Not All Steps are Created Equal: Selective Diffusion Distillation for Image Manipulation Supplemental Material

Luozhou Wang[1,4*] Shuai Yang[1,4*] Shu Liu[2] Ying-cong Chen[1,3,4†]

[1] The Hong Kong University of Science and Technology (Guangzhou). [2] SmartMore.

[3] The Hong Kong University of Science and Technology. [4] HKUST (Guangzhou) - SmartMore Joint Lab.

A. Implementation Detail

The image resolution for the human face dataset is set to 1024×1024. The image resolution for the cat face and car dataset is set to 512×512. The hyperparameters like learning rate, weight decay, loss weight, noise type, iterations, and mapper levels are adjusted individually for each prompt based on our experimental heuristics. More details can be found in our code. It is attached with the supplementary file. All experiments are conducted on an NVIDIA RTX3090, with 24GB memory. The architecture of the image manipulator is the same as the latent mapper of [5].

B. Diffusion model as a prior

As mentioned in Sec. 3.2, Diffusion models can also be used as off-the-shelf modules in some scenarios, where one model may be a prior for another conditional model. A typical example is a diffusion model $p(x)$ trained on MNIST digits and an off-the-shelf classifier $c(x, y)$ where y is the class label. Then we can use the diffusion model to generate data of a specific class for the classifier. In theory, this means that we want to deduce $p(x \mid y)$ given $p(x)$ and $c(x, y)$. One solution is to introduce an approximate variational posterior $q(x)$ to approximate the posterior distribution $p(x|y)$, and minimize:

$$F = -\mathbb{E}_{q(x)}[\log p(x) - \log q(x)] - \mathbb{E}_{q(x)}[\log c(x, y)] \tag{1}$$

When extending this formula to the scenario of diffusion model with latent variable $x_1, ..., x_T$, we can define this approximate variational posterior $q(x)$ as point estimate $q(\mathbf{x}) = \delta(\mathbf{x} - \eta)$ [2], and then minimize:

$$\begin{aligned} F = \sum_t \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)}[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2] - \mathbb{E}_{q(\mathbf{x})}[\log c(\eta, y)], \\ x_t = \sqrt{\bar{\alpha}_t}\eta + \sqrt{1 - \bar{\alpha}_t}\epsilon. \end{aligned} \tag{2}$$

This equation optimizes η, which has the same dimensionality as data. We can regard this equation as directly sampling pixels using diffusion models to get a sample that satisfies the condition y.

Another situation is that $c(x, y)$ is a hard and non-differentiable conditional model, say a deterministic function $x = f(y)$. Then the gradient descent steps will be performed concerning $\mathbf{y}$ on

$$\sum_t \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)}[\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}f(y) + \sqrt{1 - \bar{\alpha}_t}\epsilon, t)\right\|_2^2]. \tag{3}$$

For example, $f()$ can be a latent-variable model that takes latent y as input and generates a sample x. Another example is that we can also use techniques of differentiable image parameterization. In [6], y can be parameters of a 3D volume, and f is a volumetric renderer. We can regard this equation as sampling y instead of directly sampling images using diffusion models and inputting y to this conditional model. We will get a sample from the diffusion model.

Optimizing simultaneously for all t makes it difficult to guide the sample toward a mode. Thus existing methods eighter anneal t from high to low values [2], or random select t [6]. So the actual optimization process is slightly changed to

$$\mathbb{E}_{\epsilon,t}[\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}f(y) + \sqrt{1 - \bar{\alpha}_t}\epsilon, t)\right\|_2^2] \tag{4}$$

C. Fine-grained control of image manipulation

Another benefit of our approach is that the control capability of the image manipulator can be used to control the semantics more precisely. We use StyleGAN [3] as the final generator of images, and the hierarchical nature of StyleGAN allows us to decompose more complex manipulations into different levels of manipulation. For example, we can adjust the manipulation effect of the image in three levels: coarse, medium, and fine. For better control, we used the most controlled StyleSpace [7] during the experiments. For more details, please watch the supplementary video.

*Equal contribution

†Corresponding author

D. More visual samples

In this section, we provide more visual samples (Fig. 1, Fig. 2, Fig. 3 and Fig. 4) from multiple domains. For different domains, the StyleGAN generator is pre-trained using different datasets [1,4,8]. We conduct various manipulation for the human face domain, including attribute and identity translation. Both results demonstrate the effectiveness of our methods.

E. Video

We summarize the analysis of selection and distillation, the fine-grained control of our method, and more visual samples in a supplementary video.

References

[1] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 2

[2] Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. *arXiv preprint arXiv:2206.09012*, 2022. 1

[3] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019. 1

[4] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 2015. 2

[5] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *CVPR*, pages 2085–2094, 2021. 1

[6] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1

[7] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1

[8] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv:1506.03365*, 2015. 2

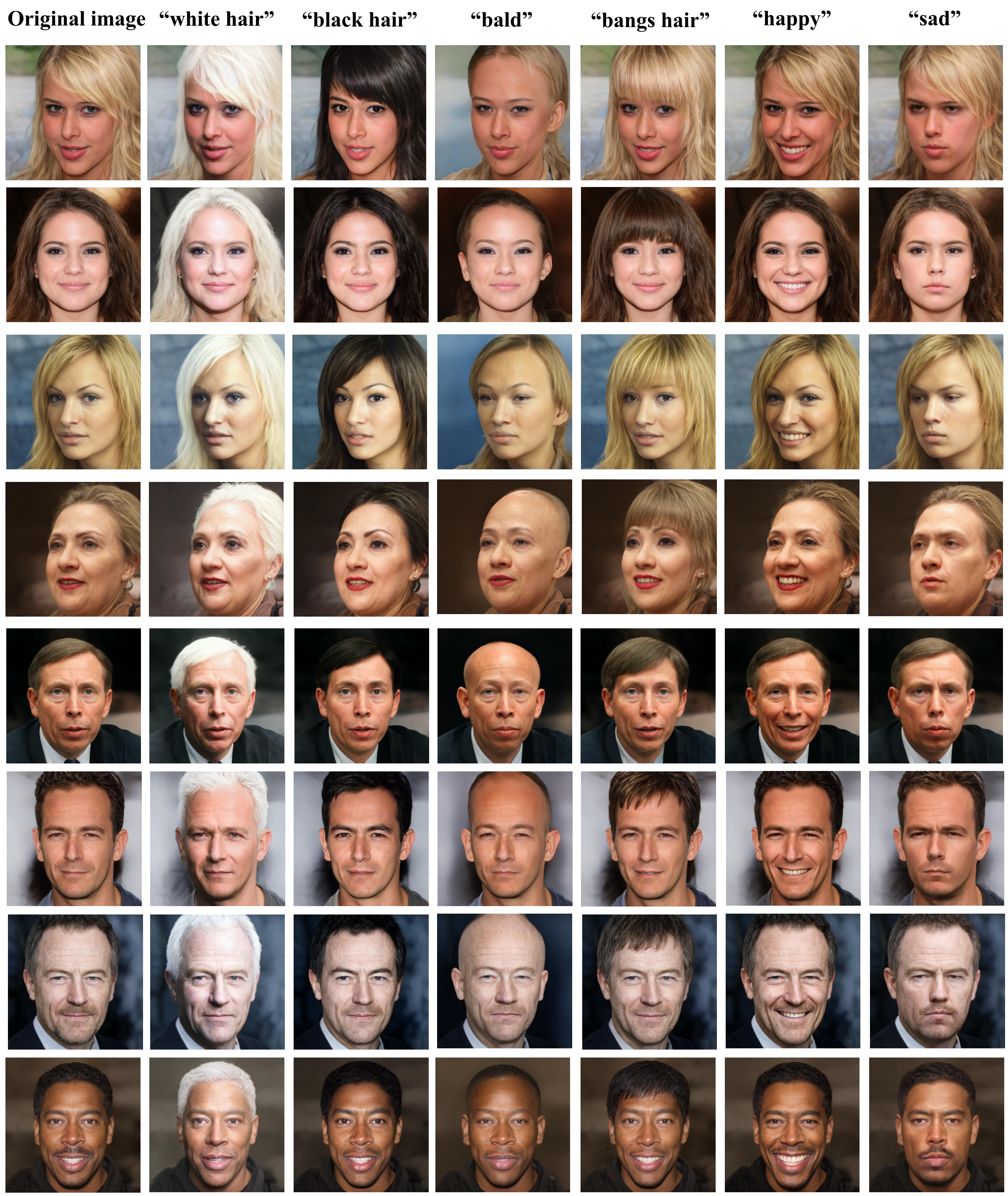

Figure 1. Additional results for SDD face manipulation, in the aspect of hair color, hairstyle, and facial expression

Figure 2. Additional results for SDD face manipulation, in the aspect of attributes addition and celebrities conversion

Figure 3. Additional results for SDD cat face manipulation

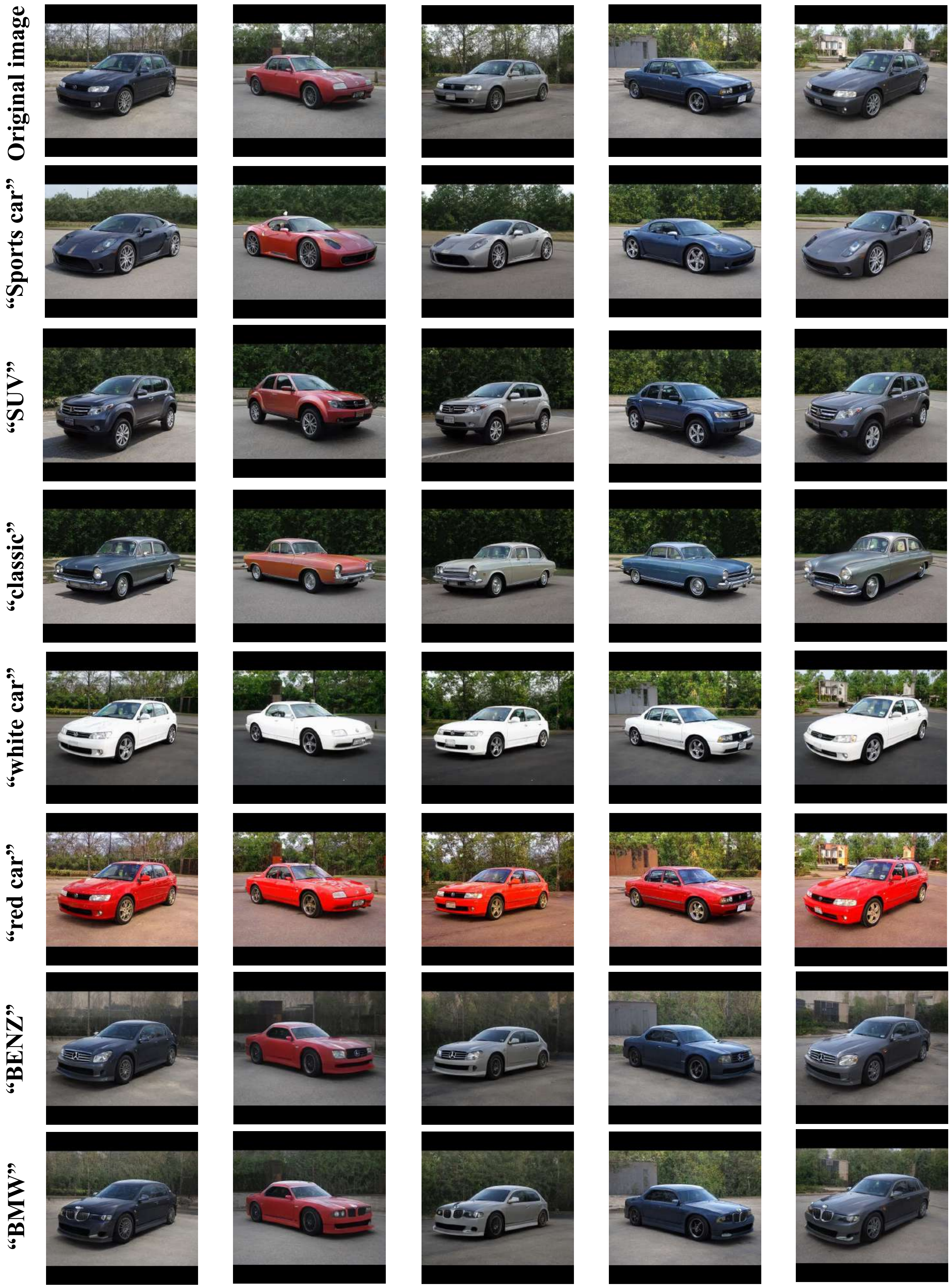

Figure 4. Additional results for SDD car manipulation