HeadSculpt: Crafting 3D Head Avatars with Text (*Supplementary Material*)

Contents

A Implementation details   **2**

    A.1 Details about 3D scene models 2

    A.2 Details about textual inversion 2

B Further analysis   **3**

    B.1 Effectiveness of textual inversion on 2D generation 3

    B.2 Inherent bias in 2D diffusion models 3

C Additional qualitative comparisons   **5**

List of Tables

    1 Hyper-parameters of HeadSculpt 2

List of Figures

    1 Samples of the tiny dataset collected for textual inversion 3

    2 Analysis of the learned `<back-view>` on 2D image generation 4

    3 Analysis of the inherent bias in 2D diffusion models 4

    4 Additional comparisons with existing text-to-3D methods (Part 1) 5

    5 Additional comparisons with existing text-to-3D methods (Part 2) 6

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Table 1: **Hyper-parameters of HeadSculpt**.

Camera setting	θ range	(20, 110)
	Radius range	(1.0, 1.5)
	FoV range	(30, 50)
Render setting	Resolution for coarse	(64, 64)
	Resolution for fine	(512, 512)
	Max num steps sampled per ray	1024
	Iter interval to update extra status	16
Diffusion setting	Guidance scale	100
	t range	(0.02, 0.98)
	$\omega(t)$	$\sqrt{\alpha_t}(1 - \alpha_t)$
Training setting	#Iterations for coarse	$70k$
	#Iterations for fine	$50k$
	Batch size	4
	LR of grid encoder	1e-3
	LR of NeRF MLP	1e-3
	LR of s_i and $\Delta \mathbf{v}_i$ in DMTET	1e-2
	LR scheduler	constant
	Warmup iterations	$20k$
	Optimizer	Adam (0.9, 0.99)
	Weight decay	0
	Precision	fp16
Hardware	GPU	$1 \times$ Tesla V100 (32GB)
	Training duration	1h (coarse) + 1h (fine)

A Implementation details

A.1 Details about 3D scene models

In the coarse stage, we make use of the grid frequency encoder $\gamma(\cdot)$ from the publicly available Stable DreamFusion [7]. This encoder maps the input $\mathbf{x} \in \mathbb{R}^3$ to a higher-frequency dimension, yielding $\gamma(\mathbf{x}) \in \mathbb{R}^{32}$. The MLP within our NeRF model consists of three layers with dimensions [32, 64, 64, 3+1+3]. Here, the output channels '3', '1', and '3' represent the predicted normals, density value, and RGB colors, respectively. In the fine stage, we directly optimize the signed distance value $s_i \in \mathbb{R}$, along with a position offset $\Delta \mathbf{v}_i \in \mathbb{R}^3$ for each vertex $\mathbf{v}_i$. We found that fitting s_i and $\mathbf{v}_i$ into MLP, as done by Fantasia3D [8], often leads to diverged training.

To ensure easy reproducibility, we have included all the hyperparameters used in our experiments in Tab 1. The other hyper-parameters are set to be the default of Stable-DreamFusion [7].

A.2 Details about textual inversion

In the main paper, we discussed the collection of a tiny dataset consisting of 34 images depicting the back view of heads. This dataset was used to train a special token, `<back-view>`, to address the ambiguity associated with the back view of landmarks. The images in the dataset were selected to encompass a diverse range of gender, color, age, and other characteristics. A few samples from the dataset are shown in Fig. 1. While our simple selection strategy has proven effective in our specific case, we believe that a more refined collection process could further enhance the controllability of the learned `<back-view>` token. We use the default training recipe provided by HuggingFace Diffusers [1], which took us 1 hour on a single Tesla V100 GPU.

---

[1]https://github.com/huggingface/diffusers/blob/main/examples/textual_inversion

2

Figure 1: **Samples of the tiny dataset collected for learning** `<back-view>` **token**.

B Further analysis

B.1 Effectiveness of textual inversion on 2D generation

To show the effectiveness of the learned `<back-view>` token, we conduct an analysis of its control capabilities in the context of 2D generation results. Specifically, we compare two generation results using Stable Diffusion [6], with both experiments sharing the same random seed. One experiment has the plain text prompt appended with the plain phrase "back view," while the other experiment utilizes the learned special token `<back-view>` in the prompt. We present a selection of randomly generated results in Fig. 2. The observations indicate that the `<back-view>` token effectively influences the pose of the generated heads towards the back, resulting in a distinct appearance. Remarkably, the `<back-view>` token demonstrates a notable generalization ability, as evidenced by the Batman case, despite not having been trained specifically on back views of Batman in the textual inversion process.

B.2 Inherent bias in 2D diffusion models

In our main paper, we discussed the motivation behind our proposed identity-aware editing score distillation (IESD), which can be attributed to two key factors. Firstly, the limitations of prompt-based editing [4, 2] are due to the inherent bias present in Stable Diffusion (SD). Secondly, while InstructPix2Pix (IP2P) [1] offers a solution by employing instruction-based editing to mitigate bias, it often results in identity loss. To further illustrate this phenomenon, we showcase the biased 2D outputs of SD and ControlNet-based IP2P in Fig. 3. Modified descriptions and instructions are utilized in these respective methods to facilitate the editing process and achieve the desired results. The results provide clear evidence of the following: (1) SD generates biased outcomes, with a tendency to underweight the "older" aspect and overweight the "skull" aspect in the modified description; (2) IP2P demonstrates the ability to edit the image successfully, but it faces challenges in preserving the identity of the avatar.

The aforementioned inherent biases are amplified in the domain of 3D generation (refer to Fig. 7 in the main paper) due to the optimization process guided by SDS loss, which tends to prioritize view consistency at the expense of sacrificing prominent features. To address this issue, our proposed IESD approach combines two types of scores: one for editing and the other for identity preservation. This allows us to strike a balance between preserving the initial appearance and achieving the desired editing outcome.

3

w/ **"back view"**	w/ `<back-view>`	w/ **"back view"**	w/ `<back-view>`	w/ **"back view"**	w/ `<back-view>`
seed: 413		*seed: 16772*		*seed: 40805*	

a DSLR portrait of Obama

seed: 50682		*seed: 93440*		*seed: 96458*	

a DSLR portrait of Hillary Clinton

seed: 2367		*seed: 19656*		*seed: 62156*	

a DSLR portrait of a boy with facial painting

seed: 53236		*seed: 62424*		*seed: 72649*	

a DSLR portrait of Batman

Figure 2: **Analysis of the learned `<back-view>` on 2D image generation**. For each pair of images, we present two 2D images generated with the same random seed, where the left image is conditioned on the plain text "back view" and the right image is conditioned on the `<back-view>` token.

Landmark Map	**Stable Diffusion**		**Reference Image**	**InstructPix2Pix**	
	seed: 19056	*seed: 72854*		*seed: 50233*	*seed: 64136*

Modified description: *a DSLR portrait of +[older] Saul Goodman*			**Instruction**: *make him older*		
	seed: 5427	*seed: 91282*		*seed: 60104*	*seed: 88141*

Modified description: *a ~~DSLR portrait~~ skull of Vincent van Gogh*			**Instruction**: *turn his face into a skull*		
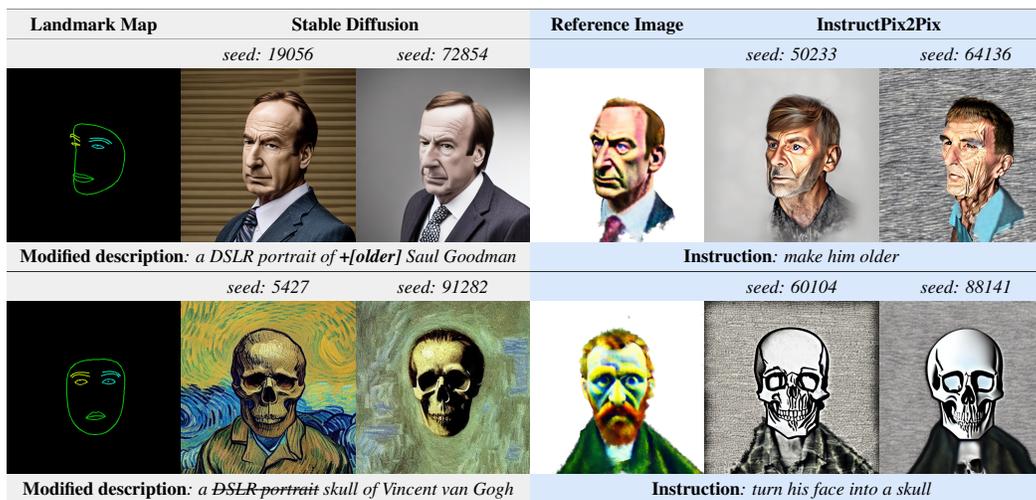
Figure 3: **Analysis of the inherent bias in 2D diffusion models**. For each case, we display several 2D outputs of SD and IP2P, utilizing modified descriptions and instructions, respectively, with reference images from our coarse-stage NeRF model to facilitate the editing process.

4

C Additional qualitative comparisons

We provide more qualitative comparisons with four baseline methods[7, 3, 5, 8] in Fig. 4 and Fig. 5. These results serve to reinforce the claims made in Sec 4.1 of the main paper, providing further evidence of the superior performance of our HeadSculpt in generating high-fidelity head avatars. These results showcase the ability of our method to capture intricate details, realistic textures, and overall visual quality, solidifying its position as a state-of-the-art solution in this task.

Notably, to provide a more immersive and comprehensive understanding of our results, we include multiple outcomes of our HeadSculpt in the form of 360° **rotating videos**. These videos can be accessed in the accompanying **HTML file**, enabling viewers to observe the generated avatars from various angles and perspectives.
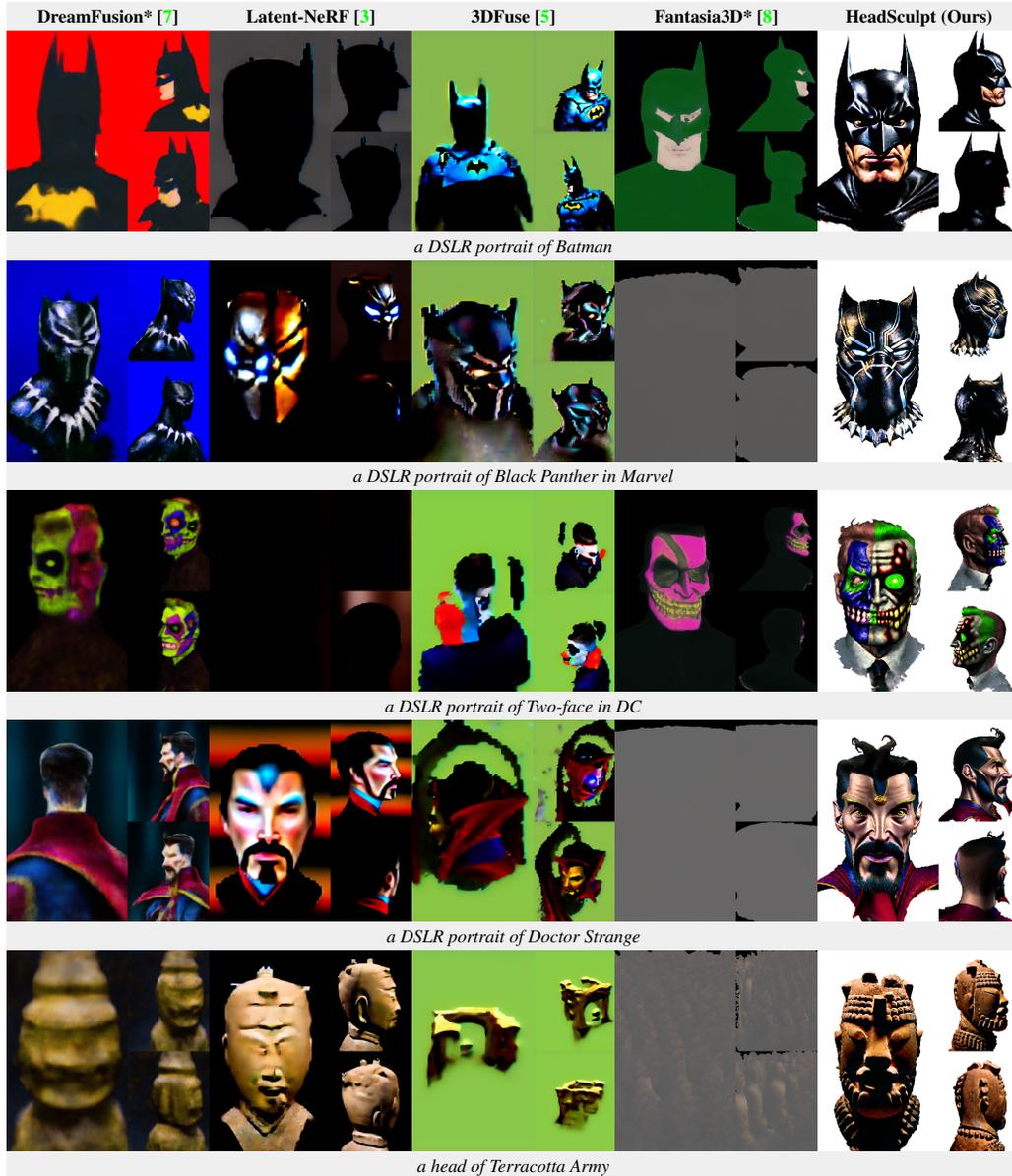
Figure 4: **Additional comparisons with existing text-to-3D methods**. *Non-official implementation.

5

| DreamFusion* [7] | Latent-NeRF [3] | 3DFuse [5] | Fantasia3D* [8] | HeadSculpt (Ours) |

a head of Simpson in the Simpsons

a head of Naruto Uzumaki

a DSLR portrait of Napoleon Bonaparte

a DSLR portrait of Leo Tolstoy

a DSLR portrait of Audrey Hepburn

a DSLR portrait of Obama with a baseball cap

a DSLR portrait of Taylor Swift

Figure 5: **Additional comparisons with existing text-to-3D methods**. *Non-official implementation.

6

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 3

[2] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, 2023. 3

[3] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022. 5, 6

[4] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2022. 3

[5] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. *arXiv preprint arXiv:2303.07937*, 2023. 5, 6

[6] Stability.AI. Stable diffusion. https://stability.ai/blog/stable-diffusion-public-release, 2022. 3

[7] Jiaxiang Tang. Stable-dreamfusion: Text-to-3d with stable-diffusion. https://github.com/ashawkey/stable-dreamfusion, 2022. 2, 5, 6

[8] Jiaxiang Tang. Fantasia3d.unofficial. https://github.com/ashawkey/fantasia3d.unofficial, 2023. 2, 5, 6

7