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

Supplementary Material

A Additional Implementation Details

A.1 Text to Animatable Avatar Optimization

For each text prompt, we first initialize an animatable head Gaussian via the super-dense Gaussian initialization. Each iteration of HeadStudio performs the following: (1) randomly sample a camera and animation inputs (pose and expression); (2) drive the animatable head Gaussian with the given pose and expression and render an image from that camera; (3) compute the gradients of the animation-based text-to-3D distillation; (4) compute the loss of the adaptive geometry regularization; At the end of an iteration, we update the animatable head Gaussian parameters using an optimizer.

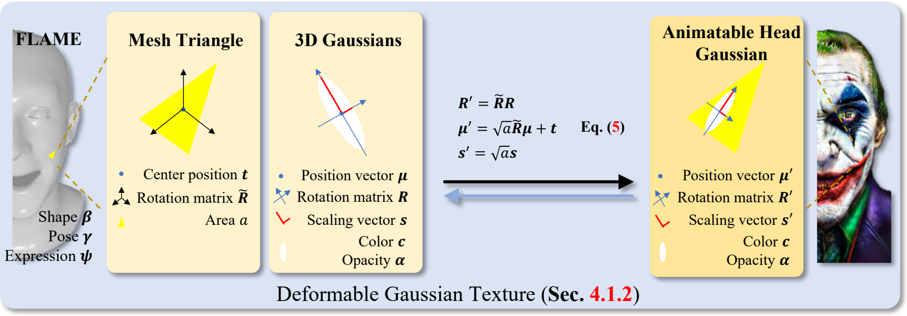

FLAME
Shape $\boldsymbol{\beta}$
Pose $\boldsymbol{\gamma}$
Expression $\boldsymbol{\psi}$
Mesh Triangle
Center position $\boldsymbol{t}$
Rotation matrix $\widetilde{\boldsymbol{R}}$
Area a
3D Gaussians
Position vector $\boldsymbol{\mu}$
Rotation matrix $\boldsymbol{R}$
Scaling vector $\boldsymbol{s}$
Color $\boldsymbol{c}$
Opacity $\boldsymbol{\alpha}$
$\boldsymbol{R}' = \widetilde{\boldsymbol{R}}\boldsymbol{R}$
$\boldsymbol{\mu}' = \sqrt{a}\widetilde{\boldsymbol{R}}\boldsymbol{\mu} + \boldsymbol{t}$
$\boldsymbol{s}' = \sqrt{a}\boldsymbol{s}$
Eq. (5)
Animatable Head Gaussian
Position vector $\boldsymbol{\mu}'$
Rotation matrix $\boldsymbol{R}'$
Scaling vector $\boldsymbol{s}'$
Color $\boldsymbol{c}$
Opacity $\boldsymbol{\alpha}$
Deformable Gaussian Texture (Sec. 4.1.2)

Fig. 11: The Details of Deformable Gaussian Texture. Animatable head Gaussian uses the mesh triangle's center position, rotation matrix and area to translate, rotate and scale the corresponding rigged 3D Gaussians, resulting in a deformed 3D Gaussians.

0. Initialization. We evenly sample $K = 10$ points per triangle from FLAME with the standard pose, and initialize the scaling via the square root of the mean distance of K-nearest neighbor points. The 3D Gaussians rigged with a large mesh triangle are initialized with a larger radius, compared to the ones rigged with a small mesh. As a result, it initializes 3D Gaussians that can thoroughly cover the head model. The further discussion of K selection can be found in Sec. B.2.

1. Random camera and animation sampling. At each iteration, the animation inputs, pose and expression are sampled from the FLAME sequences

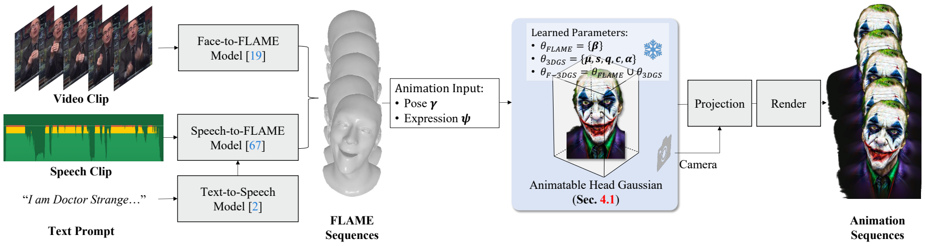


Fig. 12: The Pipeline of HeadStudio's Application. The head avatar (fixed animatable head Gaussian) can be driven by video, speech, and text using FLAME pose and expression as control.

(pre-calculated based on the real-world talk show videos [71]). Meanwhile, a camera position is randomly sampled as described in Sec. 4.3.3.

2. Deform and render animatable head Gaussian. We detail the deformation process in Fig. 11. Given the pose and expression, FLAME with learnable shape is driven according to Eq. (4), deforming the mesh triangles. Then, we utilize the mesh triangle's center position, rotation matrix and area to translate, rotate and scale the corresponding rigged 3D Gaussians (Eq. (5)). Following this, we render the deformed 3D Gaussians at a resolution of 1024 × 1024 based on the sampled camera pose.

3. Optimization with animation-based text-to-3D distillation. Based on the FLAME model, we initially draw a facial landmark map in MediaPipe format as the diffusion condition. Then, we calculate the gradients of Eq. (6) w.r.t. the animatable head Gaussian parameters, which force the rendering to satisfy the text prompt in any pose, expression, and camera view.

4. Optimization with geometry regularization. We constrain the position and radius of 3D Gaussians w.r.t. the size of their rigged mesh triangle according to the Eq. (8). Furthermore, an adaptive scaling factor is introduced in Eq. (9) for modeling elements outside the space of FLAME. The impact of the regularization is discussed in Sec. B.2.

A.2 The Pipeline of HeadStudio's Application

We present the pipeline of HeadStudio's application in Fig. 12. Once optimized, the parameters of the avatar remain fixed. Given a pose and expression, it can be deformed and rendered in a novel view. Combined with advanced techniques, such as face-to-FLAME model [16], speech-to-FLAME model [71] and text-to-speech model [2], the video, speech and text can be converted into FLAME animation inputs. HeadStudio processes the input frame by frame and produces the animation sequences, which can then be merged into a video. Consequently, HeadStudio can be driven by multi-modality and achieves real-world applications (as shown in supplementary videos).

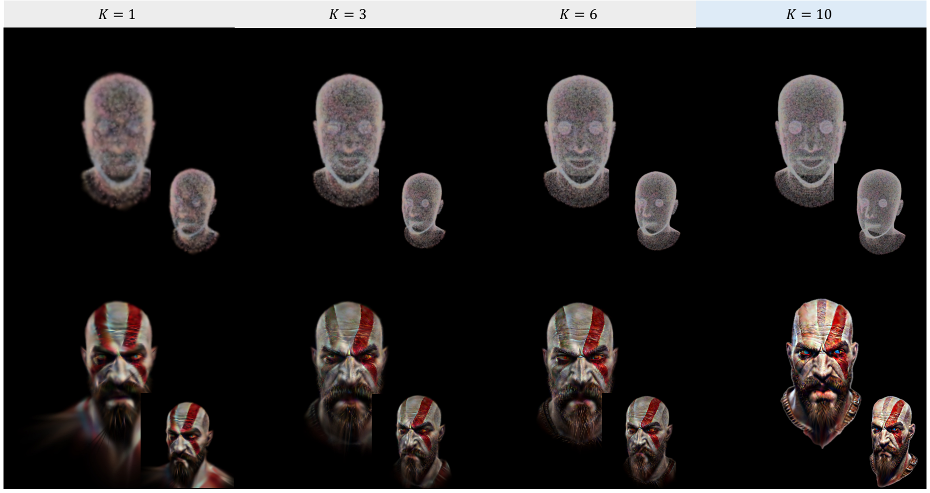


Fig. 13: Evaluation on K in super-dense Gaussian initialization. The cloning and splitting strategy can not handle the generation well. Increasing K improves generation results with dense initialization.

B Additional Experiments

B.1 Temporal Stable Diffusion

Temporal stable diffusion, such as AniPortrait [66], introduces motion module into the denoising UNet. As a result, it can generate a video clip with temporal consistency. It inspires us to utilize a temporal stable diffusion to improve the temporal smoothness (skin wobbles) and animation quality (never blinking). As shown in Fig. 14, the temporal information is indeed significant for generating smoother animations, and we will consider incorporating more temporal designs to enhance temporal supervision in the future.

B.2 Additional Ablations

Evaluation on different K in super-dense Gaussian initialization. We discuss the impact of the hyperparameter K in HeadStudio. As shown in Fig. 13, the proposed initialization is essential for generation. In the default configuration ($K = 1$), the animatable head Gaussian is unable to grow up through cloning and splitting [35], leading to a poor appearance. We attribute it to the sparse guidance provided by score distillation-based loss. On the other hand, the density of 3D Gaussians is similar to the resolution of the image. A denser 3D Gaussians will have a better representation ability. Therefore, with the increase of K, the dense initialization results in better generation results. However, a large K will result in additional time and memory costs. Therefore, we opt for $K = 10$ as the default experimental setup.

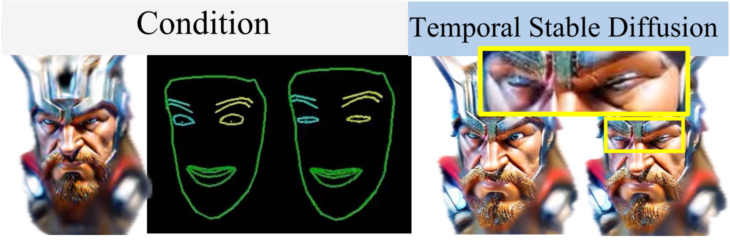


Fig. 14: Evaluation on temporal stable diffusion. The temporal information is important to improve the temporal smoothness (skin wobbles) and animation quality (never blinking).

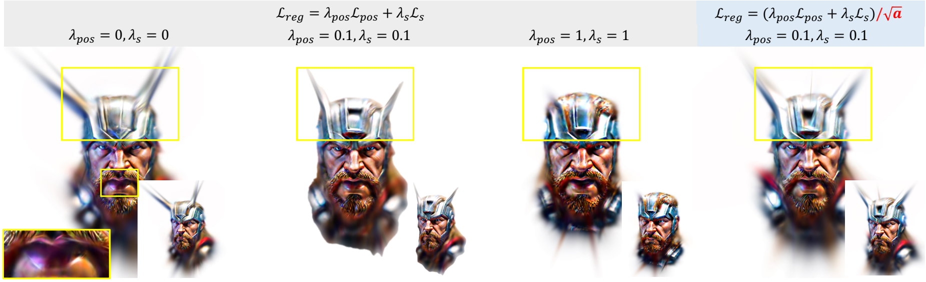

$\lambda_{pos} = 0, \lambda_s = 0$
$\mathcal{L}_{reg} = \lambda_{pos}\mathcal{L}_{pos} + \lambda_s\mathcal{L}_s$
$\lambda_{pos} = 0.1, \lambda_s = 0.1$
$\lambda_{pos} = 1, \lambda_s = 1$
$\mathcal{L}_{reg} = (\lambda_{pos}\mathcal{L}_{pos} + \lambda_s\mathcal{L}_s)/\sqrt{a}$
$\lambda_{pos} = 0.1, \lambda_s = 0.1$

Fig. 15: Evaluation on adaptive geometry regularization. Regularization is essential for semantic deformation. But the weight of regularization must find a good balance between alignment and representation. Including an adaptive scaling factor helps to combine semantic alignment and adequate representation well.

Evaluation on Adaptive Geometry Regularization. First, we investigate geometry regularization and explore the impact of its weight in HeadStudio. As depicted in Fig. 15, geometry regularization is crucial for semantic deformation. In the absence of geometry regularization ($\lambda_{pos} = 0, \lambda_s = 0$), the 3D Gaussians fail to align semantically with FLAME, resulting in the problem of mouths sticking together (first column in Fig. 15). On the other hand, the weight shows a trade-off between alignment and representation. For instance, the Thor in the third column, generated with a large constraint weight, shows good alignment in the mouth but lacks representation (the helmet is missing). Then, we analyze the proposed adaptive scaling factor. We choose the area of the mesh triangle as an adaptive scaling factor (shown in Fig. 16), which is small around the eyes and mouth, and large on jaw and over head. With the help of the adaptive scaling factor, the generation demonstrates semantic alignment and adequate representation simultaneously (fourth column in Fig. 15). It highlights the importance of the adaptive scaling factor in geometry regularization, which effectively balances the alignment and representation.

Evaluation on Animal Character. We evaluate the generalization of HeadStudio with various animal character prompts. As shown in Fig. 18 and Fig. 17,

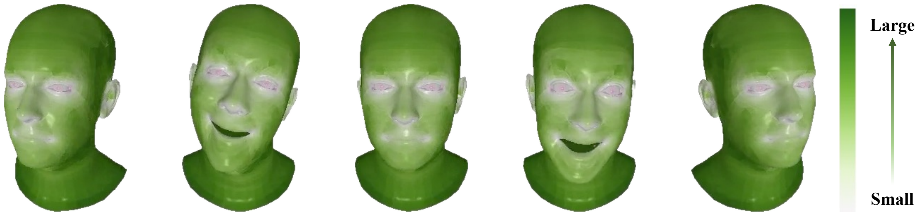
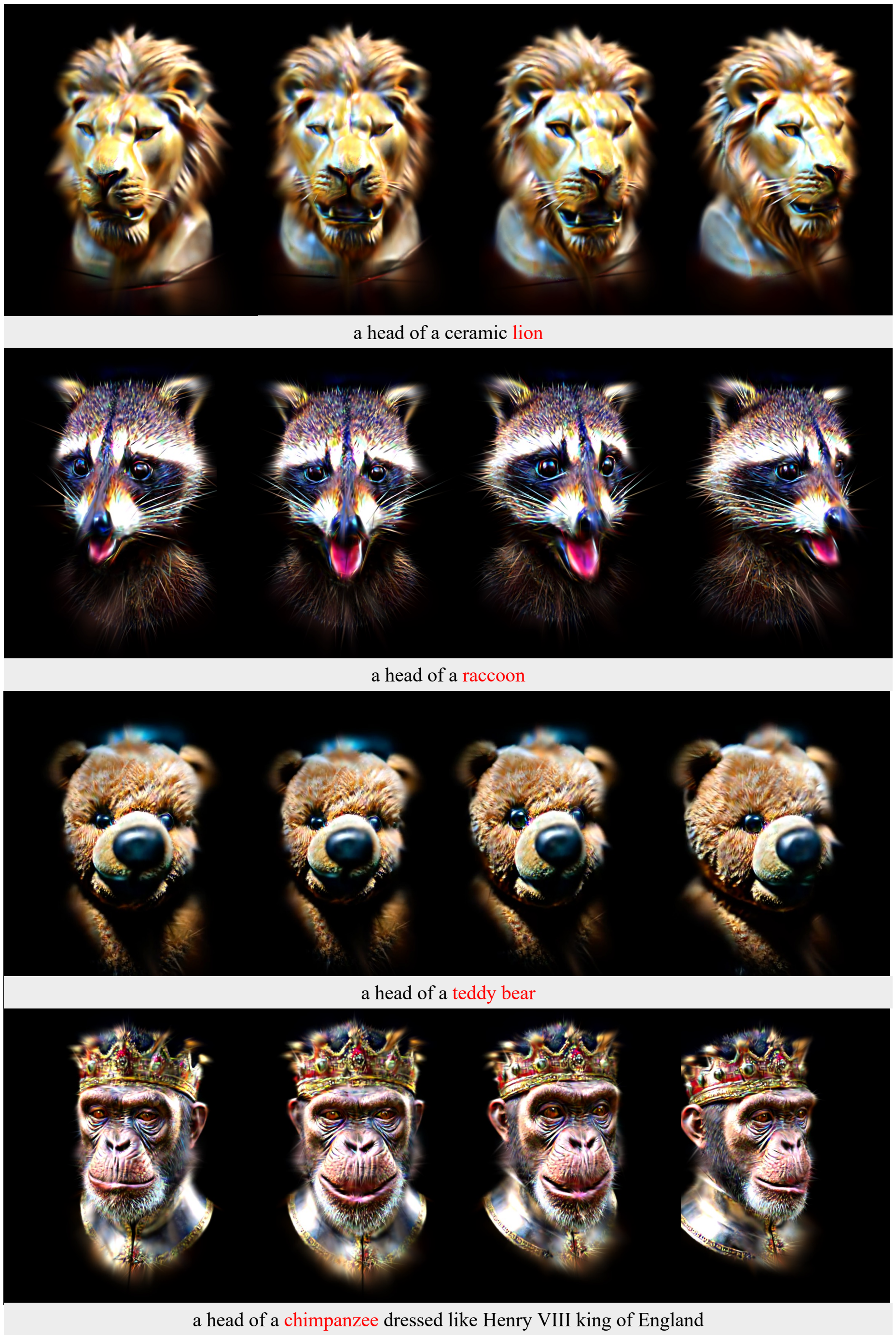


Fig. 16: More Visualization of Mesh Area. We visualize the area of mesh triangle, where small mesh is white and large mesh is green. The mesh around the eyes, noise, mouth and ears is small, while the mesh on the jaw and above the head is relatively larger.

HeadStudio effectively generates animal characters, such as the lion, corgi, bear, raccoon and chimpanzee. However, we believe that the human head prior model, FLAME [39], could limit the animation quality. In the future, replacing FLAME with an animal prior model like SMAL [84] in HeadStudio could improve animal avatar generation.

C Limitations

HeadStudio can create animatable head avatars from text for easier avatar production. However, certain challenges need to be addressed before using avatars in applications. For instance, it is important to develop a real-time driving and presentation system to integrate avatars into live broadcasts, which suited to 3DGS rendering pipeline. For instance, to enable an avatar for live broadcasting, a real-time driving and presentation system suitable for 3DGS rendering should be developed. Engineering issues such as complex workflows and audio-visual synthesis need to be carefully addressed. Additionally, our method faces some limitations inherited from FLAME, particularly in representing teeth and hair. Recent advancements in the teeth [54] and hair modeling [46] could offer solutions to these limitations.



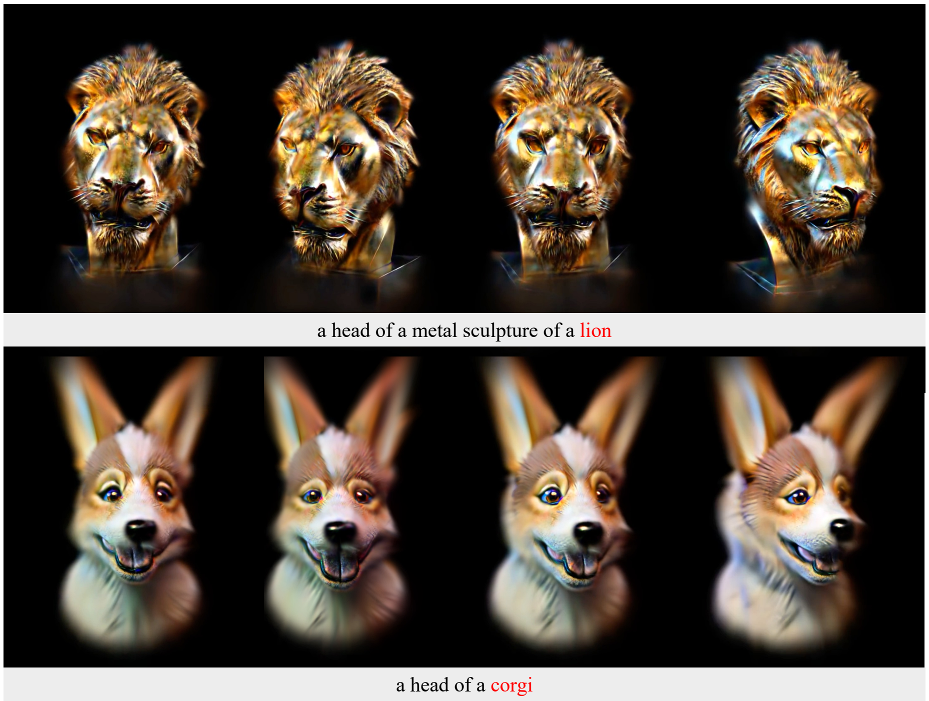
a head of a ceramic lion

a head of a raccoon

a head of a teddy bear

a head of a chimpanzee dressed like Henry VIII king of England

Fig. 17: Evaluation on Animal Character.



a head of a metal sculpture of a lion

a head of a corgi

Fig. 18: Evaluation on Animal Character. HeadStudio effectively generates animal characters, showing its versatility.