CVPR
#3250

GazeGaussian: High-Fidelity Gaze Redirection with 3D Gaussian Splatting

Supplementary Material

A. Overview

The supplementary material encompasses the subsequent components. Please visit the anonymous website https://gazegaussian.github.io/ for additional visual comparisons of novel view and novel gaze synthesis.

- Supplementary experiments
  - Ablation study on cross-dataset
  - Personal calibration for gaze estimation
  - Comparison between GazeNeRF + expression-guided neural renderer and GazeGaussian
- Additional visualization results
  - Visualization for ablation study
  - Visualization for cross-dataset comparison
  - Visualization for identity morphing
  - Visualization for transformed Gaussians
- Implementation details
- Dataset and pre-processing details
- Ethical consideration and limitations

B. Supplementary experiments

B.1. Ablation study on cross-dataset

To further validate the effectiveness of each proposed component, we conduct an ablation study on the cross-dataset evaluation to assess the generalization capability of our full pipeline. As shown in Tab. 1, the results are consistent with the ablation study in the main text. The proposed Gaussian eye rotation representation significantly improves eye redirection accuracy while ensuring robust redirection across cross-domain datasets. Additionally, the expression-guided neural renderer enhances the fidelity of the synthesized images, preserving the identity characteristics of the input image. From the ablation study on cross-dataset, we can further validate the importance of each component.

B.2. Personal calibration for gaze estimation

Following GazeNeRF, we perform personal calibration to demonstrate the benefits of our method for downstream gaze estimation tasks. Specifically, given a few calibration samples from person-specific test sets, we augment these real samples with gaze-redirected samples generated by GazeGaussian. We then fine-tune a gaze estimator pre-trained on ETH-XGaze using these augmented samples and compare its performance with a baseline model fine-tuned only on real samples. To ensure a fair comparison, the total number of augmented samples is fixed at 200 (real + generated samples), and we vary the number of real samples used for fine-tuning during the evaluation.

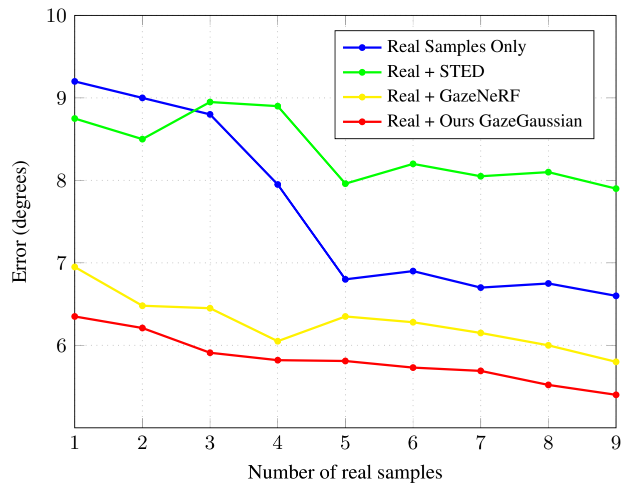

Figure 1. Error comparison based on number of real samples.

As shown in Fig. 1, the x-axis represents the number of real samples used, and the y-axis shows the gaze estimation error in degrees on the ETH-XGaze person-specific test set. We evaluate up to nine real samples in the few-shot setting. Fine-tuning the pre-trained gaze estimator with real and generated samples from GazeGaussian achieves the lowest gaze estimation error across all settings. Compared to GazeNeRF, GazeGaussian demonstrates a clear advantage, especially when fewer real samples are available, indicating that the generated samples from GazeGaussian are of higher fidelity and more effective for improving downstream gaze estimation accuracy. In contrast, samples generated by GazeNeRF lead to higher errors, while STED performs the worst, showing a notable limitation in leveraging 2D generative models for this task. This is due to the lack of consideration for the 3D nature of gaze redirection in STED, which is critical for high-quality sample generation and effective downstream adaptation.

B.3. Comparison between GazeNeRF + expression-guided neural renderer and GazeGaussian

We compare the performance of GazeNeRF, GazeNeRF enhanced with the expression-guided neural renderer (EGNR), and our proposed GazeGaussian on the ETH-XGaze dataset. As shown in Tab. 2, integrating EGNR into GazeNeRF leads to noticeable improvements in gaze redirection accuracy and image quality. This demonstrates the versatility of the proposed expression-guided neural renderer in enhancing facial synthesis and better capturing identity-specific expressions. However, even with the added EGNR, GazeNeRF's performance remains limited compared to GazeGaussian.

Table 1. Component-wise ablation study of GazeGaussian on the ColumbiaGaze, MPIIFaceGaze and GazeCapture datasets.

Two-stream	Gaus. Eye Rep.	Exp. Guided	ColumbiaGaze				MPIIFaceGaze				GazeCapture			
			Gaze↓	Head↓	LPIPS↓	ID↑	Gaze↓	Head↓	LPIPS↓	ID↑	Gaze↓	Head↓	LPIPS↓	ID↑
✓			8.996	4.494	0.325	49.286	19.787	8.491	0.321	34.483	15.697	13.740	0.260	33.393
✓		✓	9.143	4.509	0.324	49.805	16.689	8.578	0.303	34.194	15.926	14.869	0.261	33.004
✓	✓		7.799	3.754	0.284	57.252	11.938	6.860	0.257	35.614	10.339	8.208	0.216	40.458
	✓	✓	7.710	3.899	0.280	58.969	12.559	6.188	0.246	37.444	11.296	8.460	0.224	42.294
✓	✓	✓	**7.415**	**3.332**	**0.273**	**59.788**	**10.943**	**5.685**	**0.224**	**41.505**	**9.752**	**7.061**	**0.209**	**44.007**

Table 2. Comparison between GazeNeRF + expression-guided neural renderer and GazeGaussian on ETH-xgaze

Method	Gaze↓	Head Pose↓	SSIM↑	PSNR↑	LPIPS↓	FID↓	Identity Similarity↑	FPS↑
GazeNeRF	6.944	3.470	0.733	15.453	0.291	81.816	45.207	46
GazeNeRF + EGNR	6.854	3.025	0.764	16.147	0.258	67.219	50.268	44
GazeGaussian (Ours)	**6.622**	**2.128**	**0.823**	**18.734**	**0.216**	**41.972**	**67.749**	**74**

The fundamental constraint lies in GazeNeRF's representation, which lacks the explicit modeling of gaze and facial expression dynamics offered by GazeGaussian's two-stream Gaussian structure. GazeNeRF struggles to achieve fine-grained expression synthesis and accurate gaze alignment, which are critical for high-fidelity gaze redirection.

In contrast, GazeGaussian leverages the strengths of the expression-guided neural renderer with its specialized Gaussian-based eye rotation representation and two-stream structure, enabling superior expression modeling and gaze control. This allows GazeGaussian to achieve higher fidelity, identity preservation, and rendering accuracy compared to GazeNeRF, even when enhanced with the expression-guided neural renderer. These results highlight the importance of combining advanced neural rendering techniques with a robust facial and eye modeling framework for state-of-the-art performance.

C. Supplementary visualization

C.1. Visualization for ablation study

Fig. 2 presents additional qualitative results from our ablation study conducted on the ETH-XGaze dataset. These visualizations highlight the importance of each proposed component in GazeGaussian.

Without the Gaussian eye rotation representation, the model struggles to achieve accurate eye control, resulting in noticeable deviations in gaze direction and reduced realism in the eye region. This demonstrates the critical role of the Gaussian eye rotation representation in enabling precise and realistic gaze redirection. Additionally, the absence of the expression-guided neural renderer leads to a significant loss in facial detail and expression fidelity. With the renderer included, the synthesized images exhibit finer facial details and improved consistency with the target identity, showcasing the renderer's effectiveness in enhancing the overall quality of face synthesis. These results confirm that both components contribute significantly to the superior performance and visual fidelity of GazeGaussian.

C.2. Visualization for cross-dataset comparison

We provide additional cross-dataset comparison visualizations for MPIIFaceGaze (Fig. 4), ColumbiaGaze (Fig. 5) and GazeCapture (Fig. 6). Compared to the baseline, GazeGaussian achieves high-fidelity gaze redirection with superior image synthesis quality.

C.3. Visualization for identity morphing

Fig. 7 showcases identity morphing results on the ETH-XGaze dataset. For this experiment, we randomly select two subjects with identical gaze directions and head poses. By interpolating their latent codes, we generate a smooth transition between the two identities while keeping the gaze direction and head pose consistent.

This visualization demonstrates the capability of GazeGaussian to preserve gaze alignment and head orientation during synthesis, even as the facial features gradually change according to the interpolated latent codes. The results highlight the robustness of GazeGaussian in maintaining high-fidelity gaze redirection while adapting facial characteristics as required. This ability to control identity-specific details while preserving gaze and pose consistency underscores the flexibility and effectiveness of the proposed method.

C.4. Visualization for transformed Gaussians

To demonstrate the advantages of GazeGaussian's explicit incorporation of head pose and gaze direction for rotating Gaussians in the head and eye regions, we visualize the Gaussians after deformation from the canonical space. As shown in Fig. 3, the explicit support for rotation and translation in GazeGaussian allows the deformed Gaussians to form a reasonable spatial distribution and accurate color representation. This capability enables precise geometric control and

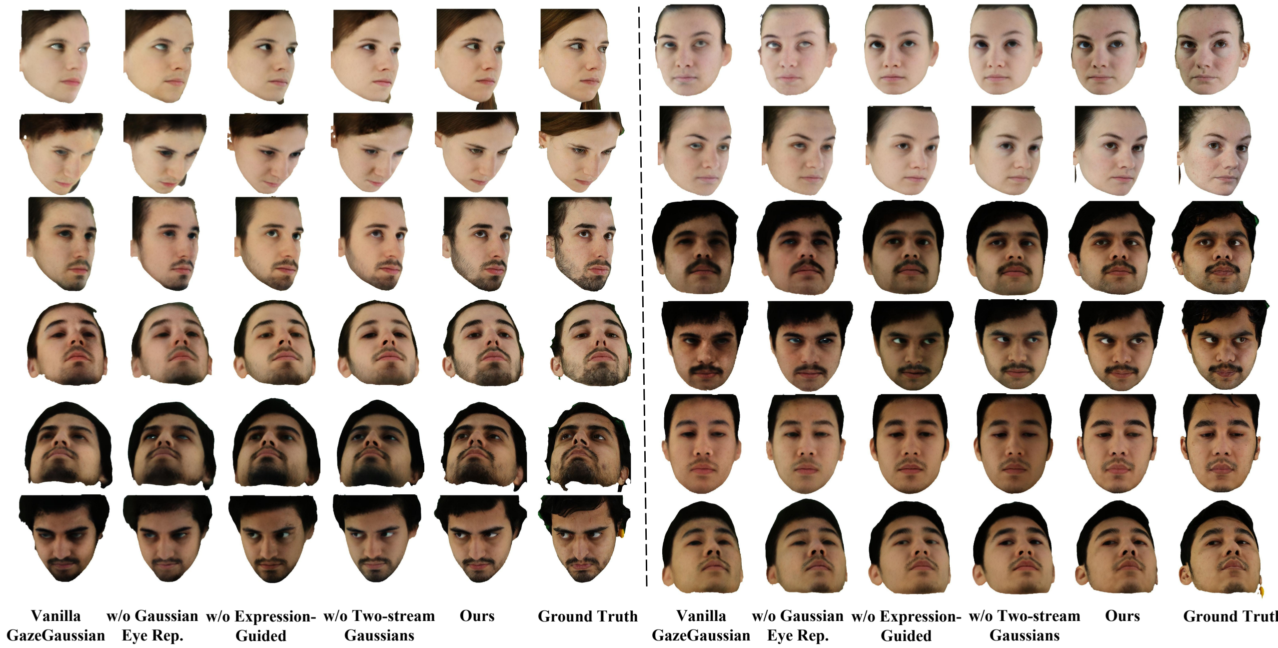

Figure 2. Additional qualitative ablation study on the ETH-XGaze dataset.

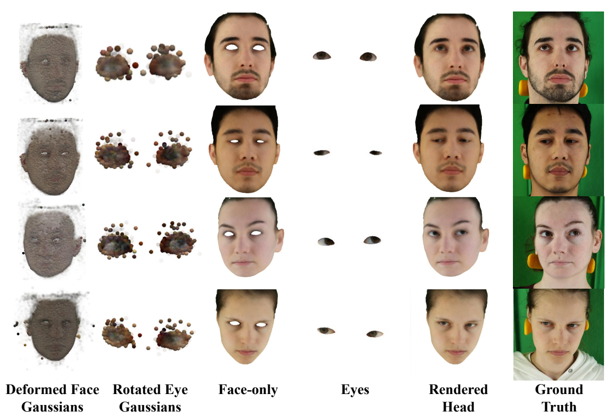

Figure 3. Visualization of transformed two-stream Gaussians after deformation from the canonical space.

high-fidelity image rendering. In contrast, GazeNeRF performs rotations only on the feature map level, failing to fully deform in 3D space, which limits its performance compared to our method.

D. Implementation details

We use the Adam optimizer [19], with a learning rate that follows an exponential decay schedule, starting at 1×10^{-4}. We use the VGG-based network pre-trained on ImageNet, as provided by the GazeNeRF [36] implementation, and fine-tune it on the ETH-XGaze training set for the functional loss $\mathcal{L}_G$ as the pre-trained gaze estimator. Additionally, we utilize the ResNet50 backbone from the GazeNeRF [36] framework, trained on the ETH-XGaze training set, to output gaze and head pose for evaluation purposes. All experiments were conducted on an NVIDIA 4090 GPU. We first train an SDF network to extract the neutral mesh and initialize the two-stream Gaussian parameters in 10 epochs. The full pipeline was then trained for an additional 20 epochs until convergence. The loss weights follow the same configuration as described in the method section of the main text.

E. Dataset and pre-processing details

Following the baseline GazeNeRF [36], all experiments are conducted on four widely used datasets.

ETH-XGaze [59] is a large-scale gaze estimation dataset featuring high-resolution images across a wide range of head poses and gaze directions. Captured with a multi-view camera setup under varying lighting conditions, it includes 756,000 frames from 80 subjects for training. Each frame contains images from 18 different camera perspectives. Additionally, a person-specific test set includes 15 subjects, each with 200 images provided with ground-truth gaze data.

ColumbiaGaze [39] contains 5,880 high-resolution images from 56 subjects. For each subject, images were taken in five distinct head poses, with each pose covering 21 preset gaze directions, allowing for detailed gaze estimation in controlled conditions.

MPIIFaceGaze [56, 57] is tailored for appearance-based gaze prediction. MPIIFaceGaze offers 3,000 face images for each of 15 subjects, paired with two-dimensional gaze labels to facilitate gaze estimation research.

GazeCapture [21] is a large-scale dataset collected through crowd-sourcing, featuring images captured across different poses and lighting conditions. For cross-dataset comparison, we use only the test portion, which includes data from 150

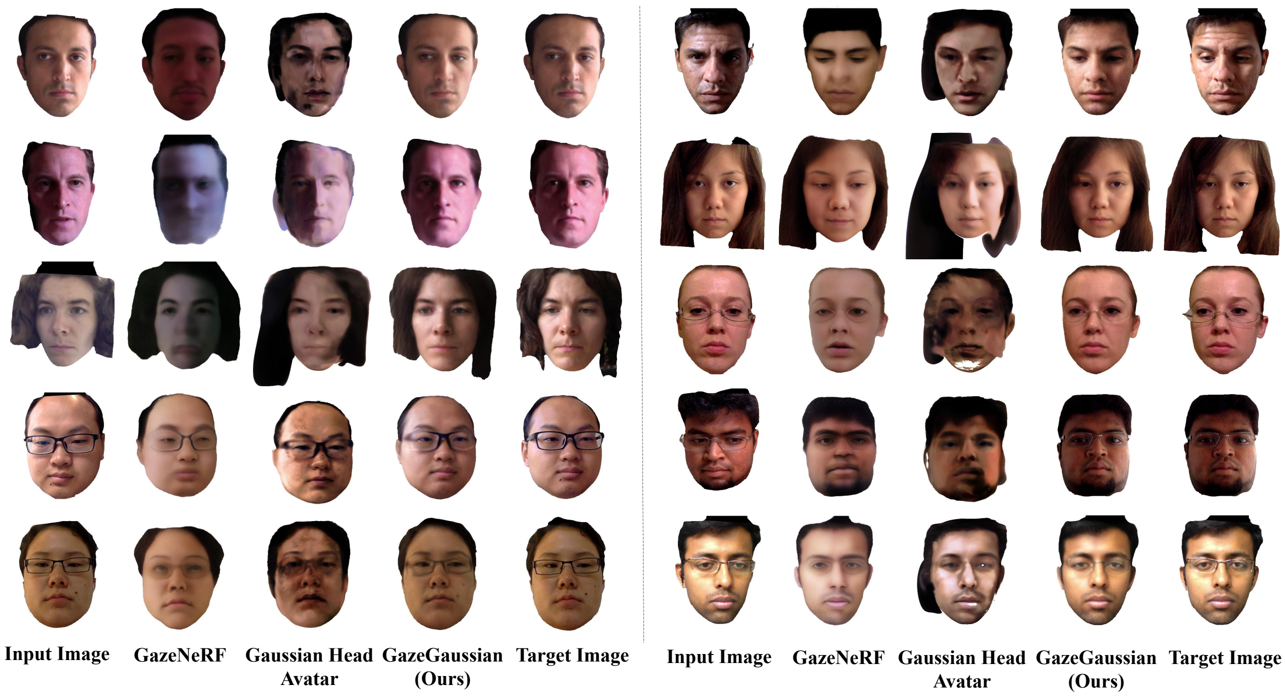

Figure 4. Cross-dataset comparison: Visualization of generated images from the MPIIFaceGaze using our GazeGaussian, GazeNeRF, and Gaussian Head Avatar.

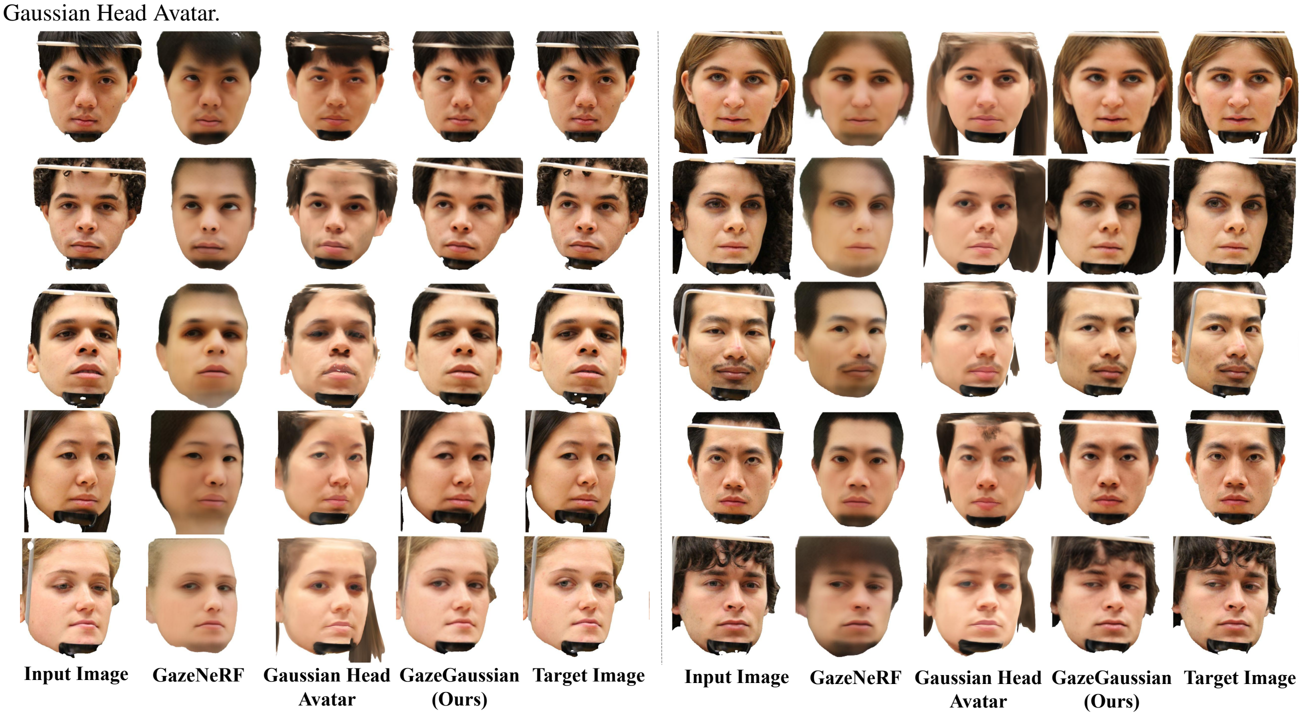

Figure 5. Cross-dataset comparison: Visualization of generated images from the ColumbiaGaze using our GazeGaussian, GazeNeRF, and Gaussian Head Avatar.

distinct subjects.

Pre-processing. We follow the preprocessing steps in GazeNeRF [36] and Gaussian Head Avatar [51]. The original resolution of ETH-XGaze [59] images is 6K × 4K, while images from other datasets vary in resolution. To standardize, we preprocess all images using the normalization method, aligning the rotation and translation between the camera and face coordinate systems. The normalized distance from the camera to the face center is fixed at 680mm. To extract 3DMM parameters and generate masks for the eyes and face-

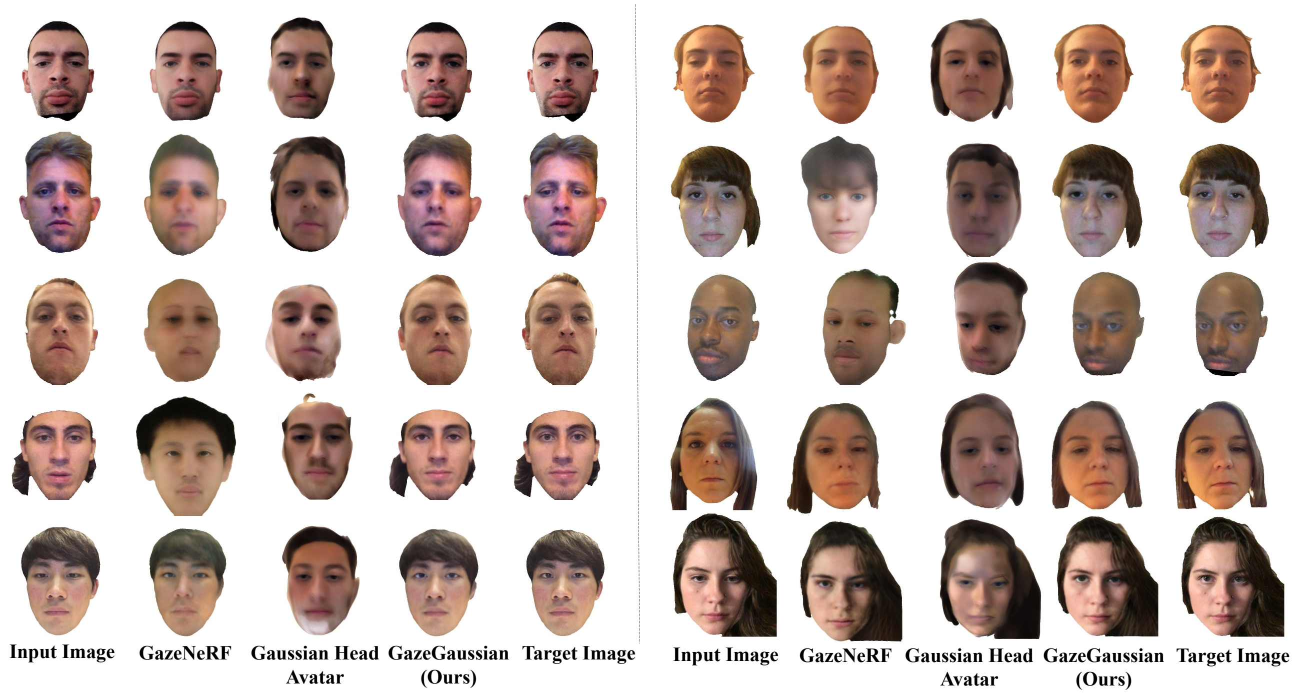

Figure 6. Cross-dataset comparison: Visualization of generated images from the GazeCapture using our GazeGaussian, GazeNeRF, and Gaussian Head Avatar.

Figure 7. Face morphing results on the ETH-XGaze dataset.

only regions, we utilize the face parsing model from [64]. GazeGaussian is trained on a single NVIDIA 4090 GPU for 20 epochs on the train set from ETH-XGaze. During inference, GazeGaussian fine-tunes on a single input image, taking approximately 30 seconds for fine-tuning and 0.2 seconds per image for generation.

CVPR
#3250

F. Ethical consideration and limitations

Our method enables the generation of highly realistic portrait videos, which, if misused, could contribute to the spread of misinformation, manipulate public opinion, and undermine trust in media sources, with significant societal consequences. Therefore, it is essential to develop reliable methods to differentiate between authentic and fabricated content. We strongly condemn the unauthorized or malicious use of this technology and emphasize the importance of considering ethical implications in its deployment.

Figure 8. Example of a failure case.

While GazeGaussian represents a significant advancement in gaze redirection quality, there is still one unresolved issue. Due to limitations in facial tracking models such as FLAME, it remains challenging to accurately model accessories such as glasses, earrings, and even hair details as shown in Fig. 8. An existing method [26] has attempted to use cylindrical Gaussian representations to capture the movement of long hair. To further enhance the diversity of character generation, improving the 3DGS facial representation will be a key focus of our future work.