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

Zhenglin Zhou[1,2], Fan Ma[2], Hehe Fan[2], Zongxin Yang[2], and Yi Yang[1,2†]

[1] State Key Laboratory of Brain-machine Intelligence, Zhejiang University, China
[2] ReLER, CCAI, Zhejiang University, China
{zhenglinzhou, mafan, hehefan, yangzongxin, yangyics}@zju.edu.cn

Abstract. Creating digital avatars from textual prompts has long been a desirable yet challenging task. Despite the promising results achieved with 2D diffusion priors, current methods struggle to create high-quality and consistent animated avatars efficiently. Previous animatable head models like FLAME have difficulty in accurately representing detailed texture and geometry. Additionally, high-quality 3D static representations face challenges in semantically driving with dynamic priors. In this paper, we introduce **HeadStudio**, a novel framework that utilizes 3D Gaussian splatting to generate realistic and animatable avatars from text prompts. Firstly, we associate 3D Gaussians with animatable head prior model, facilitating semantic animation on high-quality 3D representations. To ensure consistent animation, we further enhance the optimization from initialization, distillation, and regularization to jointly learn the shape, texture, and animation. Extensive experiments demonstrate the efficacy of HeadStudio in generating animatable avatars from textual prompts, exhibiting appealing appearances. The avatars are capable of rendering high-quality real-time ($\geq$ 40 fps) novel views at a resolution of 1024. Moreover, These avatars can be smoothly driven by real-world speech and video. We hope that HeadStudio can enhance digital avatar creation and gain popularity in the community. Code is at: `https://github.com/ZhenglinZhou/HeadStudio`.

Keywords: Head avatar animation · Text-guided generation · 3D Gaussian splatting

1 Introduction

With the development of deep learning, head avatar generation has improved significantly in recent years. At first, the image-based methods [11,81] are proposed to reconstruct the photo-realistic head avatar of a person, given one or more views. Recently, generative models (*e.g.* diffusion [55,73]) 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 [21,42],

[†] Corresponding author.

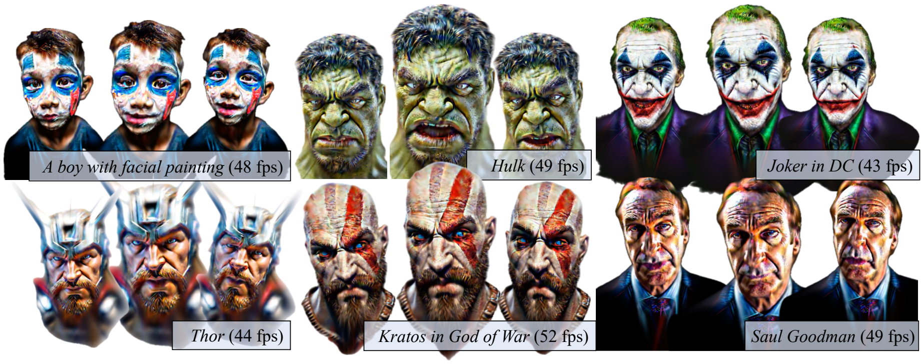

Fig. 1: Text-based animatable avatars generation by **HeadStudio**. With only one end-to-end training stage of 2 hours on 1 NVIDIA A6000 GPU, HeadStudio is able to generate animatable, high-fidelity and real-time rendering ($\geq$ 40 fps) head avatars using text inputs.

which have shown superiority over image-based methods in convenience and generalization.

However, current text-based methods cannot combine high-quality and animation effectively. For instance, HeadSculpt [21] leverages DMTet [58] for high-quality optimization and creates highly detailed head avatars but is unable to be animated. TADA [41] employs SMPL-X [51] to generate animatable digital characters but sacrifices appearance quality. There is always a trade-off between static quality and dynamic animation within current methods. We attribute it to two prominent drawbacks: (1) **Limitations in representation**: the animatable head prior model struggles to model high-quality texture and geometry (refer to Fig. 5 and Fig. 6); (2) **Challenges in optimization**: aligning the static representation with the dynamic head prior is difficult (refer to Fig. 8).

In this paper, we propose a novel text-based generation framework, named **HeadStudio**, by fully exploiting 3D Gaussian splatting (3DGS) [35], which achieves superior rendering quality and real-time performance for novel-view synthesis. Our method comprises two components: (1) **Animatable Head Gaussian**: We first arm FLAME [39], an animatable head prior model, with 3D Gaussian splatting by rigging each 3D Gaussian point to a mesh. As an animatable head Gaussian model, we use the head prior model, to deform 3D Gaussians and employ them for high-quality texture and geometry modeling. (2) **Text to Avatar Optimization**: We enhance the optimization from initialization, distillation, and regularization to jointly learn the shape, texture, and animation, improving the visual appearance and animated quality. In specific, we introduce super-dense Gaussian initialization to thoroughly cover the head model for faster convergence and improved representation. To ensure the consistency of the control signal during animation-based training, we denoise the score distillation and utilize the MediaPipe [45] facial landmark map obtained from FLAME as a fine-grained condition for the diffusion model. To further improve the fidelity of our method, we utilize an adaptive geometry regularization, which gives animatable head Gaussian the ability to employ strict constraints for semantic deformation and represent elements beyond the FLAME space, such as helmets and mustaches simultaneously.

Extensive experiments have shown that HeadStudio is highly effective and superior to state-of-the-art methods in generating dynamic avatars from text [21, 41, 48, 52, 64, 72]. Moreover, our methods can be easily extended to driving generated 3D avatars via both speech-based [69] and video-based [16] 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 arms animatable head prior model with 3DGS and enhances its optimization for creating high-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 [54] and diffusion models [27, 60], great advancements have been made in text-to-image generation (T2I) [26, 50, 70]. In particular, Stable Diffusion [55] is a notable framework that trains the diffusion models on latent space, leading to reduced complexity and detail preservation. With the emergence of text-to-2D models, more applications have been developed [46, 68, 75], such as spatial control [61, 73, 78], concept control [18, 40, 56], and image editing [8].

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) [5, 49] 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 [31], DreamFusion [52], and SJC [62], have been proposed to generate 3D objects guided by text prompt [38, 79]. Moreover, the recent advancement of text-to-3D generation also inspired multiple applications, including text-guided scenes generation [15, 29], text-guided 3D editing [22, 33], and text-guided avatar generation [10, 32, 47, 66].

3D Head Generation and Animation. Previous 3D head generation is primarily based on statistical models, such as 3DMM [7] and FLAME [39], while current methods utilize 3D-aware Generative Adversarial Networks (GANs) [4, 11, 12, 57, 59, 65, 74]. Benefiting from advancements in dynamic scene representation [9, 17, 19], animatable head avatars reconstruction has been improved. Given a monocular video or multi-view videos, these methods [37, 53, 67, 76, 77, 81] reconstruct a photo-realistic head avatar, and animate it based on FLAME. Specifically, our method was inspired by the technique [53, 81] of deforming 3D

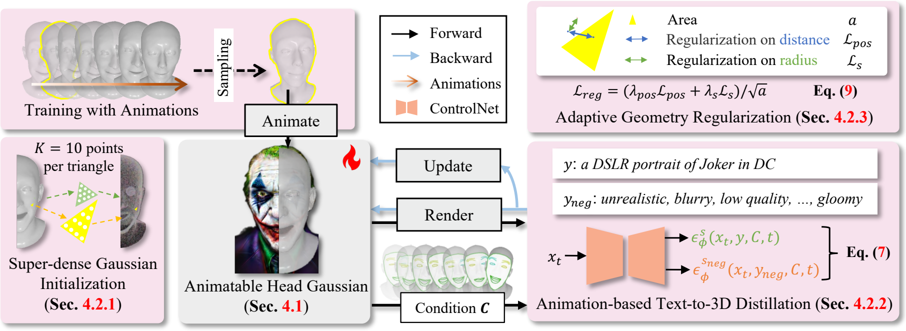
Training with Animations
Sampling
Animate
Forward
Backward
Animations
ControlNet
Area a
Regularization on distance $\mathcal{L}_{pos}$
Regularization on radius $\mathcal{L}_s$
$\mathcal{L}_{reg} = (\lambda_{pos}\mathcal{L}_{pos} + \lambda_s\mathcal{L}_s)/\sqrt{a}$ Eq. (9)
Adaptive Geometry Regularization (Sec. 4.2.3)
$K = 10$ points per triangle
Super-dense Gaussian Initialization (Sec. 4.2.1)
Animatable Head Gaussian (Sec. 4.1)
Update
Render
Condition $\boldsymbol{C}$
y: a DSLR portrait of Joker in DC
y_{neg}: unrealistic, blurry, low quality, ..., gloomy
x_t
$\epsilon_\phi^s(x_t, y, C, t)$
$\epsilon_\phi^{s_{neg}}(x_t, y_{neg}, C, t)$
Eq. (7)
Animation-based Text-to-3D Distillation (Sec. 4.2.2)

Fig. 2: Framework of HeadStudio, which integrates animatable head prior model into 3D Gaussian splatting and score distillation sampling. 1) **Animatable Head Gaussian**: each 3D point is rigged to a mesh, and then rotated, scaled, and translated by the mesh deformation. 2) **Text to Avatar Optimization**: enhance the optimization from initialization, distillation and regularization, including: super-dense Gaussian initialization, animation-based text-to-3D distillation, and adaptive geometry regularization.

points through rigging with FLAME mesh. We enhance its deformation and optimization to adapt to score distillation-based learning. On the other hand, the text-to-static head avatar methods [21,42,63,72] show superiority in convenience and generalization. These methods demonstrate impressive texture and geometry, but are not animatable, limiting their practical application. Furthermore, TADA [41] and Bergman *et al.* [6] explore the text-to-dynamic head 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 avatar 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.

Score Distillation Sampling has been proposed in DreamFusion [52]. 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

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

where $x_t = \alpha_t x_0 + \sigma_t \epsilon$ and $w(t)$ is a weighting function, and s is a pre-defined scalar of classifier-free guidance (CFG) [28]. The loss estimates and update direction that follows the score function of the diffusion model to move x to the text description region.

3D Gaussian Splatting [35] is an efficient 3D representation. It reconstructs a static scene with anisotropic 3D Gaussians, using paired image and camera pose. Each point is defined by a covariance matrix $\boldsymbol{\Sigma}$ centered at point $\boldsymbol{\mu}$:

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

Kerbl *et al.* [35] construct the semi-definite covariance matrix by defining an ellipse using a scaling matrix $\boldsymbol{S}$ and a rotation matrix $\boldsymbol{R}$, ensuring that the points have meaningful representations:

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

The shape and position of a Gaussian point can be represented by a position vector $\boldsymbol{\mu} \in \mathbb{R}^3$, a scaling vector $\boldsymbol{s} \in \mathbb{R}^3$, and a quaternion $\boldsymbol{q} \in \mathbb{R}^4$. Note that we refer $\boldsymbol{R}$ to represent the corresponding rotation matrix. Meanwhile, each 3D Gaussian point has additional parameters: color $\boldsymbol{c}$ and opacity $\boldsymbol{\alpha}$, used for splatting-based rendering. Therefore, a scene can be represented by 3DGS as $\theta_{3DGS} = \{\boldsymbol{\mu}, \boldsymbol{s}, \boldsymbol{q}, \boldsymbol{c}, \boldsymbol{\alpha}\}$. Given a camera view, the scene can be rendered by the 2D projection of Gaussians via a differentiable tile rasterizer. In optimization, the gradient of Gaussians is utilized to guide the densification and prune of Gaussians. We refer readers to [13, 35] for more details.

4 Method

HeadStudio is a text-to-dynamic head avatar geneartion method. The created head avatars can be animated by text, speech, and video. As illustrated in Fig. 2, the generation pipeline has two key components, including (1) the animatable head Gaussian in Sec. 4.1, and (2) text to avatar optimization in Sec. 4.2. Implementation details are discussed in Sec. 4.3

4.1 Animatable Head Gaussian

Animatable Head Prior Model. FLAME [39] 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:

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

where $\boldsymbol{\beta} \in \mathbb{R}^{|\boldsymbol{\beta}|}$, $\boldsymbol{\gamma} \in \mathbb{R}^{|\boldsymbol{\gamma}|}$ and $\boldsymbol{\psi} \in \mathbb{R}^{|\boldsymbol{\psi}|}$ are the shape, pose and expression parameters, respectively (we refer readers to [39, 44] for the blendshape details).

Recent works have successfully achieved semantic alignment between FLAME and various modalities, such as speech [23, 69] and talking videos [16, 80]. Therefore, existing text-to-dynamic avatar generation methods [6, 41] commonly choose FLAME [39] as the base model. As a result, the created avatars can be semantically animated. However, the mesh number of FLAME is struggled to model complex textures. For example, Bergman *et al.* [6] learns one color for each mesh. It inspires us to arm FLAME with 3D Gaussian points [35] for high-quality texture modeling.

Deformable Gaussian Texture. To mitigate the limitations of animatable head prior model, we use 3D Gaussian points to model the texture. The key point is to make sure these points can be deformed semantically by the head prior model. Following Qian *et al.* [53], we assume every 3D Gaussian point is connected with a FLAME mesh. The FLAME mesh moves and deforms the corresponding points. Given pose and expression, the FLAME mesh can be calculated by Eq. (4). Then, we quantify the mesh triangle by its center position $\boldsymbol{t}$, rotation matrix $\tilde{\boldsymbol{R}}$ and area a, 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. Formally, we deform the corresponding 3D Gaussian point as

$$\boldsymbol{R}' = \tilde{\boldsymbol{R}}\boldsymbol{R}, \qquad \boldsymbol{\mu}' = \sqrt{a}\tilde{\boldsymbol{R}}\boldsymbol{\mu} + \boldsymbol{t}, \qquad \boldsymbol{s}' = \sqrt{a}\boldsymbol{s}, \tag{5}$$

where $\boldsymbol{\mu}'$, $\boldsymbol{s}'$ and $\boldsymbol{R}'$ are the position vector, scaling vector and rotation matrix of the deformed Gaussian for rendering. Intuitively, the 3D Gaussian point will be rotated, scaled, and translated by the mesh triangle. In this way, Gaussians can be seen as a residual term of FLAME to represent intricate geometry and texture. As a result, FLAME enables the 3DGS to animate semantically, while 3DGS improves the texture representation and rendering efficiency of FLAME.

Joint Learning of Shape, Texture, Animation. The intricate texture can be modeled by the deformable Gaussian texture θ_{3DGS}. Besides, we assume the shape of head prior model $\theta_{\text{FLAME}} = \{\boldsymbol{\beta}\}$ is learnable. The learnable shape allows for modeling character more precisely. For example, characters like the Hulk in Marvel have larger heads, whereas characters like Elsa in Frozen have thinner cheeks. Meanwhile, we notice that excessive shape updates can negatively impact the learning process of 3DGS due to deformation changes. Thus, we stop the shape update after a certain number of training steps to ensure stable learning of 3DGS. As a result, a head avatar can be represented by an animatable head Gaussian as $\theta = \theta_{\text{FLAME}} \cup \theta_{\text{3DGS}}$.

4.2 Text to Avatar Optimization

To jointly learn the shape, texture, and animation of an animatable head Gaussian, we enhance its optimization from initialization, distillation, and regularization, respectively.

Super-dense Gaussian Initialization. The supervision signal of SDS loss [52] in head avatar generation is sparse. It inspires us to initialize 3D Gaussians that thoroughly cover the head model for faster convergence and improved representation. In specific, each mesh triangle is initialized with K evenly distributed points. The positions of the deformed 3D Gaussians $\boldsymbol{\mu}'$ are calculated by sampling on the FLAME model (with standard pose), with all mesh triangles sharing the same sampling weight. The deformed scaling $\boldsymbol{s}'$ is the square root of the mean distance of its K-nearest neighbor points. Then, we initialize the position and scaling of 3D Gaussians by the inversion of Eq. (5): $\boldsymbol{\mu}_{init} = \tilde{\boldsymbol{R}}^{-1}((\boldsymbol{\mu}' - \boldsymbol{t})/\sqrt{a}); \boldsymbol{s}_{init} = \boldsymbol{s}'/\sqrt{a}$. The other learnable parameters in θ_{3DGS} are initialized following vanilla 3DGS [35].

Animation-based Text-to-3D Distillation. The vanilla text-to-3D distillation [52] produces satisfactory performance in static but falls short in animation. We attribute it to the absence of new poses and expressions in training. Therefore, we design a new text-to-3D distillation that adapts to animation.

Training with Animations. We first incorporate the new pose and expression into training [41,71]. Specifically, we sample pose and expression from real-world motion sequences, such as TalkSHOW [69], to ensure that the avatar satisfies the textual prompts with a diverse range of animation.

FLAME-based Control Generation. Training with animations is crucial for dynamic avatar generation. However, the direct introduction of new pose and expression results in Janus (multi-faces) problem [30], due to the data bias in the diffusion model. This issue, represented as portrait bias with front-view, straight-looking, and closed mouths, hinders its application in animation-based distillation. To address this issue, we introduce the MediaPipe [45] facial landmark map C, a fine-grained control signal marking the regions of upper lips, lower lips, eye boundary, eyeballs, and facial boundary [21,42], for more precise and detailed guidance. It can be extracted from an animatable head Gaussian, which ensures that the control signal aligns well with the Gaussian points when the shape, pose, and expression change. The loss gradient is formulated as:

$$\nabla_\theta \mathcal{L}_{\text{SDS}} = \mathbb{E}_{t,\epsilon,\boldsymbol{\gamma},\boldsymbol{\psi}}[w(t)(\epsilon^s_\phi(x_t;y,C,t) - \epsilon)\frac{\partial_x}{\partial_\theta}]. \tag{6}$$

Denoised Score Distillation. According to our experiments, we find the generated avatars have non-detailed and over-smooth textures. To solve this issue, we consider the distilled score to be noisy [24,34,64]. Hertz *et al.* [24] indicates that the score can be seen as the noise when the rendered image matches the textual prompt. Following NFSD [34], we assume the score with a large timestep $t \geq 200$ is noisy, and the rendered image can be seen as matching the negative textural prompts, such as $y_{\text{neg}} =$ "*unrealistic, blurry, low quality, out of focus, ugly, low contrast, dull, dark, low-resolution, gloomy*". Besides, the score with a small timestep $t < 200$ is relatively clean. As a result, we reorganize the SDS into a piece-wise function:

$$\nabla_\theta \mathcal{L}_{\text{SDS}} = \begin{cases} \mathbb{E}_{t,\epsilon,\boldsymbol{\gamma},\boldsymbol{\psi}}[w(t)\epsilon^s_\phi(x_t;y,C,t)\frac{\partial_x}{\partial_\theta}], & \text{t} < 200, \\ \mathbb{E}_{t,\epsilon,\boldsymbol{\gamma},\boldsymbol{\psi}}[w(t)(\epsilon^s_\phi(x_t;y,C,t) - \epsilon^{s_{neg}}_\phi(x_t;y_{\text{neg}},C,t))\frac{\partial_x}{\partial_\theta}], & \text{t} \geq 200, \end{cases} \tag{7}$$

where s_{neg} is a pre-defined CFG scalar for negative textual prompts. Intuitively, we get a cleaner score to improve the avatar's texture.

Adaptive Geometry Regularization. To deform semantically, the 3D Gaussians should closely align with the rigged mesh triangle. Introducing a regularization term for the 3D Gaussians, such as $\|\boldsymbol{\mu}\|_2$, will lead to the 3D Gaussians being overly concentrated around the mesh center. Thus, the regularization should inversely scale with the triangle size. For instance, in the eye and mouth region, where the mesh triangle is small, the rigged Gaussians should have a relatively small scaling and position. Following Qian *et al.* [53], we introduce the position and scaling regularization. For each triangle, we initially calculate the maximum distance among its center $\boldsymbol{t}$ and three vertices, termed as τ, to describe the triangle size. Then, we formulate the regularization term as:

$$\mathcal{L}_{\text{pos}} = \| \max(\|\sqrt{a}\boldsymbol{R}'\boldsymbol{\mu}\|_2, \tau_{\text{pos}})\|_2, \qquad \mathcal{L}_{\text{s}} = \| \max(\sqrt{a}\boldsymbol{s}, \tau_{\text{s}})\|_2, \tag{8}$$

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 effectively aligns 3D Gaussians with FLAME. It ensures that the 3D Gaussians are positioned around the mesh triangle and can be semantically deformed. However, it also restricts animatble head Gaussian from modeling elements outside the space of FLAME in some cases, such as Thor's helmet and Kratos's long mustache, which are essential parts of their identities. On the other hand, as shown in Fig. 3, we observe that these elements are located on mesh triangles with large areas. This observation inspires us to introduce the area a as an adaptive factor:

$$\mathcal{L}_{\text{reg}} = (\lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{s}}\mathcal{L}_{\text{s}})/\sqrt{a}, \tag{9}$$

where $\lambda_{\text{pos}} = 0.1$ and $\lambda_{\text{s}} = 0.1$. Through regularization, the avatar demonstrates its ability for semantic deformation and modeling complex appearance.

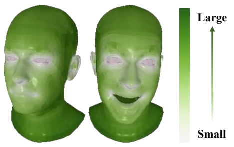

Fig. 3: Visualization of Mesh Area.

4.3 Implementation Details

Animatable Head Gaussian Details. In 3DGS, Kerbl *et al.* [35] 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 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 [53]. The densification and pruning iterations setting are following [43]. The FLAME's shape size is $|\boldsymbol{\gamma}| = 300$, the expression size is $|\boldsymbol{\psi}| = 100$ and the pose size is $|\boldsymbol{\gamma}| = 3 \times 4$ (neck, jaw, left eyeball and right eyeball).

Text to Avatar Optimization Details. We initialize animatable head Gaussian with $K = 10$ per triangle. Besides, we commonly set $s = 7.5$ and $s_{neg} = 1$ in animation-based text-to-3D distillation [34]. In our experiment, we default to using Realistic Vision 5.1 (RV5.1) [3] and ControlNetMediaPipeFace [1,73]. To alleviate the multi-face Janus problem, we also use the view-dependent prompts [30].

Training Details. The framework is implemented in PyTorch and threestudio [20]. 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 $[-30°, 30°]$,

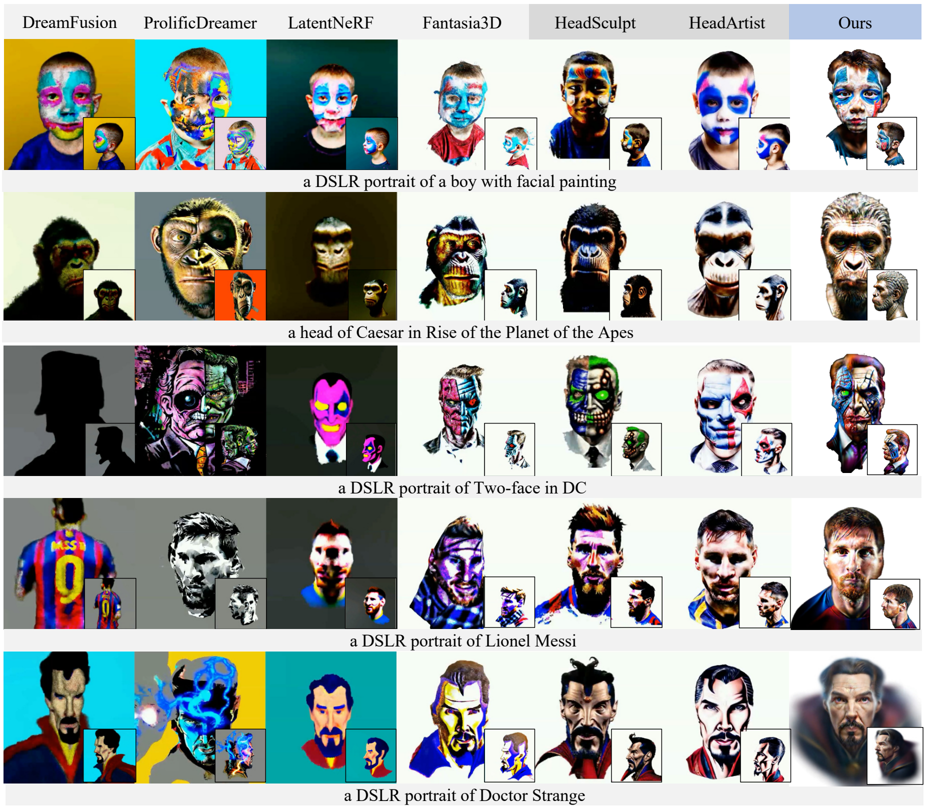
DreamFusion
ProlificDreamer
LatentNeRF
Fantasia3D
HeadSculpt
HeadArtist
Ours
a DSLR portrait of a boy with facial painting
a head of Caesar in Rise of the Planet of the Apes
a DSLR portrait of Two-face in DC
a DSLR portrait of Lionel Messi
a DSLR portrait of Doctor Strange

Fig. 4: Comparison with the text-to-static avatar generation methods. Our approach excels at producing high-fidelity head avatars, yielding superior results.

and an azimuth range of $[-180°, 180°]$. We train head avatars with a resolution of 1024 and a batch size of 8. The entire training consists of 10,000 iterations. The overall framework is trained using the Adam optimizer [36], 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 $\boldsymbol{\mu}$, scaling factor vs, rotation quaternion $\boldsymbol{q}$, color $\boldsymbol{c}$, opacity $\boldsymbol{\alpha}$, and FLAME shape $\boldsymbol{\beta}$, respectively [43]. Note that we stop the FLAME shape optimization after 8,000 iterations. 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 head avatars*: driving an avatar with FLAME sequences sampled in TalkSHOW [69].

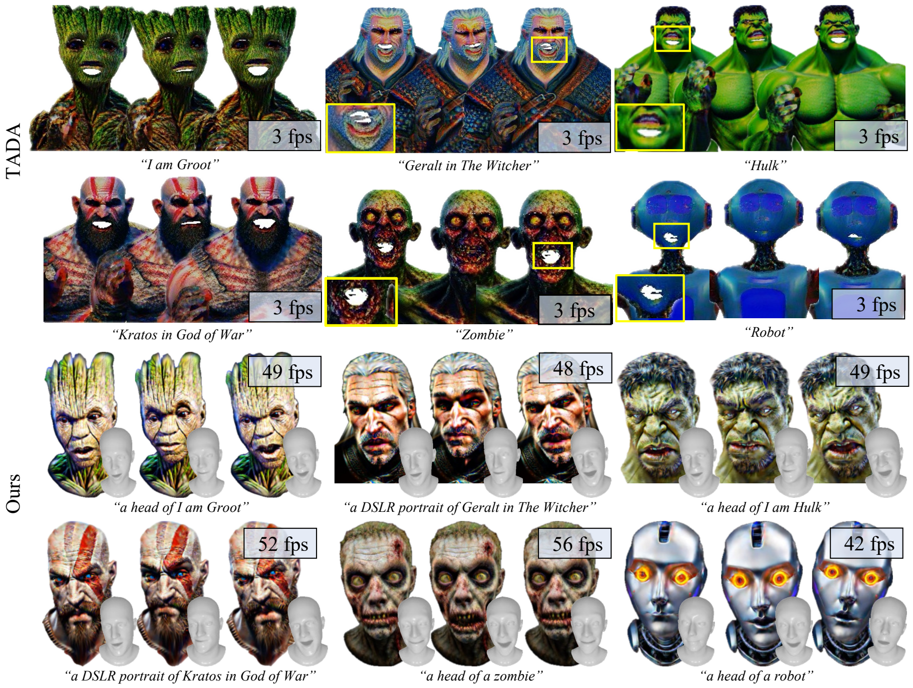

Fig. 5: Comparison with the text-to-dynamic avatar generation method TADA [41] in terms of semantic alignment and rendering speed. The yellow circles indicate semantic misalignment in the mouths, resulting in misplaced mouth texture. The rendering speed evaluation on the same device is reported in the 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.

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 [52]	0.244	0.302	0.300
LatentNeRF [48]	0.248	0.299	0.303
Fantasia3D [14]	0.267	0.304	0.300
ProlificDreamer [64]	0.268	0.320	0.308
HeadSculpt [21]	0.264	0.306	0.305
HeadArtist [42]	0.272	0.318	0.313
Ours	**0.275**	**0.322**	**0.317**

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 [52], LatentNeRF [48], Fantasia3D [14] and ProlificDreamer [64], HeadSculpt [21] and HeadArtist [42]. It is worth noting that HeadSculpt [21] and

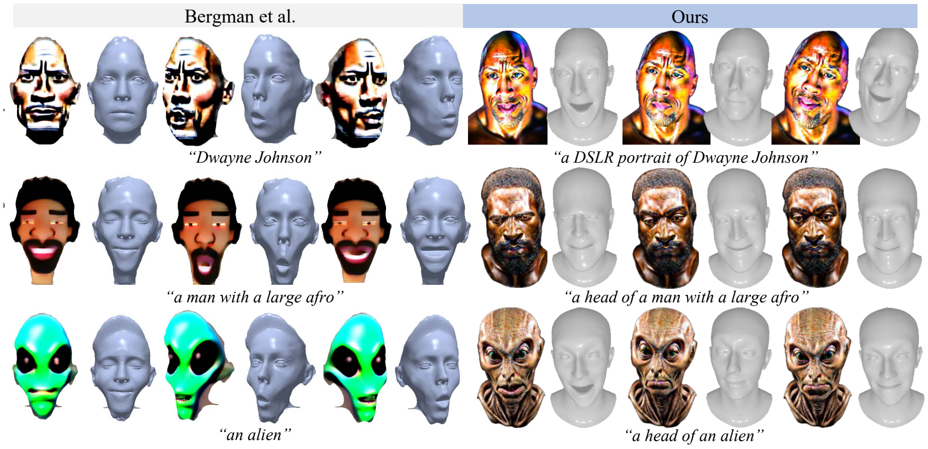
Bergman et al.
Ours
"Dwayne Johnson"
"a DSLR portrait of Dwayne Johnson"
"a man with a large afro"
"a head of a man with a large afro"
"an alien"
"a head of an alien"

Fig. 6: Comparison with the text-to-dynamic avatar generation method, Bergman *et al.* [6]. The FLAME mesh of the avatar is visualized on the bottom right. Our method demonstrates superior appearance and geometric modeling.

HeadArtist [42] specialize in text-to-static head avatar generation. 2) *dynamic head avatars*: We evaluate the efficacy of avatar animation by comparing it with TADA [41] and Bergman *et al.* [6]. Both approaches are based on FLAME and utilize it for animation.

5.1 Head Avatar Generation

Static Head Avatar Generation. We evaluate the avatar generation quality in terms of geometry and texture. In Fig. 4, 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 [14, 48, 52, 64]. This improvement can be attributed to the integration of FLAME, a reliable head structure prior, which mitigates the multi-face Janus problem [30] and enhances the geometry.

On the other hand, we evaluate the texture through quantitative experiments using the CLIP score [25]. 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.* [42], we report the average CLIP score of 10 text prompts. Tab. 1 demonstrates that HeadStudio outperforms other methods in three different CLIP variants [54]. Overall, HeadStudio excels at producing high-fidelity head avatars, outperforming the state-of-the-art text-based methods.

Dynamic Head Avatar Generation. 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 [69]. In Fig. 5, we compare HeadStudio with TADA [41].

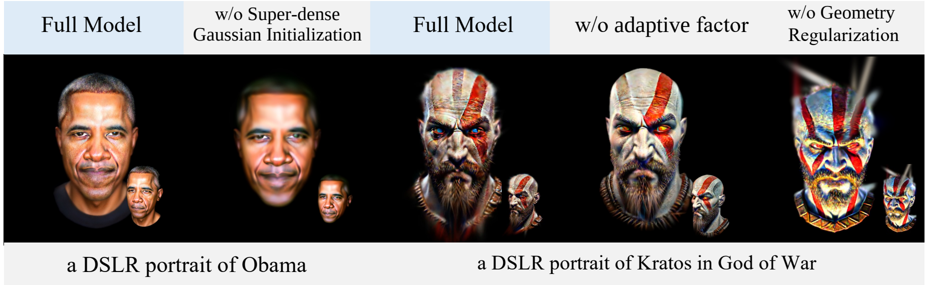
Full Model
w/o Super-dense Gaussian Initialization
Full Model
w/o adaptive factor
w/o Geometry Regularization
a DSLR portrait of Obama
a DSLR portrait of Kratos in God of War

Fig. 7: Ablation Study of Super-dense Gaussian Initialization and Adaptive Geometry Regularization. Super-dense Gaussian initialization enhances the representation ability. Geometry regularization imposes a strong restriction to reduce the outline points. The adaptive factor in geometry regularization balances restriction and expressiveness.

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 achieves 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 *vs.* 3 fps), our method demonstrates its potential in augmented or virtual reality applications. Furthermore, the comparison in Fig. 6 indicates the semantic alignment of the method proposed by Bergman *et al.* [6]. But it lacks in terms of its representation of appearance and geometry.

5.2 Ablation Study

We isolate the various contributions and conducted a series of experiments to assess their impact. In particular, we examine the design of super-dense Gaussian initialization, animation-based text-to-3D distillation, and adaptive geometry regularization. At last, we discuss the effect of different diffusion models.

Effect of Super-dense Gaussian Initialization. In Fig. 7, we present the effect of super-dense Gaussian initialization. Since the SDS supervision signal is sparse, super-dense Gaussian initialization enhances point coverage on the head model, leading to a favorable initialization and improved avatar fidelity.

Effect of Animation-based Text-to-3D Distillation. As illustrated in Fig. 8, we visualize the effect of each component in text to avatar optimization. Our method shows the improvements in the following three aspects: 1) Shape (a *vs.* c): FLAME offers precise control signals to address multi-face issues, ensuring ID consistency. 2) Texture (a *vs.* d): Denoised score distillation alleviates the over-smoothing problem in texture by eliminating unnecessary gradients. 3) Animation (a *vs.* b): Training with animations is crucial for artifact elimination (highlighted in yellow box) in deformation.

Effect of Adaptive Geometry Regularization. In Fig. 7, we also present the effect of adaptive geometry regularization. Firstly, adaptive geometry regulariza-

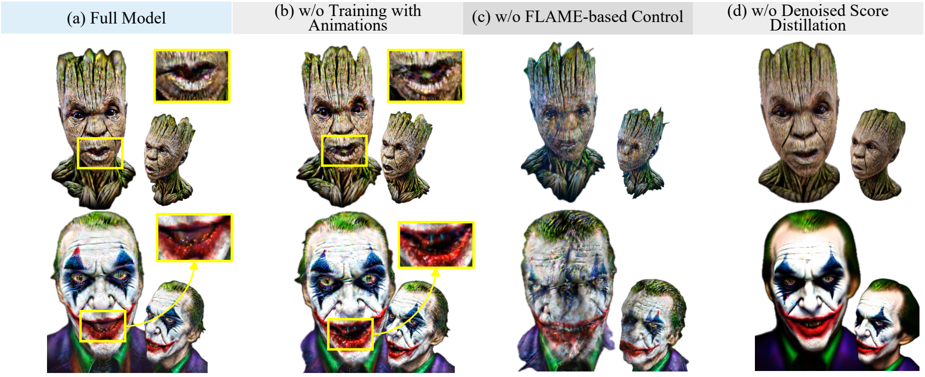
(a) Full Model
(b) w/o Training with Animations
(c) w/o FLAME-based Control
(d) w/o Denoised Score Distillation

Fig. 8: Ablation Study of Animation-based Text-to-3D Distillation. 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, animation-based text-to-3D distillation achieves an effective alignment, leading to an accurate expression deformation.

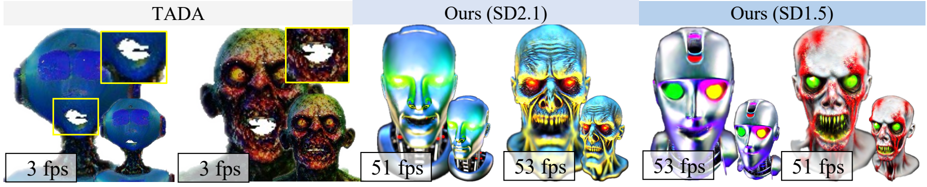

Fig. 9: Ablation Study of Different Diffusion Models. We investigate the effects of different diffusion models, including the Stable Diffusion v2.1 (SD2.1) and Stable Diffusion v1.5 (SD1.5).

tion could reduces the outline points. Nevertheless, overly strict regularization weaken the representation ability of animatable head Gaussian, such as the beard of Kratos (fourth column in Fig. 7). To address this, we introduce an adaptive 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 Different Diffusion Models. In this paper, we use Realistic Vision 5.1 (RV5.1) [3] as the default diffusion model. Compared to SD2.1 [55] and SD1.5, we observe that RV5.1 is capable of producing head avatars with a more visually appealing appearance. Meanwhile, we show the results of using SD2.1 (same as TADA [41]) and SD1.5 in Fig. 9. HeadStudio can generate avatars with better semantic alignment (texture alignment in mouths) and faster rendering speed (53 fps *vs*. 3 fps) compared with TADA [41].

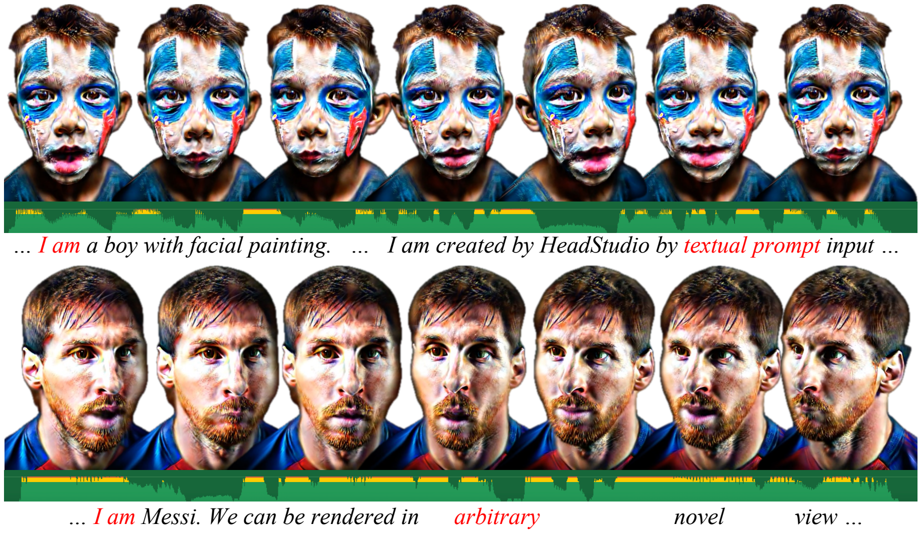
... I am a boy with facial painting. ... I am created by HeadStudio by textual prompt input ...
... I am Messi. We can be rendered in arbitrary novel view ...

Fig. 10: Application of HeadStudio. We expand our framework by employing TalkSHOW [69] to translate human speech to FLAME sequences. From bottom to top: the text input, the corresponding speech clip, and the animated head avatar.

5.3 Application of HeadStudio.

We further explore the applications of HeadStudio. **Audio-based animation** is a widely used technology in conference calls and virtual social presence. To realize it, we combine our framework with TalkSHOW [69] to translate human speech to FLAME sequences. **Text-based animation** can be used for creating talking head videos. We further expand the audio-based animation framework with a text-to-speech method PlayHT [2]. As shown in Fig. 10, the animation results are semantically aligned with the text input, showing its potential for real-world applications. We recommend the reader evaluate the performance through the supplementary videos.

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 arm the animatable head prior model with 3DGS for intricate texture and geometry modeling. Additionally, we enhance its optimization process from initialization, distillation, and regularization to simultaneously learn shape, texture, and animation, resulting in visually pleasing and high-quality animated avatars. Extensive evaluations demonstrated that our HeadStudio produces high-fidelity and animatble avatars with real-time rendering, outperforming state-of-the-art methods significantly.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2022ZD0160101, the National Natural Science Foundation of China (U2336212), the Fundamental Research Funds for the Central Universities (No. 226-2022-00051), the Fundamental Research Funds for the Central Universities (No. 226-2024-00058), the Fundamental Research Funds for the Zhejiang Provincial Universities (No. 226-2024-00208), the "Leading Goose" R&D Program of Zhejiang Province under Grant 2024C01101, and the China Postdoctoral Science Foundation (524000-X92302).

References

1. Controlnetmediapipeface, `https://huggingface.co/CrucibleAI/ControlNetMediaPipeFace`
2. Playht, `https://play.ht/`
3. Realistic vision 5.1, `https://huggingface.co/stablediffusionapi/realistic-vision-51`
4. An, S., Xu, H., Shi, Y., Song, G., Ogras, U.Y., Luo, L.: Panohead: Geometry-aware 3d full-head synthesis in 360°. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20950–20959 (June 2023)
5. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5470–5479 (2022)
6. Bergman, A.W., Yifan, W., Wetzstein, G.: Articulated 3d head avatar generation using text-to-image diffusion models. arXiv preprint arXiv:2307.04859 (2023)
7. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces. In: SIGGRAPH (1999). `https://doi.org/10.1145/311535.311556`
8. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18392–18402 (2023)
9. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. CVPR (2023)
10. Cao, Y., Cao, Y.P., Han, K., Shan, Y., Wong, K.Y.K.: Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. arXiv preprint arXiv:2304.00916 (2023)
11. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., Mello, S.D., Gallo, O., Guibas, L., Tremblay, J., Khamis, S., Karras, T., Wetzstein, G.: Efficient geometry-aware 3D generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
12. Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: 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. pp. 5799–5809 (2021)
13. Chen, G., Wang, W.: A survey on 3d gaussian splatting. arXiv preprint arXiv:2401.03890 (2024)

14. Chen, R., Chen, Y., Jiao, N., Jia, K.: 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) (October 2023)
15. Cohen-Bar, D., Richardson, E., Metzer, G., Giryes, R., Cohen-Or, D.: Set-the-scene: Global-local training for generating controllable nerf scenes. arXiv preprint arXiv:2303.13450 (2023)
16. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3D face model from in-the-wild images. ACM Transactions on Graphics, (Proc. SIGGRAPH) **40**(8) (2021), https://doi.org/10.1145/3450626.3459936
17. Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-planes: Explicit radiance fields in space, time, and appearance. In: CVPR (2023)
18. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
19. Gao, C., Saraf, A., Kopf, J., Huang, J.B.: Dynamic view synthesis from dynamic monocular video. In: Proceedings of the IEEE International Conference on Computer Vision (2021)
20. Guo, Y.C., Liu, Y.T., Shao, R., Laforte, C., Voleti, V., Luo, G., Chen, C.H., Zou, Z.X., Wang, C., Cao, Y.P., Zhang, S.H.: threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio (2023)
21. Han, X., Cao, Y., Han, K., Zhu, X., Deng, J., Song, Y.Z., Xiang, T., Wong, K.Y.K.: Headsculpt: Crafting 3d head avatars with text. arXiv preprint arXiv:2306.03038 (2023)
22. Haque, A., Tancik, M., Efros, A., Holynski, A., Kanazawa, A.: Instruct-nerf2nerf: Editing 3d scenes with instructions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
23. He, S., He, H., Yang, S., Wu, X., Xia, P., Yin, B., Liu, C., Dai, L., Xu, C.: Speech4mesh: Speech-assisted monocular 3d facial reconstruction for speech-driven 3d facial animation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14192–14202 (2023)
24. Hertz, A., Aberman, K., Cohen-Or, D.: Delta denoising score. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2328–2337 (2023)
25. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
26. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
27. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems (NeurIPS) **33**, 6840–6851 (2020)
28. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
29. Höllein, L., Cao, A., Owens, A., Johnson, J., Nießner, M.: Text2room: Extracting textured 3d meshes from 2d text-to-image models. arXiv preprint arXiv:2303.11989 (2023)
30. Hong, S., Ahn, D., Kim, S.: Debiasing scores and prompts of 2d diffusion for robust text-to-3d generation. arXiv preprint arXiv:2303.15413 (2023)
31. Jain, A., Mildenhall, B., Barron, J.T., Abbeel, P., Poole, B.: Zero-shot text-guided object generation with dream fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

32. Jiang, R., Wang, C., Zhang, J., Chai, M., He, M., Chen, D., Liao, J.: Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control. arXiv preprint arXiv:2303.17606 (2023)
33. Kamata, H., Sakuma, Y., Hayakawa, A., Ishii, M., Narihira, T.: Instruct 3d-to-3d: Text instruction guided 3d-to-3d conversion. arXiv preprint arXiv:2303.15780 (2023)
34. Katzir, O., Patashnik, O., Cohen-Or, D., Lischinski, D.: Noise-free score distillation. arXiv preprint arXiv:2310.17590 (2023)
35. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 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/`
36. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
37. Kirschstein, T., Giebenhain, S., Nießner, M.: Diffusionavatars: Deferred diffusion for high-fidelity 3d head avatars. arXiv preprint arXiv:2311.18635 (2023)
38. Li, C., Zhang, C., Waghwase, A., Lee, L.H., Rameau, F., Yang, Y., Bae, S.H., Hong, C.S.: Generative ai meets 3d: A survey on text-to-3d in aigc era. arXiv preprint arXiv:2305.06131 (2023)
39. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph. **36**(6), 194–1 (2017)
40. Liang, C., Ma, F., Zhu, L., Deng, Y., Yang, Y.: Caphuman: Capture your moments in parallel universes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6400–6409 (2024)
41. Liao, T., Yi, H., Xiu, Y., Tang, J., Huang, Y., Thies, J., Black, M.J.: Tada! text to animatable digital avatars. arXiv preprint arXiv:2308.10899 (2023)
42. Liu, H., Wang, X., Wan, Z., Shen, Y., Song, Y., Liao, J., Chen, Q.: Headartist: Text-conditioned 3d head generation with self score distillation. arXiv preprint arXiv:2312.07539 (2023)
43. Liu, X., Zhan, X., Tang, J., Shan, Y., Zeng, G., Lin, D., Liu, X., Liu, Z.: Humangaussian: Text-driven 3d human generation with gaussian splatting. arXiv preprint arXiv:2311.17061 (2023)
44. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM Trans. Graph. **34**(6), 248:1–248:16 (Oct 2015)
45. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., et al.: Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019)
46. Ma, F., Jin, X., Wang, H., Xian, Y., Feng, J., Yang, Y.: Vista-llama: Reliable video narrator via equal distance to visual tokens (2023)
47. Ma, Y., Lin, Z., Ji, J., Fan, Y., Sun, X., Ji, R.: X-oscar: A progressive framework for high-quality text-guided 3d animatable avatar generation. arXiv preprint arXiv:2405.00954 (2024)
48. Metzer, G., Richardson, E., Patashnik, O., Giryes, R., Cohen-Or, D.: Latent-nerf for shape-guided generation of 3d shapes and textures. arXiv preprint arXiv:2211.07600 (2022)
49. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
50. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)

51. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., 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). pp. 10975–10985 (Jun 2019), `http://smpl-x.is.tue.mpg.de`
52. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
53. Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. arXiv preprint arXiv:2312.02069 (2023)
54. 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: Proceedings of the International Conference on Machine Learning (ICML). pp. 8748–8763 (2021)
55. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022)
56. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arXiv preprint arxiv:2208.12242 (2022)
57. Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A.: Graf: Generative radiance fields for 3d-aware image synthesis. Advances in Neural Information Processing Systems **33**, 20154–20166 (2020)
58. Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. Advances in Neural Information Processing Systems **34**, 6087–6101 (2021)
59. Shen, X., Ma, J., Zhou, C., Yang, Z.: Controllable 3d face generation with conditional style code diffusion. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4811–4819 (2024)
60. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pp. 2256–2265. PMLR (2015)
61. Voynov, A., Aberman, K., Cohen-Or, D.: Sketch-guided text-to-image diffusion models. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
62. Wang, H., Du, X., Li, J., Yeh, R.A., 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 (CVPR) (2022)
63. Wang, T., Zhang, B., Zhang, T., Gu, S., Bao, J., Baltrusaitis, T., Shen, J., Chen, D., Wen, F., Chen, Q., et al.: Rodin: A generative model for sculpting 3d digital avatars using diffusion. arXiv preprint arXiv:2212.06135 (2022)
64. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. arXiv preprint arXiv:2305.16213 (2023)
65. Wu, Y., Xu, H., Tang, X., Chen, X., Tang, S., Zhang, Z., Li, C., Jin, X.: Portrait3d: Text-guided high-quality 3d portrait generation using pyramid representation and gans prior. ACM Trans. Graph. **43**(4) (Jul 2024). `https://doi.org/10.1145/3658162`, `https://doi.org/10.1145/3658162`
66. Xu, Y., Yang, Z., Yang, Y.: Seeavatar: Photorealistic text-to-3d avatar generation with constrained geometry and appearance. arXiv preprint arXiv:2312.08889 (2023)

67. Xu, Y., Wang, L., Zhao, X., Zhang, H., Liu, Y.: Avatarmav: Fast 3d head avatar reconstruction using motion-aware neural voxels. In: ACM SIGGRAPH 2023 Conference Proceedings (2023)
68. Yang, Z., Chen, G., Li, X., Wang, W., Yang, Y.: Doraemongpt: Toward understanding dynamic scenes with large language models (exemplified as a video agent). In: ICML (2024)
69. Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., Black, M.J.: Generating holistic 3d human motion from speech. In: CVPR (2023)
70. Zhang, C., Zhang, C., Zhang, M., Kweon, I.S.: Text-to-image diffusion model in generative ai: A survey. arXiv preprint arXiv:2303.07909 (2023)
71. Zhang, J., Zhang, X., Zhang, H., Liew, J.H., Zhang, C., Yang, Y., Feng, J.: Avatarstudio: High-fidelity and animatable 3d avatar creation from text. arXiv preprint arXiv:2311.17917 (2023)
72. Zhang, L., Qiu, Q., Lin, H., Zhang, Q., Shi, C., Yang, W., Shi, Y., Yang, S., Xu, L., Yu, J.: Dreamface: Progressive generation of animatable 3d faces under text guidance. arXiv preprint arXiv:2304.03117 (2023)
73. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)
74. Zhang, X., Zheng, Z., Gao, D., Zhang, B., Yang, Y., Chua, T.S.: Multi-view consistent generative adversarial networks for compositional 3d-aware image synthesis. International Journal of Computer Vision **131**(8), 2219–2242 (2023)
75. Zhang, Y., Fan, H., Yang, Y.: Prompt-aware adapter: Towards learning adaptive visual tokens for multimodal large language models. arXiv preprint arXiv:2405.15684 (2024)
76. Zheng, Y., Abrevaya, V.F., Bühler, M.C., Chen, X., Black, M.J., Hilliges, O.: I M Avatar: Implicit morphable head avatars from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
77. Zheng, Y., Yifan, W., Wetzstein, G., Black, M.J., Hilliges, O.: Pointavatar: Deformable point-based head avatars from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
78. Zhou, D., Li, Y., Ma, F., Zhang, X., Yang, Y.: Migc: Multi-instance generation controller for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6818–6828 (2024)
79. Zhuo, W., Ma, F., Fan, H., Yang, Y.: Vividdreamer: Invariant score distillation for hyper-realistic text-to-3d generation. In: ECCV (2024)
80. Zielonka, W., Bolkart, T., Thies, J.: Towards metrical reconstruction of human faces. In: European Conference on Computer Vision (2022)
81. Zielonka, W., Bolkart, T., Thies, J.: Instant volumetric head avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)