Supplementary Materials for ECCV 2024 paper Few-Shot Image Generation by Conditional Relaxing Diffusion Inversion

Yu Cao and Shaogang Gong

Queen Mary University of London, UK
{yu.cao, s.gong}@qmul.ac.uk

A Target Domain Samples

We first display the given samples across all three target domains, Babies [5], Sunglasses [5] and MetFaces [3], in Fig. 1.

Fig. 1: Given samples across three target domains: Babies, Sunglasses and MetFaces

B Additional Components Analysis

In the main paper, we conduct a quantitative analysis of the impact of removing the core components, namely SGE and Noise Perturbation, on model performance. Here, we provide a qualitative analysis of the removal of these components and their effects on model generation. As shown in Figure 2, omitting each process significantly influences the ability of the model to generate images. Without SGE and Noise Perturbation, our model effectively becomes an unconditional diffusion model (first row). If we replace SGE with minimal noise (second row), the generated images appear to have more details, however, most of these details are illogical. Removing the noise perturbation entirely (third row) reduced the process to the same setting as reconstruction.

Table 1: Evaluating the impact of removing each components of loss function by calculating the FID(↓) across three target domains. x_0, x_t and penalty correspond to the three terms in the Loss function (Eq. 1), respectively.

Loss Components			FID Score (↓)		
x_0	x_t	penalty	Babies	Sunglasses	MetFaces
✗	✓	✗	360.08	322.67	327.07
✗	✓	✓	273.80	132.13	203.27
✓	✗	✗	54.76	27.90	105.09
✓	✗	✓	54.46	27.94	124.63
✓	✓	✗	49.62	26.43	97.70
✓	✓	✓	**48.52**	**24.62**	**94.86**

Furthermore, We also provide a quantitative analysis of the impact of removing the components of loss function, given by:

$$\mathcal{L} = \|x_0 - x_0'\|^2 + \|x_{t-1} - x_{t-1}'\|^2 + ||G_\theta^i - \frac{1}{N}\sum_{j=1}^{N} G_\theta^j||^2 \tag{1}$$

We evaluated the impact of removing each components by calculating the FID scores across three target domains, quantitative results are shown in Tab. 1.

Fig. 2: Visual Examples for the impact of removing SGE and noise perturbation (Ptb.)

C Foundation Model Based Adaptation Method

In the main paper, we conduct a quantitative analysis of the foundation model based adaptation method applied on FSIG. Here we provide a qualitative analysis and implementation detail. The given samples across three target domains are shown in Fig. 1, the synthetic samples are shown in Fig. 3. DreamBooth [6]

and Textual-Inversion [1], as subject-level adaptation methods, generate samples that only semantically align with the target domain, exhibiting significant out-of-distribution issues. This problem is particularly pronounced in the MetFaces domain. Conversely, fine-tuning methods for large models, LoRA [2], demonstrate severe overfitting, resulting in generated samples lacking diversity. For implementation, we followed the official tutorials provided by the original papers, using basic prompts without extensive experimentation. The prompt we used are: "A facial image of a baby", "A facial image of a person wearing sunglasses" and "A photo of human face art painting".

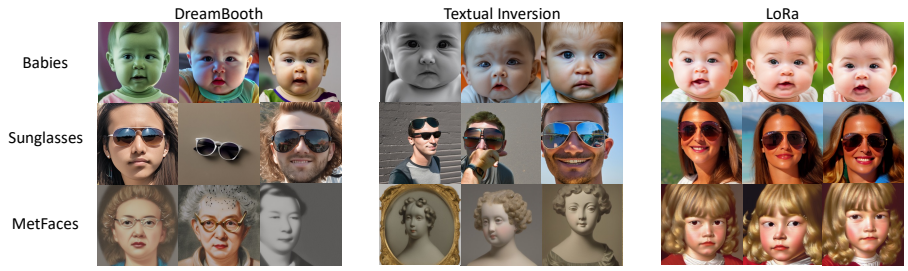

Fig. 3: Visual Examples for Foundation Model Based Adaptation Methods across three target domains: Babies, Sunglasses and MetFaces.

D Synthesis Images across Three Target Domains

In the main paper, we provided 10 synthesized images for each method as illustrative examples. This supplementary document presents an expanded visual comparison, showcasing additional images across three target domains: Babies, Sunglasses, and MetFaces. We compare our approach with two SOTA methods: RICK [7]and GenDA [4], which are representative of the fine-tuning and representation learning strategies, respectively. We present sample images from these three target domains (Ground Truth) in Fig. 1. We further demonstrate the capabilities of our method alongside RICK and GenDA through synthesized images for each domain, displayed in Figs. 4 and 5 for the Babies domain, Figs. 6 and 7 for Sunglasses, and Figs. 8 and 9 for MetFaces, thereby providing a more comprehensive visual comparison between the methods across varied domains.

Continuing from the previous discussion, we observe that for the fine-tuning methods, a common issue across the three target domains is the presence of artifacts and overfitting, with many identical images being repeatedly generated. For the representation learning methods, a shared challenge across these domains is their inability to ensure controllability over the generated samples. Specifically, in the Babies and Sunglasses domains, many samples from the source domain are produced. For MetFaces domain, fine-tuning methods, they are only able to generate a limited range of sub-domains of the MetFaces domain, failing to produce

any samples featuring crowns or ceramic. As for representation learning, the gap between the generated samples and the target domain becomes even more pronounced. In contrast, our method, while also inevitably producing some artifacts in the MetFaces domain, substantially preserves the domain-specific characteristics. We have generated images for every sub-domain provided. We believe this property is preferable in real-world applications because FSIG essentially offers a biased estimation for a target domain based on given samples. Different biases can lead to significant variations in FID scores; however, a method that maintains consistency with the given samples is safer and more logical. This adherence to sample consistency, despite potential biases, provides a more reliable and controlled approach to generating images within a specific domain.

Fig. 4: Comparison with GenDA (Left) on Target Domain Babies.

Fig. 5: Comparison with RICK (Left) on Target Domain Babies.

Fig. 6: Comparison with GenDA (Left) on Target Domain Sunglasses.

Fig. 7: Comparison with RICK (Left) on Target Domain Sunglasses.

Fig. 8: Comparison with GenDA (Left) on Target Domain MetFaces.

Fig. 9: Comparison with RICK (Left) on Target Domain MetFaces.

References

1. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
2. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
3. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. Advances in neural information processing systems **33**, 12104–12114 (2020)
4. Mondal, A.K., Tiwary, P., Singla, P., Prathosh, A.: Few-shot cross-domain image generation via inference-time latent-code learning. In: The Eleventh International Conference on Learning Representations (2022)
5. Ojha, U., Li, Y., Lu, J., Efros, A.A., Lee, Y.J., Shechtman, E., Zhang, R.: Few-shot image generation via cross-domain correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10743–10752 (2021)
6. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
7. Zhao, Y., Du, C., Abdollahzadeh, M., Pang, T., Lin, M., Yan, S., Cheung, N.M.: Exploring incompatible knowledge transfer in few-shot image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7380–7391 (2023)