OneGT: One-Shot Geometry-Texture Neural Rendering for Head Avatars

Supplementary Material

A. Training and Inference Cost

In this section, we compare the training and inference costs of our proposed method in detail with our competitive peers. The costs for the comparing mainly consist of 4 aspects: **(1)** The computational resources consumed during training, **(2)** The details of the data supporting the training, **(3)** The number of trainable parameters of the model along with the TFLOPS of one single forward process, and **(4)** The inference speed. The methods we compare include: Face vid2vid [8], AniPortrait [9], LivePortrait [3], X-Portrait [10] and Real3dPortrait [13].

Computational Resources. In Tab. 1 we list the demanded computational resources of all the tested methods on generating 512 px results. We use the officially reported information and the implementations of the above methods for testing if they are provided. We use "-" to mark the data that is not reported. Without an available official implementation, we test Face vid2vid using a popular unofficial project[1] and mark the related results with *italic* type. Considering that the computing capabilities of different GPUs vary, we use tools from [4] to measure the integrated computational resources following the official release website[2] of SD [7]. The quantitative results show that our proposed method is competitive in terms of training resource consumption, which indicates the superiority and efficiency of the designed structure of the model.

Methods	GPU Days	Carbon Emitted (kgCO$_2$eq)
Face vid2vid	*56* V100	*174.18*
AniPortrait	16 A100	66.36
LivePortrait	80 A100	331.78
X-Portrait	-	-
Real3dPortrait	59 A100	246.07
Ours (Skeleton Stage)	6 V100	18.68
Ours (Texture Stage)	32 A800	82.94

Table 1. The required training computational resources of the mentioned methods. Note that "-" means undisclosed data, and *italic* type denotes that the value is estimated using the unofficial implementation.

Training data. In Tab. 2, we list the details of the training datasets. Considering that the splitting strategy for video clips adopted by different methods is different, we compare the dataset sizes based on the approximate number of original video clips directly. Public data refers to open-source datasets, while in-house data refers to the datasets that are not released in public. "None" indicates that such type of data is not involved, and "-" indicates that the related information is not reported. Quantitatively, our method requires relatively less data, proving that our method has superior structural design and is easier to follow. As we claim in our main article, by introducing the synthetic data, the required real-world data of our method is only 10%-30% of that of other methods. Note that for the LightStage dataset [11, 12] used by LivePortrait [3] and the in-house dataset used by X-Portrait [10], these datasets are collected in the real world. Considering the devices and the conditions of the data collection as well as the portrait copyrights, they are inaccessible datasets for external users. In contrast, our in-house dataset is synthetic data generated by the rendering tools. The assets are collected from the open-source domain and the rendering software is also accessible.

Methods	Public Data/K	In-house Data/K
Face vid2vid	780	None
AniPortrait	50	None
LivePortrait	562	800
X-Portrait	None	-
Real3dPortrait	180	None
Ours	16	550

Table 2. The required training data of the mentioned methods. Note that "-" means undisclosed data, and "None" means such type of data is not involved.

Methods	Trainable Parameters/B	TFLOPS
Face vid2vid	*0.11*	*1.20*
AniPortrait	1.71	3.87
LivePortrait	0.13	1.26
X-Portrait	2.02	4.22
Real3dPortrait	0.15	1.39
Ours	1.23	2.07

Table 3. The trainable parameters and required TFLOPS for a single 512 px forward process. Note that *italic* type denotes that the value is estimated using the unofficial implementation.

Trainable Parameters and TFLOPS. In Tab. 3, we list the trainable parameters and the TFLOPS of all mentioned frameworks required for one forward pass at 512px. We use [3] to calculate the TFLOPS. Similarly, the test of Face vid2vid is done with the unofficial implementation. Clearly, models

[1] https://github.com/zhanglonghao1992/One-Shot_Free-View_Neural_Talking_Head_Synthesis

[2] https://huggingface.co/CompVis/stable-diffusion-v1-1

[3] https://www.deepspeed.ai/tutorials/flops-profiler/#flops-measurement

based on non-diffusion frameworks including Face vid2vid, LivePortrait and Real3dPortrait have smaller sizes and lower TFLOPS for one forward pass. Compared to other diffusion framework methods including AniPortrait and X-Portrait, our method has advantages in terms of computational cost and model size, aided by our lightweight Transformers-based skeleton anchoring module and efficient DiT-based texture rendering module.

Inference Speed. We test the inference speed of our method and our competitive peers' methods. The test is conducted on a single Nvidia A800 GPU at 512px. For all the methods, we use the officially reported hyper-parameters for inference. The comparisons are listed in Tab. 4. A detailed discussion is listed in the following section about our limitations.

Methods	Face vid2vid	AniPortrait	LivePortrait	X-Portrait	Real3dPortrait	**Ours**
Speed/fps	*20.0*	0.75	24.1	0.56	1.89	0.64

Table 4. The trainable parameters and required TFLOPS for a single 512 px forward process. Note that *italic* type denotes that the value is estimated using the unofficial implementation.

B. More Results

In this section, we showcase more visualization results of our model. Fig. 1 and Fig. 2 show more editing cases such as positions, shapes, and Euler angles. Fig. 3 shows more effects of the model in free-animating scenarios.

Fig. 1 shows the editing on the horizontal position, the vertical position, the zoom, the facial shapes, and the expressions of our model, which is marked as (a)-(e) in the figure respectively. We use CelebA [5] dataset for these tests. Fig. 2 shows the editing on the Euler angles. The rotation of the yaw, pitch, and roll is shown by (a)-(c) in the figure. We use images from CelebA, the official test split of VFHQ, along with a few in-the-wild portraits to complete these tests. Fig. 3 shows the animating cases of our model. (a) shows the self-reenactment cases on the VFHQ testset, and (b) shows the cross-reenactment cases for in-the-wild pictures.

In our framework, the skeleton anchoring part is responsible for learning accurate rigid and non-rigid transformations, while the texture rendering part is responsible for accurately rendering the given texture onto the corresponding skeleton. Such design exposes the interactive parts to the users, making it easier to customize their own effects. Benefiting from this decoupled design of skeleton and texture, our framework learns various mixed attributes better. In particular, compared to other existing methods, the precise skeleton structure serves as a kind of prior knowledge provided to the texture rendering module, rather than the target to learn. This ensures that the interpretability of the overall framework stays at a quite high level, allowing modules to focus more on learning the matching relationship between texture and skeleton structure, thus avoiding overfitting caused by the implicit modeling. For example, our model performs well on the stylized cases, even though our model is trained without any stylized images.

C. Limitations

Although our framework has made progress in design methods and generated qualities, there still remains room for improvement in the temporal consistency and the computational cost. In terms of the temporal consistency, our pipeline runs without any temporal-related modules. We generate our dynamic results by directly combining all the static images, and the temporal consistency is purely achieved by the basic stability of our model. In most cases, as the skeleton information has dense and intense spatial correlation with the desired rendered result, the generated temporal consistency is satisfying if adequate stable driving frames are provided. However, when the movement of the character in the frame becomes excessively large, the instability and abrupt changes in facial structure and background within the dynamic results are further exacerbated. Adding the mostly proposed temporal attention modules to the framework may ease such problems. Our framework contains a diffusion module, thus there is some gap in complexity compared to non-diffusion frameworks. The diffusion model requires multiple sampling steps to obtain the final result, while non-diffusion frameworks like GANs only require a single inference; Moreover, the diffusion model contains more computationally intensive operators such as attention. For the former point, methods like LCM [6] can be combined to shorten the model's inference time by reducing the number of sampling steps. For the latter point, more efficient operators with better principles like Mamba [1, 2] may greatly improve the overall performance of the model.

D. Further Discussion: the Adaptive Strategy

In this section, we delve a little deeper into the discussion of our proposed adaptive strategy ablation mentioned in the article. The results have shown that our proposed adaptive strategy effectively reduces the training difficulty of the model. In fact, in our framework, after converting a total of 1,107 keypoints into Gaussian distribution features, the overall features obtained own a very high complexity, which is reflected in the number of channels and spatial scales of the feature map. If all these abundant features are treated equally, the model will struggle to forcibly map each of them to multiple complex contents in the image, proven by the generated artifacts in the unexpected areas. Our adaptive strategy gives the model more flexibility, allowing the model to allocate a looser floating range for the skeletal points which are more

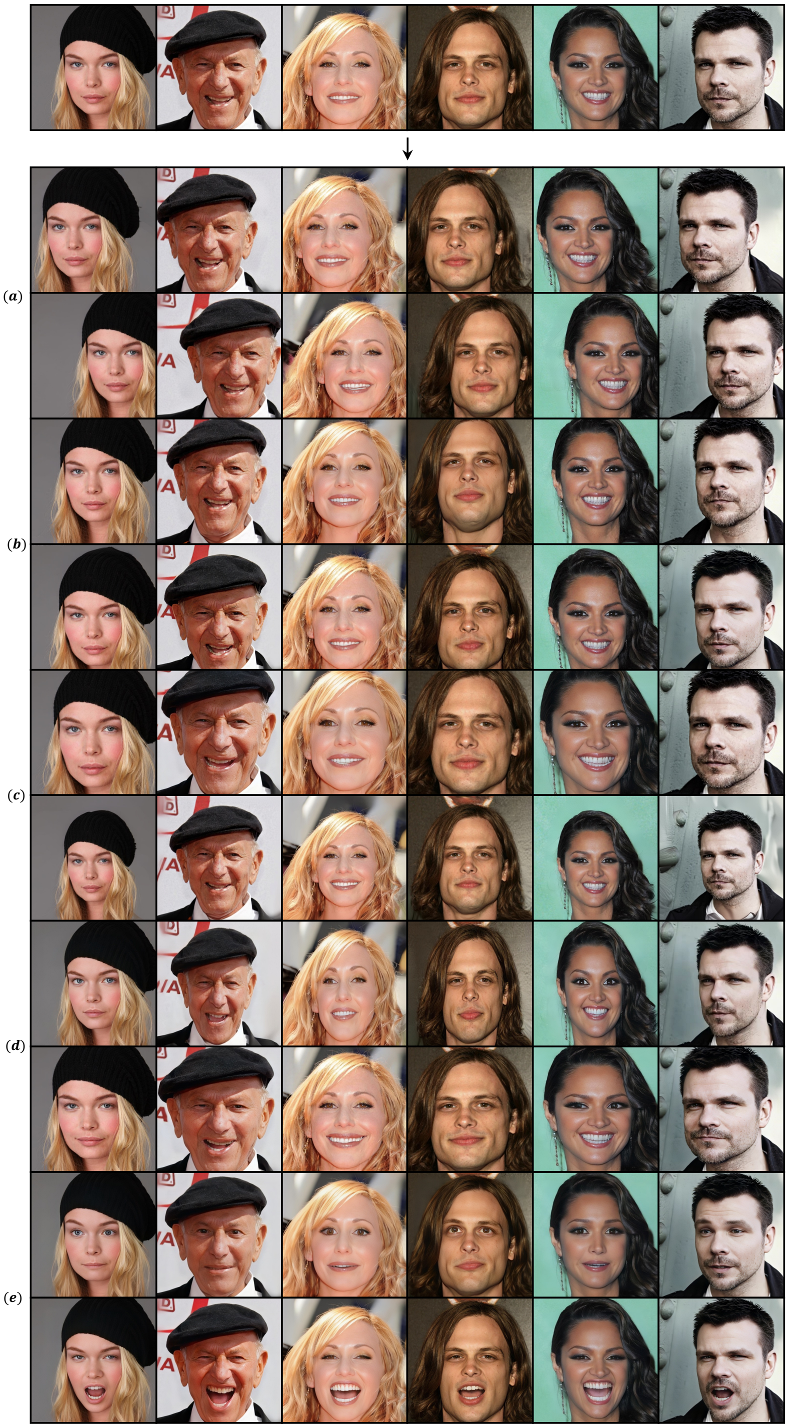

Figure 1. More qualitative editing results of our methods. (a)-(e) denotes editing on the horizontal position, the vertical position, the zoom, the facial shapes, and the expressions respectively.

difficult to learn, so that the actual responsible areas for each point can dynamically adjust, reducing the training difficulty of the model.

E. User Study

We conduct a user study for the animating results, including the self-reenactment cases and the cross-reenactment cases.

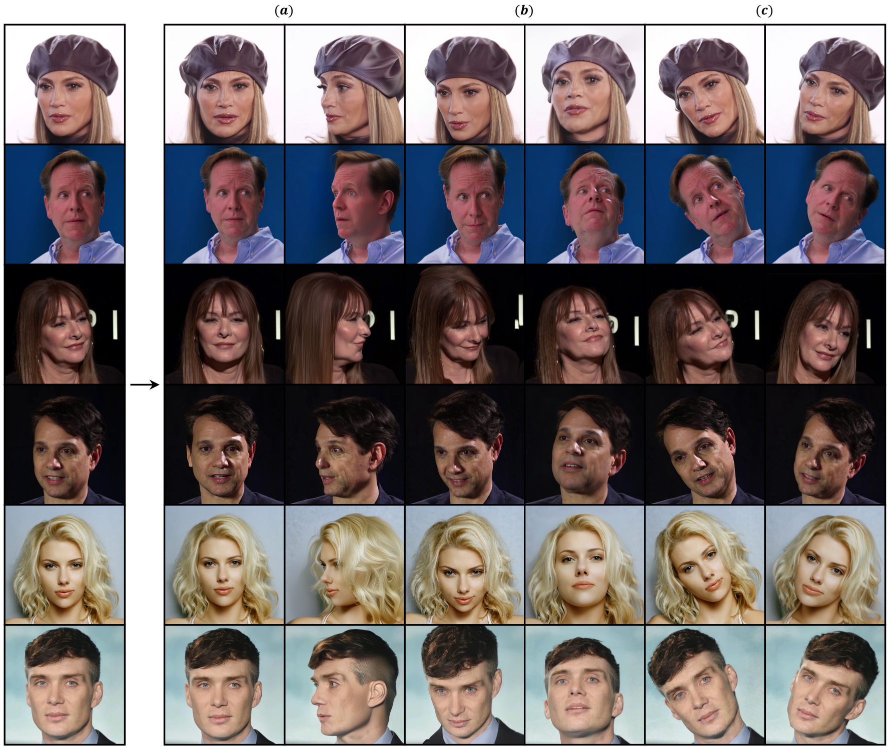

Figure 2. More qualitative results of editing the Euler angles. (a)-(c) represent editing the yaw, pitch, and roll respectively.

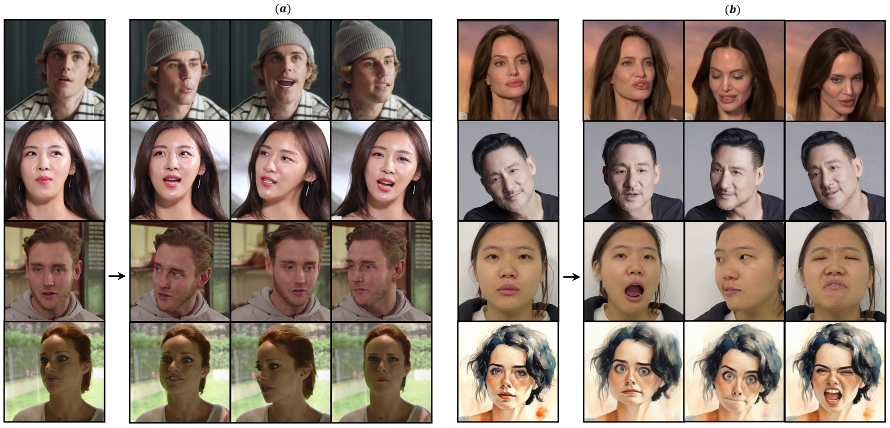

Figure 3. More qualitative results of free animating. (a) represents the self-reenactment case, and (b) represents the cross-reenactment case.

The evaluation metrics include the quality of the generated images, the fidelity in terms of poses and expressions to the target image, and the retention of the identity information of the source image in the generated results, which are abbreviated to “Quality”, “Fidelity” and “Retention” respectively. The comparisons are conducted over Face vid2vid, AniPortrait, LivePortrait, X-Portrait, Real3dPortrait and ours. We prepare 25 pairs of source-target images for self-reenactment and cross-reenactment scenarios respectively, and we distribute the tests to 150 users. For each pair of test samples, users are asked to rank the results generated by the 5 methods according to all the evaluation metrics, thus we obtain 7,500 choices for each metric. We calculate the proportion of each framework that receives the best evaluation, and such quantitative results of the self-reenactment case and the cross-reenactment case are shown in Tab. 5 and Tab. 6

respectively. From a quantitative perspective, our method outperforms its competitive peers.

Methods	Quality	Fidelity	Retention
Face vid2vid	0.1200%	2.5333%	0.0533%
AniPortrait	14.6667%	6.0267%	1.5733%
LivePortrait	14.9867%	12.0800%	16.4267%
X-Portrait	13.3867%	9.4933%	9.6267%
Real3dPortrait	0.1333%	2.0267%	6.0800%
Ours	**71.2267%**	**67.8400%**	**66.2400%**

Table 5. The quantitative results of the user study for self-reenactment cases.

Methods	Quality	Fidelity	Retention
Face vid2vid	0.9333%	0.4267%	5.8400%
AniPortrait	6.8267%	6.7200%	3.9333%
LivePortrait	11.2533%	10.5600%	7.9733%
X-Portrait	7.7867%	7.2267%	8.1333%
Real3dPortrait	1.7067%	6.0800%	2.0667%
Ours	**71.4933%**	**68.9867%**	**72.0533%**

Table 6. The quantitative results of the user study for cross-reenactment cases.

F. Bias of the Synthetic Data

As mentioned in the article, the synthetic data plays an important role in our designed pipeline, which supports the whole training data of the skeleton stage and a large part of the rendering stage. However, as discussed before, the synthetic data is born with a kind of bias: it lacks realism in terms of the texture compared with the real-world data, and there is subtle difference between the rendered geometry structures and the real ones. In our work, we have incorporated a few designs to mitigate such bias: **(1)** Our designed skeleton anchoring stage outputs **pure** keypoints, preventing the leakage of the synthetic texture into the posterior stage; **(2)** The Gaussian-like representation described in our article converts strict keypoint locations to potential position ranges, which brings in more flexibility and reduces the bias of synthetic geometry structures; **(3)** The rendering stage runs in the latent space of the VAE, which rarely has high frequency information and limits the harm from the synthetic texture. Thus we handle real-world and synthetic data through a unified framework without explicit domain adaptation. Besides, we manually add a little synthetic data to the training of the texture-rendering stage, making the model more adaptable to the skeleton stage trained on the synthetic data. The actual ratio of the synthetic data to the real-world one is 1:9.

References

[1] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024. 2

[2] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2

[3] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024. 1

[4] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019. 1

[5] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 2

[6] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023. 2

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[8] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 1

[9] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024. 1

[10] You Xie, Hongyi Xu, Guoxian Song, Chao Wang, Yichun Shi, and Linjie Luo. X-portrait: Expressive portrait animation with hierarchical motion attention. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 1

[11] Haotian Yang, Mingwu Zheng, Wanquan Feng, Haibin Huang, Yu-Kun Lai, Pengfei Wan, Zhongyuan Wang, and Chongyang Ma. Towards practical capture of high-fidelity relightable avatars. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 1

[12] Haotian Yang, Mingwu Zheng, Chongyang Ma, Yu-Kun Lai, Pengfei Wan, and Haibin Huang. Vrmm: A volumetric relightable morphable head model. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 1

[13] Zhenhui Ye, Tianyun Zhong, Yi Ren, Jiaqi Yang, Weichuang Li, Jiawei Huang, Ziyue Jiang, Jinzheng He, Rongjie Huang, Jinglin Liu, et al. Real3d-portrait: One-shot realistic 3d talking portrait synthesis. *arXiv preprint arXiv:2401.08503*, 2024. 1