2023], and text-guided 3d model editing [Haque et al. 2023; Kamata et al. 2023].

3D Head Generation and Animation. Previous 3D head generation is primarily based on statistical models, such as 3DMM [Blanz

and Vetter 1999] and FLAME [Li et al. 2017], while current methods utilize 3D-aware Generative Adversarial Networks (GANs) [An et al. 2023; Chan et al. 2022, 2021; Schwarz et al. 2020]. Benefiting from advancements in dynamic scene representation [Cao and Johnson 2023; Fridovich-Keil et al. 2023; Gao et al. 2021], reconstructing animatable head avatars has significantly improved. Given a monocular video, these methods [Qian et al. 2023; Xu et al. 2023; Zheng et al. 2022, 2023; Zielonka et al. 2023] reconstruct a photo-realistic head avatar, and animate it based on FLAME. Specifically, our method was inspired by the technique [Qian et al. 2023; Zielonka et al. 2023] of deforming 3D points through rigging with FLAME mesh. We enhance its deformation and restriction to adapt to score distillation-based learning. On the other hand, the text-based 3D head generation methods [Han et al. 2023; Liu et al. 2023a; Wang et al. 2022b; Zhang et al. 2023a] show superiority in convenience and generalization. These methods demonstrate impressive texture and geometry, but are not animatable, limiting their practical application. Furthermore, TADA [Liao et al. 2023] and Bergman *et al.* [Bergman et al. 2023] explore the text-based animatable avatar generation. Similarly, we utilize FLAME to animate the head avatar, but we use 3DGS to model texture instead of the UV-map.

3 PRELIMINARY

In this section, we provide a brief overview of text-to-head generation. The generation process can be seen as distilling knowledge from a diffusion model ϵ_ϕ into a learnable 3D representation θ. Given camera poses, the corresponding views of the scene can be rendered as images. Subsequently, the distillation method guides the image to align with the text description y. Both the distillation method and the 3D representation are important and should be carefully designed.

Score Distillation Sampling has been proposed in DreamFusion [Poole et al. 2022]. For a rendered image x from a 3D representation, SDS introduces random noise ϵ to x at the t timestep, and then uses a pre-trained diffusion model ϵ_ϕ to predict the added noise. The SDS loss is defined as the difference between predicted and added noise and its gradient is given by

$$\nabla_\theta \mathcal{L}_{\text{SDS}} = \mathbb{E}_{t,\epsilon}[w(t)(\epsilon_\phi(x_t; y, t) - \epsilon)\frac{\partial_x}{\partial_\theta}], \tag{1}$$

where $x_t = \alpha_t x_0 + \sigma_t \epsilon$ and $w(t)$ is a weighting function. The loss estimates and update direction that follows the score function of the diffusion model to move x to a text description region.

3D Gaussian Splatting [Kerbl et al. 2023] is an efficient 3D representation. It reconstructs a static scene with anisotropic 3D Gaussian points, using paired image and camera pose. Each point is defined by a covariance matrix Σ centered at point μ:

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

Kerbl *et al.* [Kerbl et al. 2023] construct the semi-definite covariance matrix by defining an ellipse using a scaling matrix S and a rotation matrix R, ensuring that the points have meaningful representations:

$$\Sigma = RSS^T R^T. \tag{3}$$

The shape and position of a Gaussian point can be represented by a position vector $\mu \in \mathbb{R}^3$, a scaling vector $s \in \mathbb{R}^3$, and a quaternion $q \in \mathbb{R}^4$. Note that we refer r to represent the corresponding rotation matrix. Meanwhile, each 3D Gaussian point has additional parameters: color $c \in \mathbb{R}^3$ and opacity α, used for splatting-based rendering (we refer readers to [Kerbl et al. 2023] for the rendering details). Therefore, a scene can be represented by 3DGS as $\theta_{3DGS} = \{\mu, s, q, c, \alpha\}$.

4 METHOD

4.1 Semantic Alignment via FLAME

3D Gaussian splatting is commonly used in static avatar generation [Tang et al. 2023; Yi et al. 2023a], but applying it to dynamic avatar generation remains a challenging task. The difficulty lies in two aspects: 1) deform 3D Gaussian points with facial expression; 2) distill knowledge with facial expression. In general, the generation process lacks semantic alignment. To address this issue, we introduce FLAME [Li et al. 2017], a statistical head model, as an intermediate representation. Recent works have successfully achieved semantic alignment between FLAME and various human communication modalities, such as speech [He et al. 2023; Yi et al. 2023b] and facial expressions [Feng et al. 2021; Zielonka et al. 2022]. Thus, our focus turns to realizing semantic alignment between 3D Gaussian points and FLAME.

FLAME with Learnable Shape. FLAME [Li et al. 2017] is a vertex-based linear blend skinning (LBS) model, with $N = 5023$ vertices and 4 joints (neck, jaw, and eyeballs). The head animation can be formulated by a function:

$$M(\beta, \theta, \psi) : \mathbb{R}^{|\beta| \times |\gamma| \times |\psi|} \rightarrow \mathbb{R}^{3N}, \tag{4}$$

where $\beta \in \mathbb{R}^{|\beta|}$, $\gamma \in \mathbb{R}^{|\gamma|}$ and $\psi \in \mathbb{R}^{|\psi|}$ are the shape, pose and expression parameters, respectively (we refer readers to [Li et al. 2017; Loper et al. 2015] for the blendshape details).

Among them, the shape $\theta_{FLAME} = \{\beta\}$ is learnable, while the others are treated as animation inputs. The learnable shape allows for a more precise character model. For example, characters like the Hulk in Marvel have larger heads, whereas characters like Elisa in Frozen have thinner cheeks. However, excessive shape updates can negatively impact the learning process of 3DGS due to deformation changes. Therefore, we stop the shape update after a certain number of training steps to ensure stable learning of 3DGS.

FLAME-based dynamic head generation. We embed FLAME into the dynamic head generation process. Specifically, we first introduce the FLAME-based 3DGS (F-3DGS), which deforms 3D Gaussian points based on the FLAME mesh. Then, we present the FLAME-based SDS (F-SDS), using a FLAME-based control signal to guide score distillation. The FLAME-based restrictions designed in both 3D representation and score distillation assist HeadStudio in creating animatable and high-fidelity head avatars.

4.2 FLAME-based 3DGS

To deform 3D Gaussian points with facial expression, we introduce the **F**LAME-based **3D G**aussian **s**platting (**F-3DGS**). We first formally present the deformation process. Then, we indicate some improvement in initialization and optimization.

FLAME-based Gaussian Deformation. We assume every 3D Gaussian point is connected with a FLAME mesh. The FLAME mesh moves and deforms the corresponding points. Given any expression and pose, the FLAME mesh can be calculated by Eq. (4). Then, we quantify the mesh triangle by its mean position T, rotation matrix R and area S, which describe the triangle's location, orientation and scaling in world space, respectively. Among them, the rotation matrix is a concatenation of one edge vector, the normal vector of the triangle, and their cross-product. Given FLAME mesh, we deform the corresponding 3D Gaussian point as

$$r' = Rr, \tag{5}$$

$$\mu' = \sqrt{S}R\mu + T, \tag{6}$$

$$s' = \sqrt{S}s. \tag{7}$$

Intuitively, the 3D Gaussian point will be rotated, scaled and translated by the mesh triangle. As a result, FLAME enables the 3DGS to deform semantically, while 3DGS improves the texture representation and rendering efficiency of FLAME.

Initialization with Uniform Super-Resolution. Compared to reconstructing avatars, generating with score distillation involves a sparser control signal. It inspires us to initialize 3D Gaussian points that can thoroughly cover the head model for faster convergence and improved representation. Therefore, we introduce the points super-resolution that uniformly samples K points on each FLAME mesh. Specifically, the deformed 3D Gaussian points μ' are uniformly sampled on the FLAME mesh, in a standard pose with zero expression and pose parameters. The deformed scaling s' is the square root of the mean distance of its K-nearest neighbor points. Then, we initialize the mean position and scaling by the inversion of Eqs. (6) and (7): $\mu_{init} = R^{-1}((\mu' - T)/\sqrt{S}); s_{init} = s'/\sqrt{S}$. The other learnable parameters in θ_{3DGS} are initialized following vanilla 3DGS [Kerbl et al. 2023].

Optimization with Mesh Regularization. To deform semantically, the 3D Gaussian should align closely with the corresponding mesh triangle. Intuitively, the range of the mean position and scaling of 3D Gaussian points should be proportional to the size of the mesh triangle. For instance, in the eye and mouth region, where the mesh triangle is small, the 3D Gaussian points rigged on this mesh should also have a small scaling s and mean position μ. Therefore, we introduce the position and scaling regularization. For each triangle, we first compute the maximum distance among its mean position T and three vertices, termed as τ. It describes the maximum range of the 3D Gaussian. Then, the regularization term can be formulated as:

$$\mathcal{L}_{\text{pos}} = \| \max(\|\sqrt{S}R\mu\|_2, \tau_{\text{pos}})\|_2, \tag{8}$$

$$\mathcal{L}_{\text{s}} = \| \max(\sqrt{S}s, \tau_{\text{s}})\|_2, \tag{9}$$

where $\tau_{\text{pos}} = 0.5\tau$ and $\tau_{\text{s}} = 0.5\tau$ are the experimental position tolerance and scaling tolerance, respectively.

The regularization term is effective in the case of small mesh triangles in the mouth and eyes. However, when it comes to larger mesh triangles like those in the jaw with a mustache or the head with a hat, the strict regularization hampers the representation ability. Therefore, we introduce the scale factor and formulate the full regularization as:

$$\mathcal{L}_{\text{reg}} = (\lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{s}}\mathcal{L}_{\text{s}})/\sqrt{S}, \tag{10}$$

where $\lambda_{\text{pos}} = 0.1$ and $\lambda_{\text{s}} = 0.1$. With the help of regularization, F-3DGS shows the ability of semantic deformation.

4.3 FLAME-based SDS

Training with Animations. A straightforward method is training F-3DGS with a fixed pose and expression. While it produces satisfactory performance in static, it falls short in animation. To address this limitation, we incorporate training with animations. During the training process, we sample pose and expression from a motion sequence, such as TalkSHOW [Yi et al. 2023b], to ensure that the avatar satisfies the textual prompts with a diverse range of animation.

FLAME-based Control Generation. The vanilla SDS loss [Poole et al. 2022] performs effectively in static avatar generation. However, the data bias in the pre-trained diffusion model hinders its application in dynamic avatar generation. For example, the diffusion model prefers to assume that the person is looking straight, and the character's mouth is closed. Consequently, this leads to ambiguous supervision, and further results in improper coloring beyond the boundaries of the eyeballs and inability to separate the mouth.

To address this issue, we introduce the MediaPipe [Lugaresi et al. 2019] facial landmark map C, a fine-grained control signal marking the regions of upper lips, upper lips, eye boundary and eye balls. The facial landmarks in MediaPipe format can be extracted from FLAME, ensuring that the control signal aligns well with the F-3DGS. The loss gradient can be formulated as:

$$\nabla_\theta \mathcal{L}_{\text{F-SDS}} = \mathbb{E}_{t,\epsilon,\gamma,\psi}[w(t)(\epsilon_\phi(x_t; y, C, t) - \epsilon)\frac{\partial x}{\partial \theta}], \tag{11}$$

where $\theta = \theta_{\text{FLAME}} \cup \theta_{\text{3DGS}}$. Compared to SDS, F-SDS introduces more precise and detailed supervision.

Denoised Score Distillation. The vanilla SDS often results in non-detailed and blurry outputs due to noise gradients [Katzir et al. 2023; Wang et al. 2023; Zeiler 2012]. In image editing, Hertz *et al.* [Hertz et al. 2023] indicate that the predicted SDS gradients act as undesired noise when the rendered image matches the textual prompt. Therefore, the key to improving avatar generation lies in identifying matched prompts and removing undesired SDS gradients. Taking inspiration from Katzir *et al.* [Katzir et al. 2023], we assume that the rendered image with a large timestep $t > 200$ matches the negative textural prompts, such as y_{neg} = "unrealistic, blurry, low quality, out of focus, ugly, low contrast, dull, dark, low-resolution, gloomy". As a result, we reorganize the F-SDS in Eq. (11) by replacing the ϵ term with $\epsilon_\phi(x_t; y_{\text{neg}}, C, t)$). Intuitively, it leads to a cleaner gradient. In our experiment, the denoised score distillation leads to a better semantic alignment, benefiting an accurate animation.

4.4 Implementation Details

F-3DGS Details. In 3DGS, Kerbl *et al.* [Kerbl et al. 2023] employs a gradient threshold to filter points that require densification. Nevertheless, the original design cannot handle textual prompts with varying gradient responses. To address this, we utilize a normalized

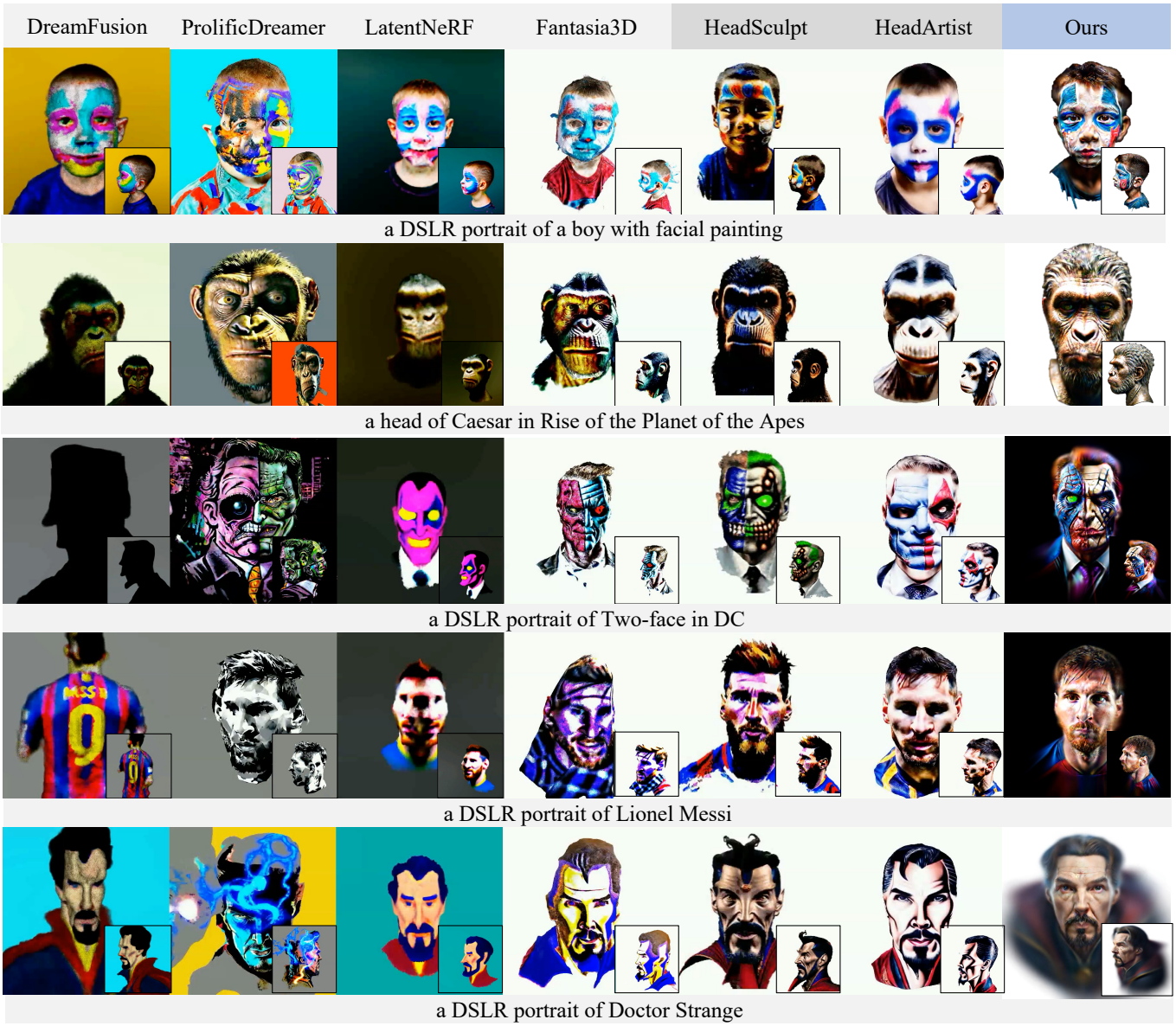

Fig. 3. Comparison with the text to static avatar generation methods. Our approach excels at producing high-fidelity head avatars, yielding superior results.

gradient to identify the points with consistent and significant gradient responses. Furthermore, the cloned and split points will inherit the same mesh triangle correspondence of their parent.

The F-3DGS is initialized with super-resolution $K = 10$. The entire 3DGS training consists of 10,000 iterations. The densification and pruning iterations setting are following [Liu et al. 2023b]. The overall framework is trained using the Adam optimizer [Kingma and Ba 2014], with betas of $[0.9, 0.99]$, and learning rates of 5e-5, 1e-3, 1e-2, 1.25e-2, 1e-2, and 1e-3 for mean position μ, scaling factor s, rotation quaternion q, color c, opacity α, and FLAME shape β, respectively. Note that we stop the FLAME shape optimization after 8,000 iterations.

F-SDS Details. In our experiment, we default to using Realistic Vision 5.1 (RV5.1) and ControlNetMediaPipFace [Zhang et al. 2023b]. Compared to Stable Diffusion 2.1 [Rombach et al. 2022], we observe that RV5.1 is capable of producing head avatars with a more visually appealing appearance. To alleviate the multi-face Janus problem, we also use the view-dependent prompts [Hong et al. 2023].

Training Details. The framework is implemented in PyTorch and threestudio [Guo et al. 2023]. We employ a random camera sampling strategy with camera distance range of $[1.5, 2.0]$, a fovy range of $[40°, 70°]$, an elevation range of $[-30°, 30°]$, and an azimuth range of $[-180°, 180°]$. We train head avatars with a resolution of 1024

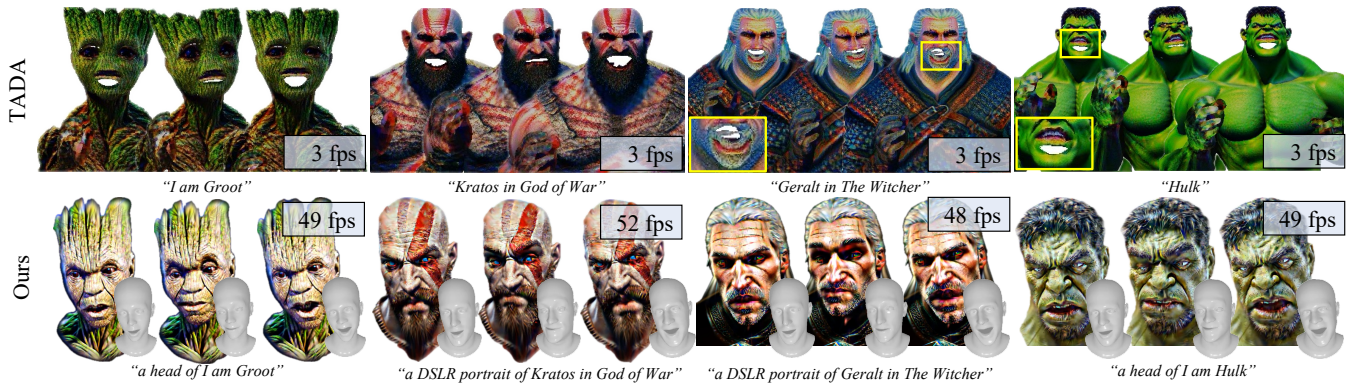

Fig. 4. Comparison with the text to dynamic avatar generation method TADA [Liao et al. 2023] in terms of semantic alignment and rendering speed. The yellow circles indicate semantic misalignment in the mouths, resulting in misplaced mouth texture. The rendering speed evaluation on the same device is reported in the blue box. The FLAME mesh of the avatar is visualized on the bottom right. Our method provides effective semantic alignment, smooth expression deformation, and real-time rendering.

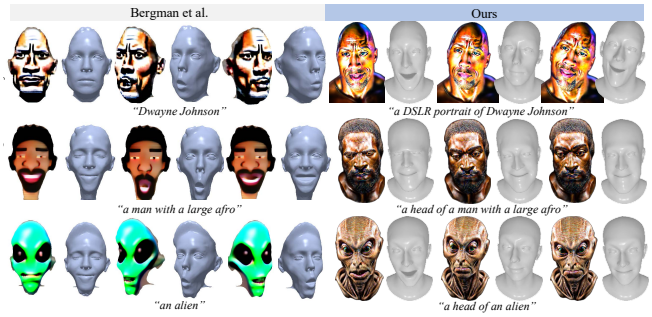

Fig. 5. Comparison with the text to dynamic avatar generation method, Bergman *et al.* [Bergman et al. 2023]. The FLAME mesh of the avatar is visualized on the bottom right. Our method demonstrates superior appearance and geometric modeling.

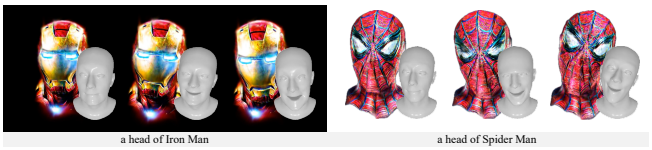

Fig. 6. Analysis of no mouth character generation. The FLAME mesh of the avatar is visualized on the bottom right. Our method effectively handles the generation of characters missing the mouth, avoiding holes in the mouth region.

Table 1. **Quantitative Evaluation.** Evaluating the coherence of generations with their caption using different CLIP models.

CLIP-Score	ViT-L/14↑	ViT-B/16 ↑	ViT-B/32 ↑
DreamFusion [Poole et al. 2022]	0.244	0.302	0.300
LatentNeRF [Metzer et al. 2022]	0.248	0.299	0.303
Fantasia3D [Chen et al. 2023a]	0.267	0.304	0.300
ProlificDreamer [Wang et al. 2023]	0.268	0.320	0.308
HeadSculpt [Han et al. 2023]	0.264	0.306	0.305
HeadArtist [Liu et al. 2023a]	0.272	0.318	0.313
Ours	**0.275**	**0.322**	**0.317**

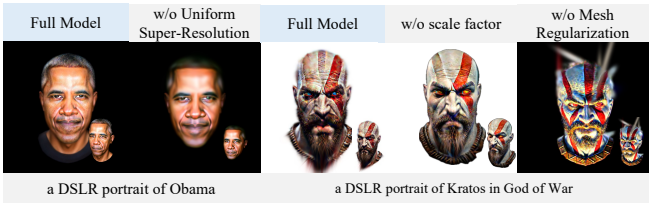

Fig. 7. **Ablation Study of F-3DGS.** We present the effect of uniform super-resolution and mesh regularization. Uniform super-resolution results in a beneficial initialization and enhances the representation ability. Mesh regularization imposes a strong restriction to reduce the outline points. The scale factor in mesh regularization balances restriction and expressiveness.

and a batch size of 8. The entire optimization process takes around two hours on a single NVIDIA A6000 (48GB) GPU.

5 EXPERIMENT

Evaluation. We evaluate the quality of head avatars with two settings. 1) *static head avatars*: producing a diverse range of avatars based on various text prompts. 2) *dynamic avatars*: driving an avatar with FLAME sequences sampled in TalkSHOW [Yi et al. 2023b].

Baselines. We compare our method with state-of-the-art methods in two settings. 1) *static head avatars*: We compare the generation results with six baselines: DreamFusion [Poole et al. 2022], LatentNeRF [Metzer et al. 2022], Fantasia3D [Chen et al. 2023a] and ProlificDreamer [Wang et al. 2023], HeadSculpt [Han et al. 2023] and HeadArtist [Liu et al. 2023a]. Among them, HeadSculpt [Han et al. 2023] and HeadArtist [Liu et al. 2023a] specialize in text to head avatar generation. 2) *dynamic head avatars*: We evaluate the efficacy of avatar animation by comparing it with TADA [Liao et al. 2023] and Bergman *et al.* [Bergman et al. 2023]. Both approaches are based on FLAME and utilize it for animation.

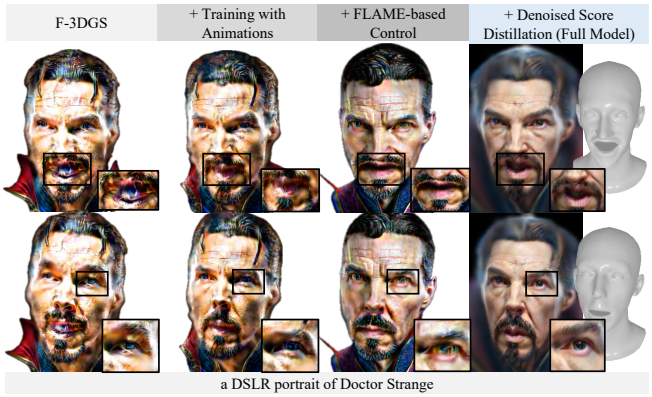

Fig. 8. **Ablation Study of F-SDS.** We investigate the effects of training with animation, FLAME-based control, and denoised score distillation. These approaches are dedicated to improving the semantic accuracy of score distillation. As a result, F-SDS achieves an effective alignment, leading to an accurate expression deformation.

5.1 Head Avatar Generation

We evaluate the avatar generation quality in terms of geometry and texture. In Fig. 3, we evaluate the geometry through novel-view synthesis. Comparatively, the head-specialized methods produce avatars with superior geometry compared to the text-to-3D methods [Chen et al. 2023a; Metzer et al. 2022; Poole et al. 2022; Wang et al. 2023]. This improvement can be attributed to the integration of FLAME, a reliable head structure prior, which mitigates the multi-face Janus problem [Hong et al. 2023] and enhances the geometry.

On the other hand, we evaluate the texture through quantitative experiments using the CLIP score [Hessel et al. 2021]. This metric measures the similarity between the given textual prompt and the generated avatars. A higher CLIP score indicates a closer match between the generated avatar and the text, highlighting a more faithful texture. Following Liu *et al.* [Liu et al. 2023a], we report the average CLIP score of 10 text prompts. Table 1 demonstrates that HeadStudio outperforms other methods in three different CLIP variants [Radford et al. 2021]. Overall, HeadStudio excels at producing high-fidelity head avatars, outperforming the state-of-the-art text-based methods.

5.2 Head Avatar Animation

We evaluate the efficiency of animation in terms of semantic alignment and rendering speed. For the evaluation of semantic alignment, we visually represent the talking head sequences, which are controlled by speech [Yi et al. 2023b]. In Fig. 4, we compare HeadStudio with TADA [Liao et al. 2023]. The yellow circles in the first row indicate a lack of semantic alignment in the mouths of Hulk and Geralt, resulting in misplaced mouth texture. Our approach utilizes F-SDS and F-3DGS, which enable excellent semantic alignment and smooth expression deformation. On the other hand, our method enables real-time rendering. When compared to TADA, such as Kratos (52 fps v.s. 3 fps), our method demonstrates its potential in augmented or virtual reality applications.

Furthermore, the comparison in Fig. 5 indicates the semantic alignment in the method proposed by [Bergman et al. 2023]. Nevertheless, it lacks in terms of its representation of appearance and geometry. Moreover, as depicted in Figure 6, our approach effectively creates animatable avatars of Iron Man and Spider Man. Our method avoids creating holes in the mouth and effectively handles the generation of characters without a mouth.

5.3 Ablation Study

We isolate the various contributions and conducted a series of experiments to assess their impact. In particular, we examine the design of F-SDS and F-3DGS. For F-3DGS, we examined the impact of uniform super-resolution and mesh regularization. Regarding F-SDS, we assessed the influence of training with animation, FLAME-based control, and denoised distillation.

Effect of FLAME-based 3DGS. In Fig 7, we present the effect of uniform super-resolution and mesh regularization. Since the F-SDS supervision signal is sparse, super-resolution enhances point coverage on the head model, leading to a favorable initialization and improved avatar fidelity. Conversely, mesh regularization reduces the outline points. Nevertheless, overly strict regularization weaken the representation ability of F-3DGS, such as the beard of Kratos (fourth column in Fig. 7). To address this, we introduce a scale factor to balance restriction and expressiveness based on the area of mesh triangle. Consequently, the restriction of Gaussian points rigged on jaw mesh has been reduced, resulting in a lengthier beard for Kratos (third column in Fig. 7).

Effect of FLAME-based SDS. As illustrated in Fig. 8, we visualize the effect of each component in F-3DS. By utilizing rigged deformation, the generated avatar can be controlled by expression sequences. However, as depicted in the first column, it exhibits noticeable artifacts. This reveals that F-3DGS, despite employing strict mesh regularization, struggles to handle the animation. This limitation arises from the semantic misalignment, where the supervision signal fails to precisely identify the intended area. We then use training with animation to separate the different areas, such as the upper lips and lower lips. Furthermore, we introduce the fine-grained semantic guidance, the Mediapipe [Lugaresi et al. 2019] facial landmark map, which is used in conjunction with ControlNet [Zhang et al. 2023b]. The result (third column in Fig. 8) shows its effectiveness in addressing the issue of mouths sticking together. Subsequently, the denoised score distillation helps eliminate undesired noise in the gradient, further enhancing performance.

6 CONCLUSION

In this paper, we propose HeadStudio, a novel pipeline for generating high-fidelity and animatable 3D head avatars using 3D Gaussian Splatting. We use FLAME as a intermediate representation, embedding into the 3D representation and score distillation. As a result, we deform 3D Gaussian points by rigging each 3D Gaussian point into a FLAME mesh. Additionally, using FLAME-based fine-grained control signal to guide score distillation. Extensive evaluations demonstrated that our HeadStudio produces high-fidelity and animatble avatars with real-time rendering, outperforming state-of-the-art methods significantly.

REFERENCES

Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. 2023. PanoHead: Geometry-Aware 3D Full-Head Synthesis in 360°. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 20950–20959.

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 5470–5479.

Alexander W Bergman, Wang Yifan, and Gordon Wetzstein. 2023. Articulated 3d head avatar generation using text-to-image diffusion models. arXiv preprint arXiv:2307.04859 (2023).

Volker Blanz and Thomas Vetter. 1999. A Morphable Model for the Synthesis of 3D Faces. In SIGGRAPH. 8 pages. https://doi.org/10.1145/311535.311556

Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. 2022. Retrieval-augmented diffusion models. Advances in Neural Information Processing Systems 35 (2022), 15309–15324.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 18392–18402.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in Neural Information Processing Systems (NeurIPS) 33 (2020), 1877–1901.

Ang Cao and Justin Johnson. 2023. HexPlane: A Fast Representation for Dynamic Scenes. CVPR (2023).

Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. 2023. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. arXiv preprint arXiv:2304.00916 (2023).

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. 2022. Efficient Geometry-aware 3D Generative Adversarial Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 5799–5809.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. 2023a. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. 2023b. Re-imagen: Retrieval-augmented text-to-image generator. Proceedings of the International Conference on Learning Representation (ICLR) (2023).

Dana Cohen-Bar, Elad Richardson, Gal Metzer, Raja Giryes, and Daniel Cohen-Or. 2023. Set-the-Scene: Global-Local Training for Generating Controllable NeRF Scenes. arXiv preprint arXiv:2303.13450 (2023).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).

Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. 2021. Learning an Animatable Detailed 3D Face Model from In-The-Wild Images. ACM Transactions on Graphics, (Proc. SIGGRAPH) 40, 8. https://doi.org/10.1145/3450626.3459936

Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. 2023. K-Planes: Explicit Radiance Fields in Space, Time, and Appearance. In CVPR.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022).

Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. 2021. Dynamic View Synthesis from Dynamic Monocular Video. In Proceedings of the IEEE International Conference on Computer Vision.

Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. 2023. threestudio: A unified framework for 3D content generation. https://github.com/threestudio-project/threestudio.

Xiao Han, Yukang Cao, Kai Han, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, Tao Xiang, and Kwan-Yee K Wong. 2023. HeadSculpt: Crafting 3D Head Avatars with Text. arXiv preprint arXiv:2306.03038 (2023).

Ayaan Haque, Matthew Tancik, Alexei Efros, Aleksander Holynski, and Angjoo Kanazawa. 2023. Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

Shan He, Haonan He, Shuo Yang, Xiaoyan Wu, Pengcheng Xia, Bing Yin, Cong Liu, Lirong Dai, and Chang Xu. 2023. Speech4Mesh: Speech-Assisted Monocular 3D Facial Reconstruction for Speech-Driven 3D Facial Animation. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 14192–14202.

Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. 2023. Delta denoising score. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 2328–2337.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021).

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems (NeurIPS) 33 (2020), 6840–6851.

Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. 2023. Text2room: Extracting textured 3d meshes from 2d text-to-image models. arXiv preprint arXiv:2303.11989 (2023).

Susung Hong, Donghoon Ahn, and Seungryong Kim. 2023. Debiasing scores and prompts of 2d diffusion for robust text-to-3d generation. arXiv preprint arXiv:2303.15413 (2023).

Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. 2022. Zero-Shot Text-Guided Object Generation with Dream Fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. 2023. AvatarCraft: Transforming Text into Neural Human Avatars with Parameterized Shape and Pose Control. arXiv preprint arXiv:2303.17606 (2023).

Hiromichi Kamata, Yuiko Sakuma, Akio Hayakawa, Masato Ishii, and Takuya Narihira. 2023. Instruct 3D-to-3D: Text Instruction Guided 3D-to-3D conversion. arXiv preprint arXiv:2303.15780 (2023).

Oren Katzir, Or Patashnik, Daniel Cohen-Or, and Dani Lischinski. 2023. Noise-free score distillation. arXiv preprint arXiv:2310.17590 (2023).

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Transactions on Graphics 42, 4 (July 2023). https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).

Chenghao Li, Chaoning Zhang, Atish Waghwase, Lik-Hang Lee, Francois Rameau, Yang Yang, Sung-Ho Bae, and Choong Seon Hong. 2023. Generative AI meets 3D: A Survey on Text-to-3D in AIGC Era. arXiv preprint arXiv:2305.06131 (2023).

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. ACM Trans. Graph. 36, 6 (2017), 194–1.

Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxaing Tang, Yangyi Huang, Justus Thies, and Michael J Black. 2023. Tada! text to animatable digital avatars. arXiv preprint arXiv:2308.10899 (2023).

Hongyu Liu, Xuan Wang, Ziyu Wan, Yujun Shen, Yibing Song, Jing Liao, and Qifeng Chen. 2023a. HeadArtist: Text-conditioned 3D Head Generation with Self Score Distillation. arXiv preprint arXiv:2312.07539 (2023).

Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, and Ziwei Liu. 2023b. HumanGaussian: Text-Driven 3D Human Generation with Gaussian Splatting. arXiv preprint arXiv:2311.17061 (2023).

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-person Linear Model. ACM Trans. Graph. 34, 6, Article 248 (Oct. 2015), 16 pages.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019).

Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. 2022. Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures. arXiv preprint arXiv:2211.07600 (2022).

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In Proceedings of the European Conference on Computer Vision (ECCV).

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021).

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 10975–10985. http://smpl-x.is.tue.mpg.de

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2022. DreamFusion: Text-to-3D using 2D Diffusion. arXiv preprint arXiv:2209.14988 (2022).

Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. 2023. GaussianAvatars: Photorealistic Head Avatars with Rigged 3D Gaussians. arXiv preprint arXiv:2312.02069 (2023).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning (ICML). 8748–8763.

Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. 2023. Make-a-story: Visual memory conditioned consistent story generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2493–2502.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 10684–10695.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation. arXiv preprint arxiv:2208.12242 (2022).

Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. 2020. Graf: Generative radiance fields for 3d-aware image synthesis. Advances in Neural Information Processing Systems 33 (2020), 20154–20166.

Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. 2021. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. Advances in Neural Information Processing Systems 34 (2021), 6087–6101.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In International Conference on Machine Learning. PMLR, 2256–2265.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. 2023. DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation. arXiv preprint arXiv:2309.16653 (2023).

Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. 2023. Sketch-guided text-to-image diffusion models. In ACM SIGGRAPH 2023 Conference Proceedings. 1–11.

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. 2022a. Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. 2022b. Rodin: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion. arXiv preprint arXiv:2212.06135 (2022).

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. 2023. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. arXiv preprint arXiv:2305.16213 (2023).

Yuelang Xu, Lizhen Wang, Xiaochen Zhao, Hongwen Zhang, and Yebin Liu. 2023. AvatarMAV: Fast 3D Head Avatar Reconstruction Using Motion-Aware Neural Voxels. In ACM SIGGRAPH 2023 Conference Proceedings.

Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. 2023b. Generating Holistic 3D Human Motion from Speech. In CVPR.

Taoran Yi, Jiemin Fang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. 2023a. Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. arXiv preprint arXiv:2310.08529 (2023).

Matthew D. Zeiler. 2012. ADADELTA: AN ADAPTIVE LEARNING RATE METHOD. arXiv preprint arXiv:1212.5701 (2012).

Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. 2023c. Text-to-image diffusion model in generative ai: A survey. arXiv preprint arXiv:2303.07909 (2023).

Longwen Zhang, Qiwei Qiu, Hongyang Lin, Qixuan Zhang, Cheng Shi, Wei Yang, Ye Shi, Sibei Yang, Lan Xu, and Jingyi Yu. 2023a. DreamFace: Progressive Generation of Animatable 3D Faces under Text Guidance. arXiv preprint arXiv:2304.03117 (2023).

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023b. Adding Conditional Control to Text-to-Image Diffusion Models.

Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. 2022. I M Avatar: Implicit Morphable Head Avatars from Videos. In Computer Vision and Pattern Recognition (CVPR).

Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. 2023. PointAvatar: Deformable Point-based Head Avatars from Videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Wojciech Zielonka, Timo Bolkart, and Justus Thies. 2022. Towards Metrical Reconstruction of Human Faces. European Conference on Computer Vision.

Wojciech Zielonka, Timo Bolkart, and Justus Thies. 2023. Instant Volumetric Head Avatars. Conference on Computer Vision and Pattern Recognition.