Response to Eurographics Reviewers

We would like to thank the reviewers for their constructive comments and suggestions. We appreciate the positive comments from all the reviewers and have carefully revised the paper to address the raised issues. Below we respond to each reviewer's concerns one by one and incorporate the changes in the main paper.

Reviewer #1

- **The paper only compares with methods from 2023, while there are many recent relevant papers such as Portrait3D (TOG 2024).**

  A: Our method focuses on generating animatable 3D Gaussian head models. In contrast, Portrait3D generates 3D head models from text prompts, but these models are static and do not support animation. We have added more discussion about the recent relevant papers in the related works.

- **Generation in UV space is easier than in 3D space, but it may introduce artifacts around the UV cutting areas (seams). Since this method only shows frontal views with small variations, I would like to see the quality of the back side.**

  A: In the training, we use the FFHQ dataset, which contains only front-view faces. Since expression parameters and head poses are both sampled from the dataset during training, the model can only guarantee the generation quality of frontal faces.

- **I recommend that the authors explore the difference between their proposed coordinates and the construction of the standard tangent space. If they are the same, this part could be reduced.**

  A: Following the reviewer's suggestion, we have shortened and streamlined the description of this part, removing redundant details and emphasizing its key properties: continuity across faces and tight coupling with mesh normals and global head pose, while clarifying its practical role in animation and generation.

- **I would like to know the decoupling quality of the static and dynamic parts. Specifically, what information is identified by the network as "dynamic"? It would be better to visualize these parts to improve the technical soundness of this section.**

  A: To show the effectiveness of the dynamic module, we add the qualitative comparison for the ablation study. Fig. A1 shows that the detailed expressions, such as blinking, are more accurately depicted by the dynamic module.

Reviewer #2

- **The dynamic module, which appears to be a core contribution, is under-described, which makes it difficult to judge the contributions soundness.**

  A: To show the effectiveness of the dynamic module, we add the qualitative comparison for the ablation study. Fig. A1 shows that the detailed expressions, such as blinking, are more accurately depicted by the dynamic module.

Figure A1: Qualitative comparison for the ablation study. The component $\mathcal{G}_m$ enhances the quality of the generated mouth, while $\mathcal{G}_d$ improves dynamic clarity and expression details.

- **Most of everything else is very similar to prior work "GaussianAvatars: Photorealistic Head Avatars with Rigged 3D Gaussians" and thus presumed to be technical sound.**

  A: Our method is a generative model that can synthesize animatable 3D Gaussian head avatars. In contrast, GaussianAvatars performs per-subject reconstruction from multi-view capture data, where each identity is captured by 16 camera views and requires per-identity optimization. Our approach adopts a discriminative learning paradigm to learn generalizable dynamic face generation from a large-scale image dataset, rather than optimizing for a specific subject.

Reviewer #3

- **In Table 1(b) (Runtime), the unit of the reported runtime values is unclear and should be explicitly specified.**

  A: Table 1(b) reports the average frames synthesized per second(FPS). The caption of Table 1(b) has been updated to Runtime(FPS).

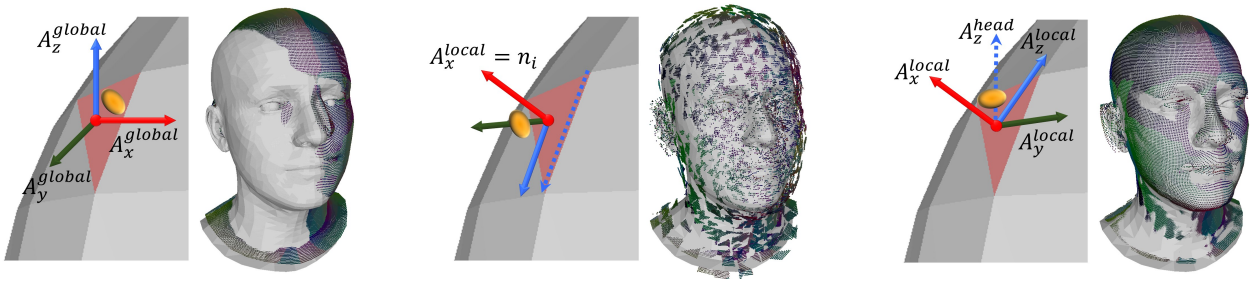
- **The training dataset of the proposed model is not clearly specified. The paper only mentions "25M images".**

  A: We conduct training on the FFHQ dataset (70k images) and follow the EG3D setting to train from scratch for 25M image iterations.

- **The paper should provide more detailed explanations of the global offset and naive local coordinate strategies used in the ablation study.**

  A: The definitions of different coordinate systems are shown in Fig.A2 (Fig.3 in draft). As shown in Fig.A2(a) is to directly adopt the global coordinate [XGGZ24] as the local coordinate, this approach leads to misplacements when expressions or head poses change. The naive local coordinate(Fig.A2(b)) leverages each triangle's normal and one of its edges to define two perpendicular axes.

Reviewer #4



(a) Global Coordinate (b) Naive Local Coord. (c) Ours

Figure A2: Different definitions of local coordinate system.

- **some terms are used in equations before they are defined/introduced (e.g. all the symbols when defining the 14-dim feature $f_{u,v}$).**

  A: Each pixel defined in 2D UV feature map is represented as $f_{u,v} = \{o_i, s_i, q_i, c_i, \sigma_i\}$, where $o \in \mathbb{R}^3$ denotes the center offset, $s \in \mathbb{R}^3$ the sacle, $q \in \mathbb{R}^4$ the rotation parameterized as quaternion , $c \in \mathbb{R}^3$ the color, and $\sigma \in \mathbb{R}$ the opacity. The above details have been added into Sec. 2.1.1 in the main paper.

- **The evaluation section also could be clearer with an explicit description of how AED/APD are computed, especially the parameter-reconstruction part.**

  A: For the Average Expression Distance (AED) and the Average Pose Distance (APD), we first leverage a pre-trained face reconstruction network to recover the corresponding FLAME coefficients from the generated results. We then compute the average distance metrics (AED/APD) by measuring the mean deviation between the driving FLAME parameters and these reconstructed FLAME coefficients.

References

[XGGZ24] XIANG J., GAO X., GUO Y., ZHANG J.: Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 1802–1812.