Expanding Expressiveness of Diffusion Models with Limited Data via Self-Distillation based Fine-Tuning: Supplementary materials

Jiwan Hur, Jaehyun Choi, Gyojin Han, Dong-Jae Lee, and Junmo Kim
School of Electrical Engineering, KAIST, South Korea
{jiwan.hur, chlwogus, gjhan0820, jhtwosun, junmo.kim}@kaist.ac.kr

1. Connection between DDIBs and Unconditional Image Generation

In this section, we provide a general analysis of Dual Diffusion Implicit Bridges (DDIBs) [9] and their connection to unconditional image generation. We first note that with a proper hyperparameter, a particular sampling process of diffusion model, denoising diffusion implicit model (DDIM) is an ordinary differential equation (ODE) process where the output is deterministically provided given the input noise. Leveraging the DDIM and trained diffusion models, the noise can be translated into the image through the *reverse process* of diffusion models $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$, where $\boldsymbol{x}_t$ is a perturbed image $\boldsymbol{x}_0$ with a time step t. Notably, DDIM can run in the forward direction $p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ to get the noise from the image. Since $p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ is also deterministic in DDIM, the sampled noise can be used as a latent of the image.

Recently, Su et al. [9] propose an unpaired domain translation method, Dual Diffusion Implicit Bridges (DDIBs), which leverages the source diffusion models ϵ^{src} and target diffusion models ϵ^{trg} to get aligned image pairs from source and target domain. The source model ϵ^{src} runs DDIM sampling in the forward direction to get the latent noise $\boldsymbol{x}_T^{src}$ from the input image of the source domain $\boldsymbol{x}_0^{src}$. Then, from the $\boldsymbol{x}_0^{src}$, the target model ϵ^{trg} runs DDIM sampling in the reverse direction to get an image $\boldsymbol{x}_0^{trg}$ from the $\boldsymbol{x}_T^{src}$. Through the above processes, DDIBs prove that $\boldsymbol{x}_0^{trg}$ is semantically aligned with the $\boldsymbol{x}_0^{src}$, without requiring a joint training of source and target pairs.

From the DDIBs, it can be derived that given the same initial noise $\boldsymbol{x}_T$, the source model ϵ^{src} and the target model ϵ^{trg} can generate aligned images, $\boldsymbol{x}_0^{src}$ and $\boldsymbol{x}_0^{trg}$, by running the deterministic DDIM sampling process in the reverse direction. As a result, unconditional generation from the same initial noise can be interpreted as a domain translation between the source and target model, generating semantically aligned images. From this point of view, the proposed method, Self-Distillation-based Fine-Tuning (SDFT), can generate a more diverse and more aligned image with the source model than the Naïve Fine-Tune model, as shown in the main manuscript.

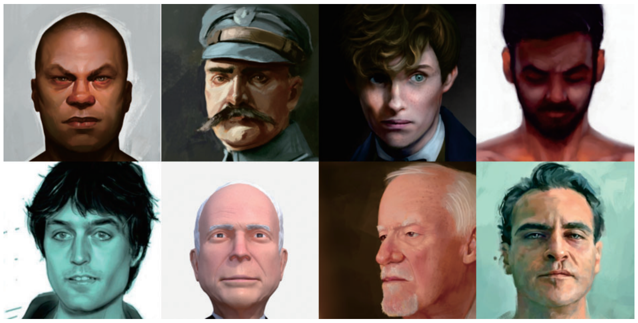

Figure 1. Randomly selected images from a limited AAHQ dataset for the training of diffusion models in the main manuscript.

2. Explanation for Preparing Dataset

We utilize source diffusion models pretrained on FFHQ, which has 70K diverse real faces with various attributes. For the target limited datasets, we utilize MetFaces [2], which has 1,336 high-quality portraits. Due to the limited samples and inherent biases, MetFaces do not or scarcely contain diverse facial attributes (e.g. smiling with teeth, sunglasses, various hairstyles *etc.*). We further exclude 10 samples which include *glasses* from the MetFaces for the more challenging scenario. For another target dataset using the same source dataset, we utilize AAHQ [5] which contains 25k high-quality artistic faces. However, AAHQ is a sufficiently large and diverse dataset. To simulate the limited, biased dataset, we select images from AAHQ using CLIP [8] following the nie et al. [7]. Specifically, we measure the cosine similarity of embedding vectors of CLIP between AAHQ images and the prompt *"A realistic painting of an expressionless man without glasses "*. Then by thresholding the similarity and manually excluding some misclassified images, we get 1,437 limited AAHQ images. As described in nie et al. [7], using CLIP similarity for preparing datasets is efficient and effective for selecting datasets based on the natural language without labor-intensive work. Fig. 1 shows the randomly selected images from a limited AAHQ dataset. Note that even though it contains a similar number of images as MetFaces, it contains more various images such as various skin colors, and various emotions. For example, MetFaces contains the faces of medieval works of art, so most samples consistently show similar styles of painting and the style of painting of the time, such as smiling a little in most samples. However, despite using CLIP to select a limited and biased dataset in AAHQ, it lacks biases such as picture style and expressions.

3. Training Details

We provide additional details for training diffusion models in the main manuscript. We use the lighter version of ADM [1] for the baseline model in all experiments and use default settings. To implement the SDFT, we use 4 hyperparameters that define the in Tab. 1. Note that since the output from the auxiliary input drastically collapses as the timestep increases, we set higher γ^{aux} for all experiments. The visualization of $w(t)$ used for SDFT according to different gamma values is provided in Fig. 2.

Datasets	MetFaces	AAHQ
$\lambda^{distill}$	0.1	0.1
λ^{aux}	0.1	0.3
$\gamma^{distill}$	3	50
γ^{aux}	3	50

Table 1. Hyperparameters for SDFT training.

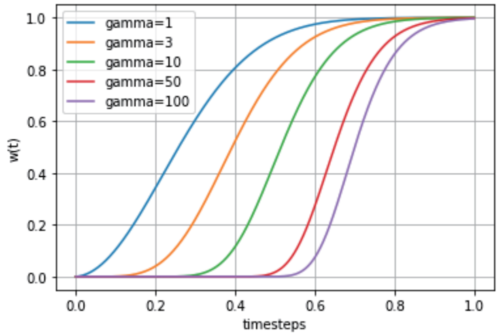

Figure 2. Different weightings according to hyperparameter γ.

4. Experimental Details in Domain Translation

For all domain translation methods [6, 10], rather than editing in all diffusion time steps, authors used partial editing time steps for the best trade-off between *realistic* and *faithful*. Note that successful domain translation should be *realistic* to fit the style of the target domain and *faithful* to ensure that the various attributes from the input image are accurately preserved. The larger the editing time steps, the more *realistic* the translated outputs, but the less *faithful* it becomes, losing the crucial attributes from the source images. As a result, various domain translation methods generally adopt an editing range of $0.5T$ in their experiments for the best trade-off between *realism* and *faithfulness*. Following this, in the main manuscript, we adopt an editing range of $0.5T$ for all experiments on MetFaces. For AAHQ experiments, we found that an editing range of $0.5T$ is less *realistic*, thus we increase the editing range into $0.625T$ for all experiments on AAHQ. For the EGSDE [10], we train domain classifier for domain-independent energy function using 10K FFHQ [3] samples and each training datasets (entire MetFaces dataset or limited AAHQ dataset used for training of diffusion models). We used the official training code provided by the author, with 1500 and 4500 training iterations, respectively. For the domain-specific energy function, we use a downsampler with a downsampling factor of 32. For the comprehensive comparison, we provide more results of domain translation in Figs. 3 and 4.

5. Experimental Details in Text-Guided Image Manipulation

In the main manuscript, we utilize Asyrp [4] for the text-guided image manipulation method. We use officially provided training and inference code to implement Asyrp and use training step 50 and inversion step 100 in all experiments. We utilize 100 training images from the MetFaces and AAHQ for Asyrp and the training epoch is set to 5. However, we found that in some facial expressions such as *sad, angry*, and *old*, training epoch 5 results in a bad bias that the manipulated faces get too old. As a result, we set the training epoch as 3 for these facial expressions. The editing range is set to $0.5T$. For quantitative evaluation, we use 500 and 400 test images for MetFaces and AAHQ, respectively, and use 5 scripts (*smiling, sad, angry, young and old*). For the comprehensive comparison, we provide more results of text-guided translation in Figs. 5 and 6 using text guidance *smiling, sad, angry, young, old, man*, and *woman*. Overall, ours(SDFT) outperforms for the various expressions, even though the input image is not close to the real human face, while successfully preserving the identities. We note that the provided results are not included in the training datasets for Asyrp.

6. More Results on Unconditional Image Generation

In Fig. 7, we provide more samples from unconditional image generation for the comprehensive comparison. As described in Sec. 1, the images generated from the same initial noise have aligned semantics. Since SDFT can preserve more diverse information from the source model, SDFT can

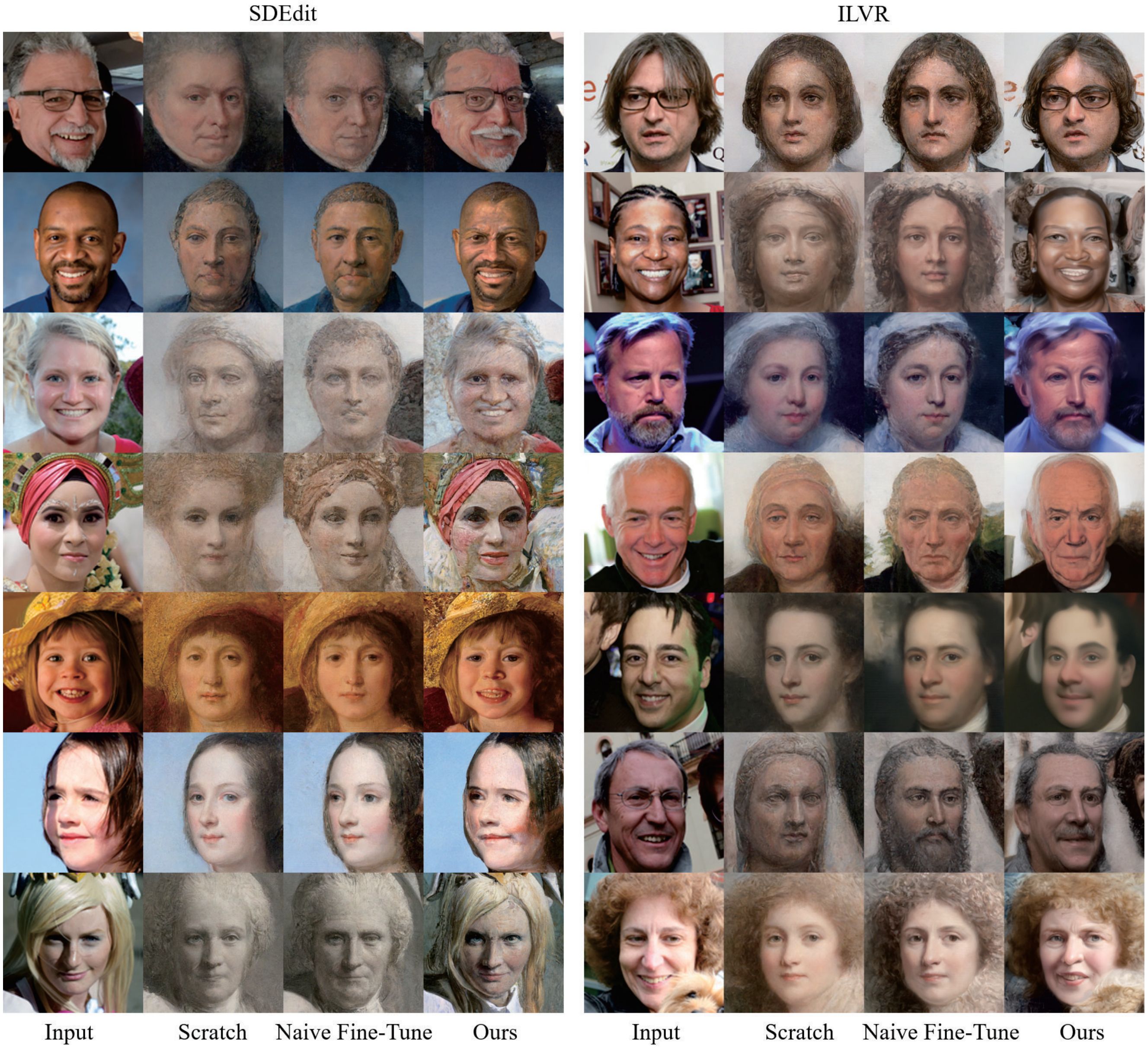

Figure 3. Domain translation outputs using SDEdit and ILVR with fine-tuned diffusion models.

generate more semantically aligned, diverse images from unconditional image generation.

References

[1] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2

[2] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020. 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*, pages 4401–4410, 2019. 2

[4] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022. 2

[5] Mingcong Liu, Qiang Li, Zekui Qin, Guoxin Zhang, Pengfei Wan, and Wen Zheng. Blendgan: Implicitly gan blending for arbitrary stylized face generation. *Advances in Neural Information Processing Systems*, 34:29710–29722, 2021. 1

[6] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-

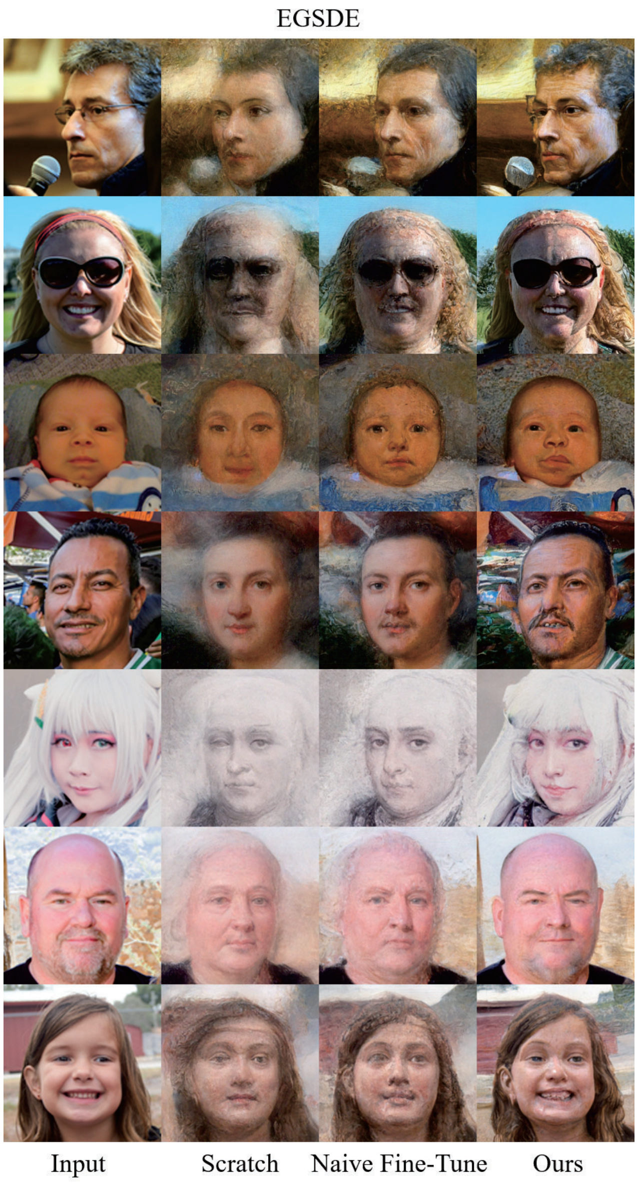

Figure 4. Domain translation outputs using EGSDE with fine-tuned diffusion models.

Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2

[7] Weili Nie, Arash Vahdat, and Anima Anandkumar. Controllable and compositional generation with latent-space energy-based models. *Advances in Neural Information Processing Systems*, 34:13497–13510, 2021. 1

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[9] Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. *arXiv preprint arXiv:2203.08382*, 2022. 1

[10] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *arXiv preprint arXiv:2207.06635*, 2022. 2

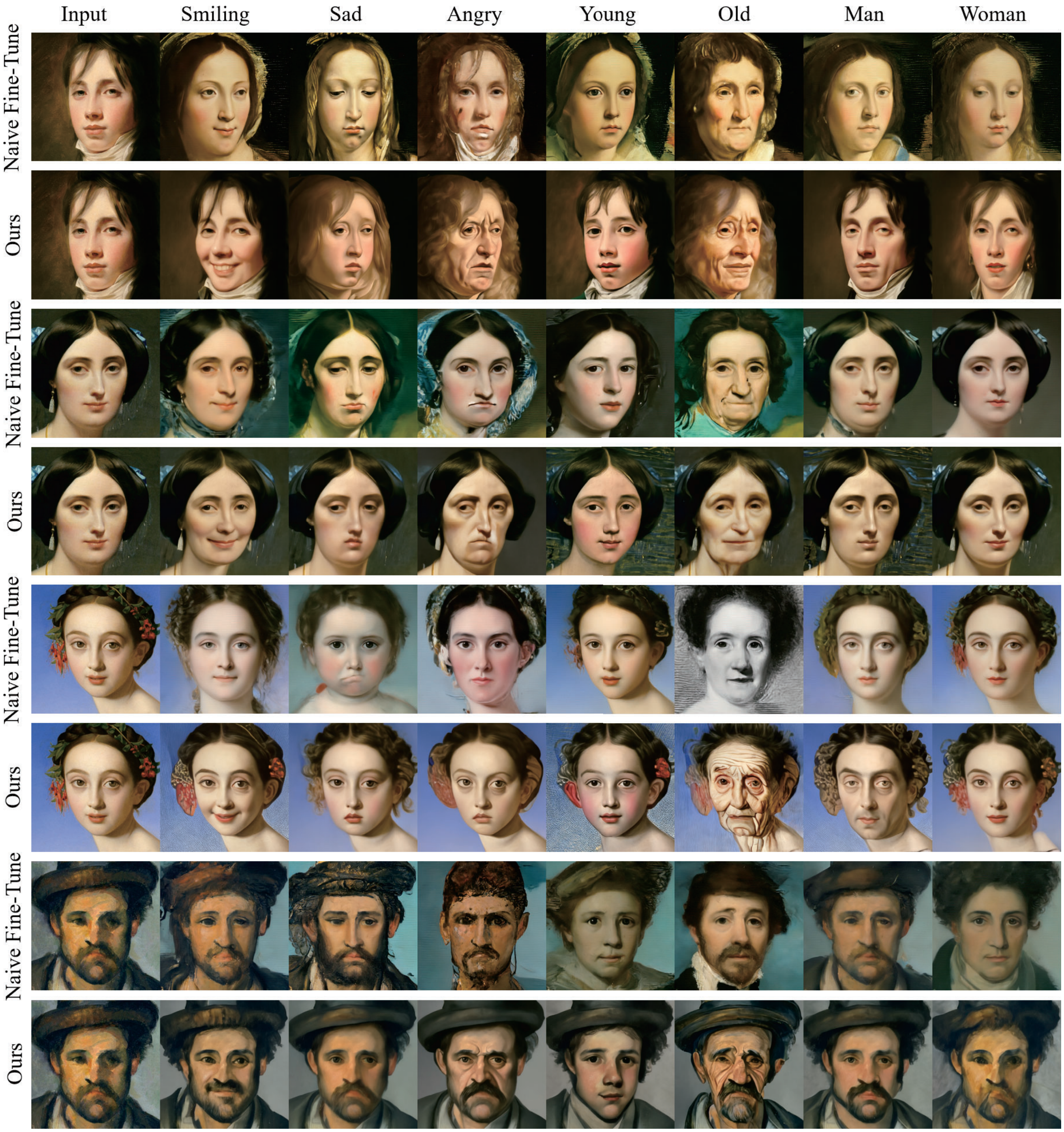

Figure 5. Various results of text-guided image manipulation, where the input is close to the human face.

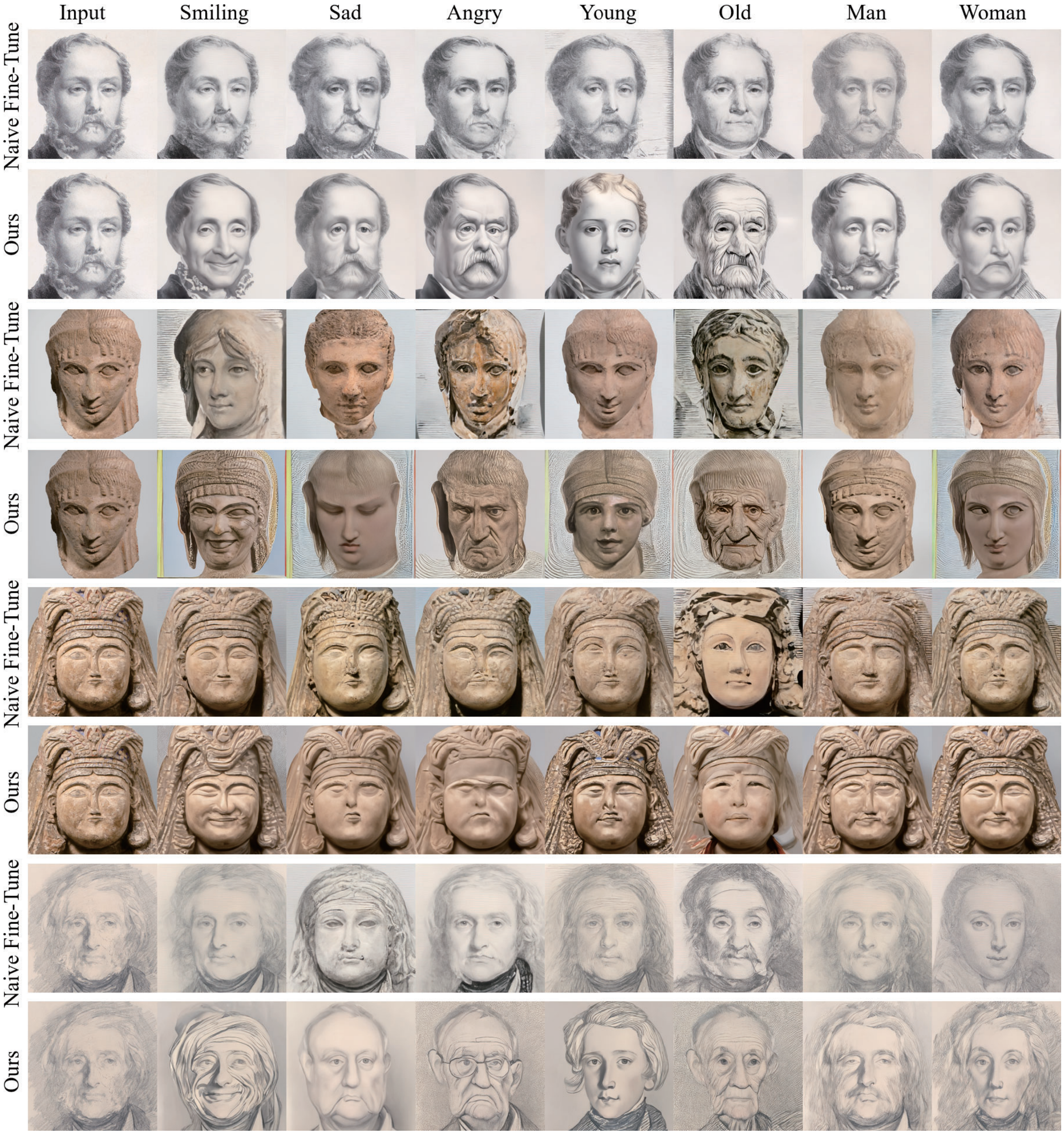

Figure 6. Various results of text-guided image manipulation, where the input is not relatively close to the human face.

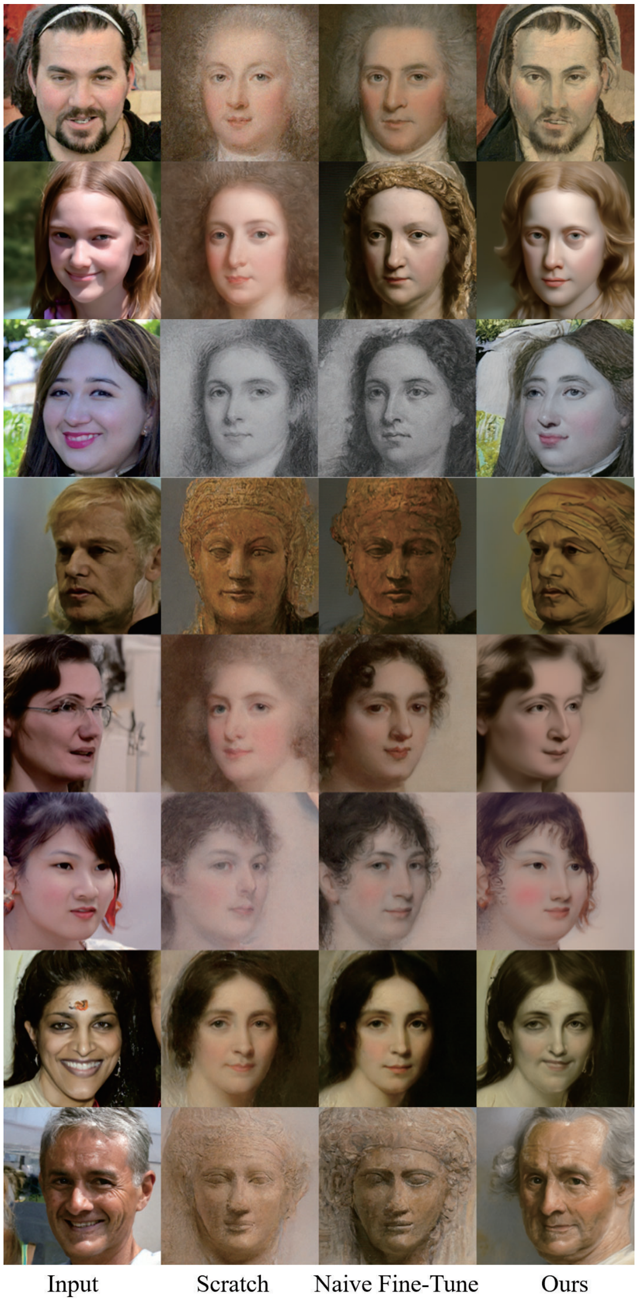

Figure 7. Various results of unconditional image generation. Images in each row are generated from the same initial noise.