EFFICIENT TRANSFER LEARNING IN DIFFUSION MODELS VIA ADVERSARIAL NOISE

Xiyu Wang[1], **Baijiong Lin**[2], **Daochang Liu**[1], **Ying-Cong Chen**[2], **Chang Xu**[1]
[1] School of Computer Science, Faculty of Engineering, The University of Sydney
[2] The Hong Kong University of Science and Technology (Guangzhou)
{xiyuwang.usyd, bj.lin.email}@gmail.com,
{daochang.liu, c.xu}@sydney.edu.au, yingcongchen@ust.hk

ABSTRACT

Diffusion Probabilistic Models (DPMs) have demonstrated substantial promise in image generation tasks but heavily rely on the availability of large amounts of training data. Previous works, like GANs, have tackled the limited data problem by transferring pre-trained models learned with sufficient data. However, those methods are hard to be utilized in DPMs since the distinct differences between DPM-based and GAN-based methods, showing in the unique iterative denoising process integral and the need for many timesteps with no-targeted noise in DPMs. In this paper, we propose a novel DPMs-based transfer learning method, TAN, to address the limited data problem. It includes two strategies: similarity-guided training, which boosts transfer with a classifier, and adversarial noise selection which adaptively chooses targeted noise based on the input image. Extensive experiments in the context of few-shot image generation tasks demonstrate that our method is not only efficient but also excels in terms of image quality and diversity when compared to existing GAN-based and DDPM-based methods.

1 INTRODUCTION

Generative models like as GANs (Brock et al., 2018), VAEs (Kingma & Welling, 2013), and autoregressive models (Van den Oord et al., 2016) have achieved significant success in images (Brock et al., 2018), text (Brown et al., 2020), and audio (Dhariwal et al., 2020), leveraging large amounts of unlabeled data. Diffusion probabilistic models (DPMs) (Sohl-Dickstein et al., 2015; Ho et al., 2020; Nichol & Dhariwal, 2021) have advanced in producing high-quality, diverse images but depend heavily on extensive data, risking overfitting with limited datasets. Transfer learning offers a solution by applying knowledge from models trained on large datasets to smaller ones, adapting GAN-based models to new domains with few samples (Wang et al., 2018; Karras et al., 2020a; Wang et al., 2020; Li et al., 2020). Techniques include data augmentation to prevent overfitting or measuring distances between images to ensure quality (Zhang et al., 2018; Karras et al., 2020a). This approach addresses data scarcity and enhances model versatility across different content types (Ojha et al., 2021; Zhao et al., 2022).

Applying GAN techniques to DPMs faces challenges due to their distinct training methods. GANs generate clean images in one step (Li et al., 2020; Ojha et al., 2021; Zhao et al., 2022), while DPMs iteratively predict less noisy images, requiring more steps for a high-quality output. This iterative process complicates model transfer, with two main issues: estimating transfer direction on noisy images and dealing with DPMs' non-targeted noise, which can affect images unevenly, causing some to overfit. The DDPM pairwise adaptation method attempts to address this by using blurry predicted images for comparison, but this can lead to inaccuracies. Furthermore, the random Gaussian noise used in diffusion models presents additional challenges, potentially requiring numerous iterations for effective transfer, especially with limited training data. As demonstrated in Figure 1, when one image (below) is just successfully transferred from the source domain to the target domain, another image (above) may have severely overfit and become too similar to the target image. Such normally distributed noise may also necessitate an extensive number of iterations to transfer, especially when the gradient direction is noisy due to limited images.

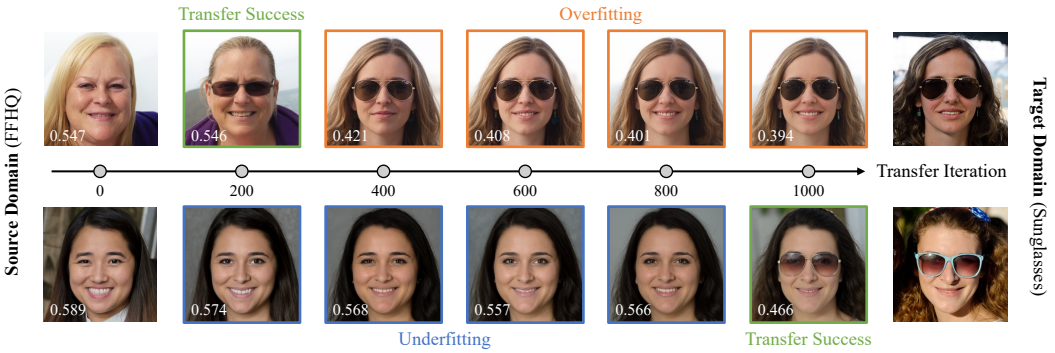


Figure 1: Two sets of images generated from corresponding fixed noise inputs at different stages of fine-tuning DDPM from FFHQ to 10-shot Sunglasses. The perceptual distance (LPIPS Zhang et al. (2018)) with the training target image is shown on each generated image. When the bottom image successfully transfers to the target domain, the top image has already suffered from overfitting.

To address direction estimation challenges in diffusion model transfers, we propose using a similarity measurement to bridge the source-target gap, introducing a ***similarity-guided training*** method. This approach, which fine-tunes pre-trained models for target domains, uses a classifier to measure divergence, leveraging source domain knowledge. It overcomes the issues of unstable gradients in few-shot settings and non-targeted noise by implicitly comparing target data with source data. Additionally, we introduce ***adversarial noise selection*** for min-max training, selecting noise based on the input image to minimize denoising failures. This method speeds up training, ensures faster convergence, and achieves high-quality, style-consistent image generation in few-shot tasks, outperforming existing GAN and DDPM techniques.

2 Preliminary

Gaussian diffusion models are used to approximate the data distribution $x_0 \sim q(x_0)$ by $p_\theta(x_0)$. The distribution $p_\theta(x_0)$ is modeled in the form of latent variable models. According to (Ho et al., 2020), the diffusion process from a data distribution to a Gaussian distribution with variance $\beta_t \in (0, 1)$ for timestep t can be expressed as:

$$q(x_t|x_0) = \mathcal{N}(x_t; \bar{\alpha}_t x_0, (1-\bar{\alpha}_t)\mathbf{I}),\ x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon\,,$$

where $\alpha_t := 1 - \beta_t$, $\bar{\alpha}_t := \prod_{i=0}^{t}(1-\beta_i)$ and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Ho et al. (2020) trains a U-Net (Ronneberger et al., 2015) model parameterized by θ to fit the data distribution $q(x_0)$ by maximizing the variational lower-bound. The DDPM training loss with model $\epsilon_\theta(x_t, t)$ can be expressed as:

$$\mathcal{L}_{\texttt{sample}}(\theta) := \mathbb{E}_{t,x_0,\epsilon}\left\|\epsilon - \epsilon_\theta(x_t, t)\right\|^2. \tag{1}$$

Based on (Song et al., 2020), the reverse process of DDPM and DDIM at timestep t can be expressed as:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\underbrace{\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t,t)}{\bar{\alpha}_t}\right)}_{\text{predicted } x_0} + \underbrace{\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\cdot\epsilon_\theta(x_t,t)}_{\text{direction pointing to } x_t} + \underbrace{\sigma_t\epsilon_t}_{\text{random noise}}\,,$$

where $\sigma_t = \eta\sqrt{(1-\bar{\alpha}_{t-1})/(1-\bar{\alpha}_t)}\sqrt{1-\bar{\alpha}_t/\bar{\alpha}_{t-1}}$ and $\eta = 0$ (Song et al., 2020) or $\eta = 1$ (Ho et al., 2020) or $\eta = \sqrt{(1-\bar{\alpha}_t)/(1-\bar{\alpha}_{t-1})}$ (Ho et al., 2020). Enhance, Dhariwal & Nichol (2021) propose the conditional reverse noise process as:

$$p_{\theta,\phi}(x_{t-1}|x_t, y) \approx \mathcal{N}(x_{t-1}; \mu_\theta(x_t,t) + \sigma_t^2\gamma\nabla_{x_t}\log p_\phi(y|x_t), \sigma_t^2\mathbf{I}), \tag{2}$$

$$\text{where}\quad \mu_\theta(x_t,t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t,t)\right), \tag{3}$$

and γ is a hyperparameter for conditional control. For the sake of clarity in distinguishing domains, this paper uses $\mathcal{S}$ and $\mathcal{T}$ to represent the source and target domain, respectively.

3 Transfer Learning in Diffusion Models via Adversarial Noise

In this section, we introduce transfer learning in diffusion models via Adversarial Noise, dubbed TAN, with similarity-guided training and adversarial noise selection for stronger transfer ability.

3.1 SIMILARITY-GUIDED TRAINING

We use similarity to measure the gap between the source and target domains using a noised image x_t at timestep t instead of the final image. Drawing inspiration from (Dhariwal & Nichol, 2021; Liu et al., 2023), we express the domain difference between the source and target in terms of the divergence in similarity measures. Initially, we assume a model that can predict noise with both source and target domains, denoted as $\theta_{(\mathcal{S},\mathcal{T})}$. As Equation 2, the reverse process for the source and target images can be written as:

$$p_{\theta_{(\mathcal{S},\mathcal{T})},\phi}(x_{t-1}|x_t, y = Y) \approx \mathcal{N}(x_{t-1}; \mu_{\theta_{(\mathcal{S},\mathcal{T})}} + \sigma_t^2 \gamma \nabla_{x_t} \log p_\phi(y = Y|x_t), \sigma_t^2 \mathbf{I}) \,, \tag{4}$$

where Y is $\mathcal{S}$ or $\mathcal{T}$ for source or target domain image generation, respectively. We can consider $\mu(x_t) + \sigma_t^2 \gamma \nabla_{x_t} \log p_\phi(y = \mathcal{S}|x_t)$ as the source model $\theta_\mathcal{S}$, which only synthesize image on the source domain respectively. For brevity, we denote $p_{\theta_\mathcal{S},\phi}(x_{t-1}^\mathcal{S}|x_t) = p_{\theta_{(\mathcal{S},\mathcal{T})},\phi}(x_{t-1}|x_t, y = \mathcal{S})$. We similar define $p_{\theta_\mathcal{T},\phi}(x_{t-1}^\mathcal{T}|x_t)$ as above by replace $\mathcal{S}$ with $\mathcal{T}$. Therefore, the KL-divergence between the output of source model $\theta_\mathcal{S}$ and the target $\theta_\mathcal{T}$ with the same input x_t at timestep t, is defined as:

$$\begin{aligned} & \mathrm{D}_{\mathrm{KL}} \left(p_{\theta_\mathcal{S},\phi}(x_{t-1}^\mathcal{S}|x_t), p_{\theta_\mathcal{T},\phi}(x_{t-1}^\mathcal{T}|x_t) \right) \\ = \; & \mathbb{E}_{t,x_0,\epsilon} \left[\left\| \nabla_{x_t} \log p_\phi(y = \mathcal{S}|x_t) - \nabla_{x_t} \log p_\phi(y = \mathcal{T}|x_t) \right\|^2 \right], \end{aligned} \tag{5}$$

where p_ϕ is a classifier to distinguish x_t. The detailed derivation is in Appendix. We consider the $\nabla_{x_t} \log p_\phi(y = \mathcal{S}|x_t)$ and $\nabla_{x_t} \log p_\phi(y = \mathcal{T}|x_t)$ as the similarity measures of the given x_t in the source and target domains, respectively. Since transfer learning primarily focuses on bridging the gap between the image generated by the current fine-tuning model and the target domain image, we disregard the first term and utilize only $p_\phi(y = \mathcal{T}|x_t^\mathcal{T})$ to guide the training process. Specifically, we employ a fixed pre-trained binary classifier that differentiates between source and target images at time step t to boost the training process. Similarly with the vanilla training loss in DPMs (Ho et al., 2020), i.e., Equation 1, we use the KL-divergence between the output of current model θ and target model $\theta_\mathcal{T}$ at time step t as:

$$\min_\theta \mathbb{E}_{t,x_0,\epsilon} \left[\left\| \epsilon_t - \epsilon_\theta(x_t, t) - \hat{\sigma}_t^2 \gamma \nabla_{x_t} \log p_\phi(y = \mathcal{T}|x_t) \right\|^2 \right] \,, \tag{6}$$

where $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, ϵ_θ is the pre-trained neural network on source domain, γ is a hyper-parameter to control the similarity guidance, $\hat{\sigma}_t = (1 - \bar{\alpha}_{t-1}) \sqrt{\frac{\alpha_t}{1-\bar{\alpha}_t}}$, and p_ϕ is the binary classifier differentiating between source and target images. Equation 6 is defended as similarity-guided DPMs train loss. The full proof is provided in the Appendix. We leverage the pre-trained classifier to indirectly compare the noised image x_t with both domain images, subtly expressing the gap between the currently generated image and the target image.

3.2 ADVERSARIAL NOISE SELECTION

Despite potentially determining the transfer direction, we still encounter a fundamental second challenge originating from the noise mechanism in diffusion models. As mentioned, the model needs to be trained to accommodate the quantity of noise ϵ_t over many iterations. However, increasing iterations with limited images may lead to overfitting of the training samples, thereby reducing the diversity of the generated samples. On the other hand, training with too few iterations might only successfully transform a fraction of the generated images into the target domain as Figure 1.

To counter these issues, we propose an adaptive noise selection method. This approach utilizes a min-max training process to reduce the actual training iterations required and ensure the generated images closely resemble the target images. After the model has been trained on a large dataset, it exhibits a strong noise reduction capability for source datasets. This implies it only needs to minimize specific types of Gaussian noise with which the trained model struggles or fails to denoise with the target domain sample. The first step in this process is to identify the maximum Gaussian noise with the current model, and then specifically minimize the model using this noise. Based on Equation 6, this can be mathematically formulated as follows:

$$\min_\theta \max_\epsilon \mathbb{E}_{t,x_0} \left[\left\| \epsilon - \epsilon_\theta(x_t, t) - \hat{\sigma}_t^2 \gamma \nabla_{x_t} \log p_\phi(y = \mathcal{T}|x_t) \right\|^2 \right] \,. \tag{7}$$

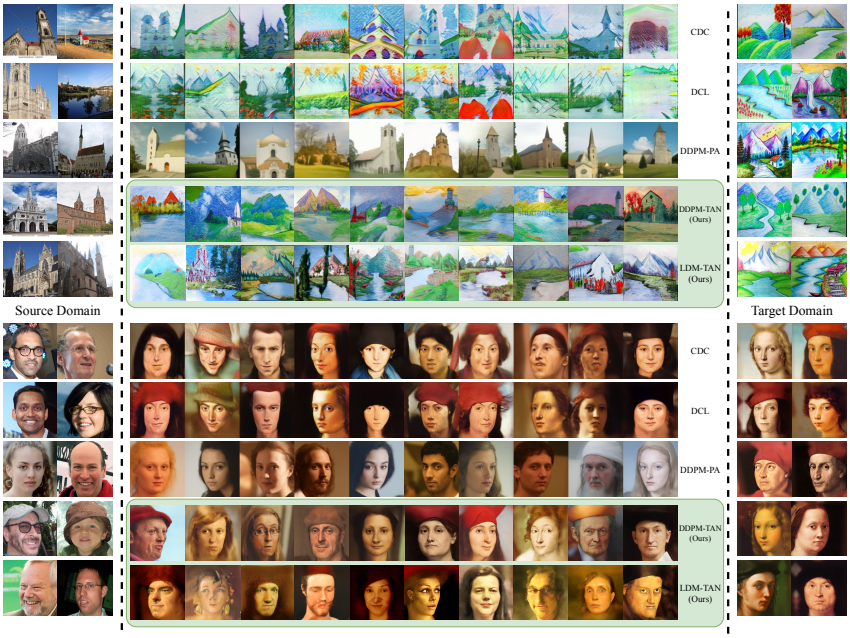


Figure 2: The 10-shot image generation samples on LSUN Church → Landscape drawings (top) and FFHQ → Raphael's paintings (bottom). When compared with other GAN-based and DDPM-based methods

Although finding the exact maximum noise is challenging as Equation 7, the projected gradient descent (PGD) strategy can be used to solve the inner maximization problem instead. Specifically, the inner maximization of Gaussian noise can be interpreted as finding the "worse-case" noise corresponding to the current neural network. Practically, the similarity-guided term is disregarded, as this term is hard to compute differential and is almost unchanged in the process. We utilize the multi-step variant of PGD with gradient ascent, as expressed below:

$$\epsilon^{j+1} = \text{Norm}\left(\epsilon^j + \omega \nabla_{\epsilon^j} \left\|\epsilon^j - \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon^j, t)\right\|^2\right), \quad j = 0, \cdots, J-1, \tag{8}$$

where ω is a hyperparameter that represents the "learning rate" of the negative loss function, and Norm is a normalization function that approximately ensures the mean and standard deviation of ϵ^{j+1} is $\mathbf{0}$ and $\mathbf{I}$ respectively. The initial value, ϵ_0, is sampled from the Gaussian distribution as $\epsilon_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We use this method to identify this worse-case noise and minimizing the "worse-case" Gaussian noise is akin to minimizing all Gaussian noises that are "better" than it.

4 Experiment

To demonstrate the effectiveness of our approach, we perform a series of few-shot image generation experiments using a limited set of just 10 training images with the same setting as DDPM-PA (Zhu et al., 2022). We compare our method against state-of-the-art GAN-based and DDPM-based techniques, assessing the quality and diversity of the generated images through both qualitative and quantitative evaluations. This comprehensive comparison provided strong evidence of the superiority of our proposed method in the context of few-shot image generation tasks.

4.1 Experimental Setup

Datasets. Following (Ojha et al., 2021), we use FFHQ (Karras et al., 2020b) and LSUN Church (Yu et al., 2015) as source datasets. For the target datasets, we employe 10-shot Sketches, Babies, Sunglasses, and face paintings by Amedeo Modigliani and Raphael Peale, which correspond to the source domain FFHQ. Additionally, we utilize 10-shot Haunted Houses and Landscape drawings as target datasets corresponding to the LSUN Church source domain.

Measurements. In our diversity assessment, we use Intra-LPIPS and FID metrics as outlined in CDC (Ojha et al., 2021). We compute Intra-LPIPS by generating 1,000 images, assigning each to the

Table 1: The Intra-LPIPS (↑) results for both DDPM-based strategies and GAN-based baselines are presented for 10-shot image generation tasks. These tasks involve adapting from the source domains of FFHQ and LSUN Church. The "Parameter Rate" column provides information regarding the proportion of parameters fine-tuned in comparison to the pre-trained model's parameters. The best results are marked as **bold**.

Methods	Parameter Rate	FFHQ → Babies	FFHQ → Sunglasses	FFHQ → Raphael's paintings	LSUN Church → Haunted houses	LSUN Church → Landscape drawings
CDC	100%	0.583±0.014	0.581±0.011	0.564±0.010	0.620±0.029	0.674±0.024
DCL	100%	0.579±0.018	0.574±0.007	0.558±0.033	0.616±0.043	0.626±0.021
DDPM-PA	100%	0.599±0.024	0.604±0.014	0.581±0.041	0.628±0.029	0.706±0.030
DDPM-TAN (Ours)	1.3%	0.592±0.016	0.613±0.023	**0.621**±0.068	0.648±0.010	0.723±0.020
LMD-TAN (Ours)	1.6%	**0.601**±0.018	**0.613**±0.011	0.592±0.048	**0.653**±0.010	**0.738**±0.026

nearest training sample via LPIPS distance, then averaging the distances within and across clusters. FID, which measures generative model quality by comparing sample and dataset distributions, can be less reliable for small datasets like our 10-shot cases. Therefore, we follow the DDPM-PA method, applying FID on larger datasets (Sunglasses and Babies) containing 2,500 and 2,700 images, respectively, to ensure stability.

Baselines. To adapt pre-trained models to target domains using a limited number of samples, we compare our work with several GAN-based and DDPMs baselines that share similar objectives. These include CDC (Ojha et al., 2021), DCL (Zhao et al., 2022), and DDPM-PA (Zhu et al., 2022). All these methods are implemented on the same StyleGAN2 (Karras et al., 2020b) codebase.

4.2 Overall Performance

Qualitative Evaluation. Figure 2 showcases samples from GAN and DDPM methods for 10-shot image generation tasks: LSUN Church to Landscape drawings and FFHQ to Raphael's paintings. GAN samples exhibit unnatural blurs and artifacts, highlighting our method's ability to manage complex transformations without losing original image features. DDPM-PA, a current DDPM approach, underfits target domain images, creating noticeable color and style differences. Our TAN method retains source shapes and outlines while capturing the target style more effectively, maintaining detail in buildings and faces, and aligning closer to the target color style than DDPM-PA. Our DDPM and LDM-based approach generates more diverse and realistic samples with richer details compared to existing methods.

Quantitative Evaluation. In Table 1, we display the Intra-LPIPS results for DPMs-TAN under various 10-shot adaptation conditions. DDPM-TAN yields a considerable improvement in Intra-LPIPS across most tasks when compared with other GAN-based and DDPMs-based methods. Furthermore, LMD-TAN excels beyond state-of-the-art GAN-based approaches, demonstrating its potent capability to preserve diversity in few-shot image generation. The FID results are presented in Table 2, where TAN also demonstrates remarkable advancements compared to other GAN-based or DPMs-based methods, especially in FFHQ → 10-shot Sunglasses with 20.06 FID. We provide more results for other adaptation scenarios in the Appendix.

Table 2: FID (↓) results of each method on 10-shot FFHQ → Babies and Sunglasses. The best results are marked as **bold**.

Methods	TGAN	ADA	EWC	CDC	DCL	PA	ADMT
Babies	104.79	102.58	87.41	74.39	52.56	48.92	**46.70**
Sunglasses	55.61	53.64	59.73	42.13	38.01	34.75	**20.06**

5 Conclusion

We present TAN, utilizing adversarial noise selection and similarity guidance to enhance DPM training efficiency. TAN speeds up training, ensures faster convergence, and generates images matching the target style with source resemblance. Tests on few-shot image generation show TAN outperforms current GAN and DDPM methods in image quality and diversity.

REFERENCES

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Neural Information Processing Systems*, 2020.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Neural Information Processing Systems*, 2021.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Neural Information Processing Systems*, 2020.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, 2019.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Neural Information Processing Systems*, 2020a.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020b.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Yijun Li, Richard Zhang, Jingwan Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. *arXiv preprint arXiv:2012.02780*, 2020.

Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 2021.

Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *IEEE/CVF International Conference on Computer Vision*, 2019.

Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2015.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Neural Information Processing Systems*, 2016.

Yaxing Wang, Chenshen Wu, Luis Herranz, Joost Van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *European Conference on Computer Vision*, 2018.

Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

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 preprint arXiv:1506.03365*, 2015.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Yunqing Zhao, Henghui Ding, Houjing Huang, and Ngai-Man Cheung. A closer look at few-shot image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Jingyuan Zhu, Huimin Ma, Jiansheng Chen, and Jian Yuan. Few-shot image generation with diffusion models. *arXiv preprint arXiv:2211.03264*, 2022.

A PROOFS

A.1 SOURCE AND TARGET MODEL DISTANCE

This subsection introduces the detailed derivation of source and target model distance, Equation equation 5 as follows,

$$
\begin{aligned}
& \mathtt{D}_{\mathrm{KL}}\left(p_{\theta_{\mathcal{S}},\phi}(x_{t-1}^{\mathcal{S}}|x_t), p_{\theta_{\mathcal{T}},\phi}(x_{t-1}^{\mathcal{T}}|x_t)\right) \\
= \; & \mathtt{D}_{\mathrm{KL}}\left(p_{\theta_{(\mathcal{S},\mathcal{T})},\phi}(x_{t-1}|x_t, y=\mathcal{S}), p_{\theta_{(\mathcal{S},\mathcal{T})},\phi}(x_{t-1}|x_t, y=\mathcal{T})\right) \\
\approx \; & \mathtt{D}_{\mathrm{KL}}(\mathcal{N}(x_{t-1}; \mu_{\theta_{(\mathcal{S},\mathcal{T})}} + \sigma_t^2\gamma\nabla_{x_t}\log p_\phi(y=\mathcal{S}|x_t), \sigma_t^2\mathbf{I}), \mathcal{N}(x_{t-1}; \mu_{\theta_{(\mathcal{S},\mathcal{T})}} + \sigma_t^2\gamma\nabla_{x_t}\log p_\phi(y=\mathcal{T}|x_t), \sigma_t^2\mathbf{I})) \\
= \; & \mathbb{E}_{t,x_0,\epsilon}\left[\frac{1}{2\sigma_t^2}\left\|\mu_{\theta_{(\mathcal{S},\mathcal{T})}} + \sigma_t^2\gamma\nabla_{x_t}\log p_\phi(y=\mathcal{S}|x_t) - \mu_{\theta_{(\mathcal{S},\mathcal{T})}} - \sigma_t^2\gamma\nabla_{x_t}\log p_\phi(y=\mathcal{T}|x_t)\right\|^2\right] \\
= \; & \mathbb{E}_{t,x_0,\epsilon}\left[C_1\left\|\nabla_{x_t}\log p_\phi(y=\mathcal{S}|x_t) - \nabla_{x_t}\log p_\phi(y=\mathcal{T}|x_t)\right\|^2\right], \qquad (9)
\end{aligned}
$$

where $C_1 = \gamma/2$ is a constant. Since C_1 is scale constant, we can ignore this scale constant for the transfer gap and Equation equation 9 is the same as Equation equation 5.

A.2 SIMILARITY-GUIDED LOSS

In this subsection, we introduce the full proof how we get similarity-guided loss, Equation equation 6. Inspired by (Ho et al., 2020), training is carried out by optimizing the typical variational limit on negative log-likelihood:

$$
\begin{aligned}
\mathbb{E}[-\log p_{\theta,\phi}(x_0|y=\mathcal{T})] \leq \; & \mathbb{E}_q\left[-\log\frac{p_{\theta,\phi}(x_{0:T}|y=\mathcal{T})}{q(x_{1:T}|x_0)}\right] \\
= \; & \mathbb{E}_q\left[-\log p(x_T) - \sum_{t\geq 1}\log\frac{p_{\theta,\phi}(x_{t-1}|x_t, y=\mathcal{T})}{q(x_t|x_{t-1})}\right] := L\,. \qquad (10)
\end{aligned}
$$

According to (Ho et al., 2020), $q(x_t|x_0)$ can be expressed as:

$$
q(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\right)\,. \qquad (11)
$$

Training efficiency can be obtained by optimizing random elements of L in Equation equation 10 using stochastic gradient descent. Further progress is made via variance reduction by rewriting L in Equation equation 10 with Equation equation 11 as Ho et al. (2020):

$$
\begin{aligned}
L = \; & \mathbb{E}_q[\underbrace{\mathtt{D}_{\mathrm{KL}}\left(q(x_T|x_0, p(x_T|y=\mathcal{T})\right)}_{L_T} + \sum_{t>1}\underbrace{\mathtt{D}_{\mathrm{KL}}\left(q(x_{t-1}|x_t, x_0), p_{\theta,\phi}(x_{t-1}|x_t, y=\mathcal{T})\right)}_{L_{t-1}} \\
& - \underbrace{\log p_{\theta,\phi}(x_0|x_1, y=\mathcal{T})}_{L_0}]\,. \qquad (12)
\end{aligned}
$$

As Dhariwal & Nichol (2021), the conditional reverse noise process $p_{\theta,\phi}(x_{t-1}|x_t, y)$ is:

$$
p_{\theta,\phi}(x_{t-1}|x_t, y) \approx \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t) + \sigma_t^2\gamma\nabla_{x_t}\log p_\phi(y|x_t), \sigma_t^2\mathbf{I}\right)\,. \qquad (13)
$$

The L_{t-1} with Equation equation 13 can be rewrited as:

$$
\begin{aligned}
L_{t-1} := \; & \mathtt{D}_{\mathrm{KL}}\left(q(x_{t-1}|x_t, x_0), p_{\theta,\phi}(x_{t-1}|x_t, y=\mathcal{T})\right) \\
= \; & \mathbb{E}_q\left[\frac{1}{2\sigma_t^2}\left\|\tilde{\mu}_t(x_t, x_0) - \mu_t(x_t, x_0) - \sigma_t^2\gamma\nabla_{x_t}\log p_\phi(y|x_t)\right\|^2\right] \\
= \; & \mathbb{E}_{t,x_0,\epsilon}\left[C_2\left\|\epsilon_t - \epsilon_\theta(x_t, t) - \hat{\sigma}_t^2\gamma\nabla_{x_t}\log p_\phi(y=\mathcal{T}|x_t)\right\|^2\right]\,, \qquad (14)
\end{aligned}
$$

where $C_2 = \frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}$ is a constant, and $\hat{\sigma}_t = (1-\bar{\alpha}_{t-1})\sqrt{\frac{\alpha_t}{1-\bar{\alpha}_t}}$. We define the L_{t-1} as similarity-guided DPMs train loss and we will ignore the C_2 for better results during training as (Ho et al., 2020).

A.3 OPTIMIZATION

For time and GPU memory saving, we implement an additional adaptor module, ψ^l, (Noguchi & Harada, 2019) to learn the shift gap as Equation 5 based on x_t in practice. During the training, we keep the parameters of θ^l constant and update the additional adaptor layer parameters ψ^l. The overall loss function can be expressed as follows,

$$L(\psi) \equiv \mathbb{E}_{t,x_0} \left\| \epsilon^\star - \epsilon_{\theta,\psi}(x_t^\star, t) - \hat{\sigma}_t^2 \gamma \nabla_{x_t^\star} \log p_\phi(y = \mathcal{T} | x_t^\star) \right\|^2 \tag{15}$$

$$\text{s.t. } \epsilon^\star = \arg\max_{\epsilon} \left\| \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t) \right\|^2 \ , \ \epsilon^\star_{\text{mean}} = \mathbf{0} \text{ and } \epsilon^\star_{\text{std}} = \mathbf{I} \, , \tag{16}$$

where $\epsilon^\star$ is the "worse-case" noise, the $x_t^\star = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon^\star$ is the corresponding noised image at the timestep t and ψ is certain extra parameter beyond pre-trained model. We link the pre-trained U-Net model with the adaptor layer (Houlsby et al., 2019) as $x_t^l = \theta^l(x_t^{l-1}) + \psi^l(x_t^{l-1})$, where x_t^{l-1} and x_t^l represents the l-th layer of the input and output, and θ^l and ψ^l denote the l-th layer of the pre-trained U-Net and the additional adaptor layer, respectively.

Algorithm 1 Training DPMs with TAN

Require: binary classifier p_ϕ, pre-trained DPMs ϵ_θ, learning rate η

repeat
  $x_0 \sim q(x_0)$;
  $t \sim \text{Uniform}(\{1, \cdots, T\})$;
  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;
  for $j = 0, \cdots, J-1$ **do**
    Update ϵ^j via Eq. 8;
  end for
  Compute $L(\psi)$ with $\epsilon^\star = \epsilon^J$ via Eq. 15;
  Update the adaptor model parameter: $\psi = \psi - \eta \nabla_\psi L(\psi)$;
until converged.

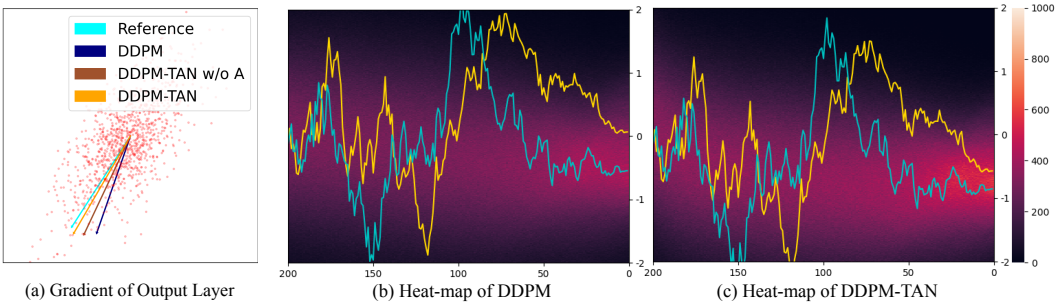
B ADDITIONAL EXPERIMENTS

In this section, we present additional experimental results, including the qualitative evaluation of FFHQ $\rightarrow$ Sunglasses and FFHQ $\rightarrow$ Babies in Appendix B.2, the quantitative evaluation of FFHQ $\rightarrow$ Sketches and FFHQ $\rightarrow$ Amedeo's paintings in Appendix B.3, effects of some key hyperparameters (i.e., similarity-guided training scale γ, adversarial noise selection scale ω, and the training iteration) in Appendix B.5, and an anonymous user study in Appendix B.6 to compare the proposed method with DDPM-PA.

B.1 ADDITIONAL VISUALIZATION ON TOY DATA

To conduct a quantitative analysis, we trained a diffusion model to generate 2-dimensional toy data with two Gaussian noise distributions. The means of the Gaussian noise distributions for the source and target are $(1, 1)$ and $(-1, -1)$, and their variances are denoted by $\mathbf{I}$. We train a simple neural network with source domain samples and then transfer this pre-trained model to target samples.

Figure 3 (a) illustrates the output layer gradient direction of four different settings in the first iteration, with the same noise and timestep t. The red line, computed with ten thousand different samples, is considered a reliable reference direction (close to 45 degrees southwest). For 10-shot samples, we repeat them a thousand times into one batch to provide a unified comparison with ten thousand different samples. The dark blue line and the sienna represent the gradient computed with the traditional DDPM as the baseline and similarity-guided training in a 10-shot sample, respectively. The orange line represents our method, DDPM-TAN, in a 10-shot sample. The gradient of our method is closer to the reliable reference direction, demonstrating that our approach can effectively correct the issue of the noisy gradient. The red points in the background symbolize "worse-case" noise, which is generated through adversarial noise selection. The graphic shows how the noise distribution transitions from a circle (representing a normal Gaussian distribution) to an ellipse. The principal axis of this ellipse is oriented along the gradient of the model parameters. This illustrates the noise



(a) Gradient of Output Layer (b) Heat-map of DDPM (c) Heat-map of DDPM-TAN

Figure 3: This Figure visualizes gradient changes and heat maps: Figure (a) shows gradient directions with various settings—the cyan line for the gradient of 10,000 samples in one step, dark blue for ten samples in one step as baseline method (trained with traditional DDPM), the sienna for our similarity-guided training, and the orange for our method DDPM-TAN, while red points at the background are "worse"-case noises by adversarial noise selection; Figure (b) and (c) depict heat-maps of the baseline and our method, with cyan and gold lines representing the generation sampling process value with the original DDPM and our method, respectively.

distribution shift under our adversarial noise selection approach, which effectively fine-tunes the model by actively targeting the “worst-case” noise that intensifies the adaptation task.

Figures 3 (b) and (c) present heatmaps of the baseline and our method in only one dimension, respectively. The cyan and gold lines denote the values of the generation sampling process using the original DDPM and our method. The heat-maps in the background illustrate the distribution of values for 20,000 samples generated by the original DDPM (baseline) and our method. The lighter the color in the background, the greater the number of samples present. There is a significantly brighter central highlight in (c) compared to (b), demonstrating that our method can learn the distribution more quickly than the baseline method. The gold and cyan lines in the two figures are approximately parallel, providing further evidence that our method can learn the gap more rapidly.

B.2 Additional Qualitative Evaluation

In Figure 4, we provide qualitative results for GAN-based and DDPM-based methods for the 10-shot FFHQ $\rightarrow$ Sunglasses and Babies task. The quantitative results are provided in Table 1. When compared to the GAN-based method (shown in the 2nd and 3rd rows), our approach (shown in the 5th and 6th rows) generates images of faces wearing sunglasses, displaying a wide variety of detailed hairstyles and facial features. Moreover, DPMs-TAN produces samples with more vivid and realistic reflections in the sunglasses. Notably, our method also manages to generate more realistic backgrounds.

B.3 Additional Quantitative Evaluation

As depicted in Table 3, our proposed DPMs-TAN method demonstrates superior performance over contemporary GAN-based and DPMs-based methods in terms of generation diversity for the given adaptation scenarios in FFHQ $\rightarrow$ Sketches and FFHQ $\rightarrow$ Amedeo’s paintings. Especially, we achieve 0.544 ± 0.025 for the FFHQ $\rightarrow$ Sketches, far more better than other methods.

Table 3: The Intra-LPIPS ($\uparrow$) results for both DDPM-based strategies and GAN-based baselines are presented for 10-shot image generation tasks. The best results are marked as **bold**.

Methods	FFHQ $\rightarrow$ Sketches	FFHQ $\rightarrow$ Amedeo’s paintings
CDC	0.454 ± 0.017	0.620 ± 0.029
DCL	0.461 ± 0.021	0.616 ± 0.043
DDPM-PA	0.495 ± 0.024	0.626 ± 0.022
DDPM-ANT (Ours)	0.544 ± 0.025	0.620 ± 0.021

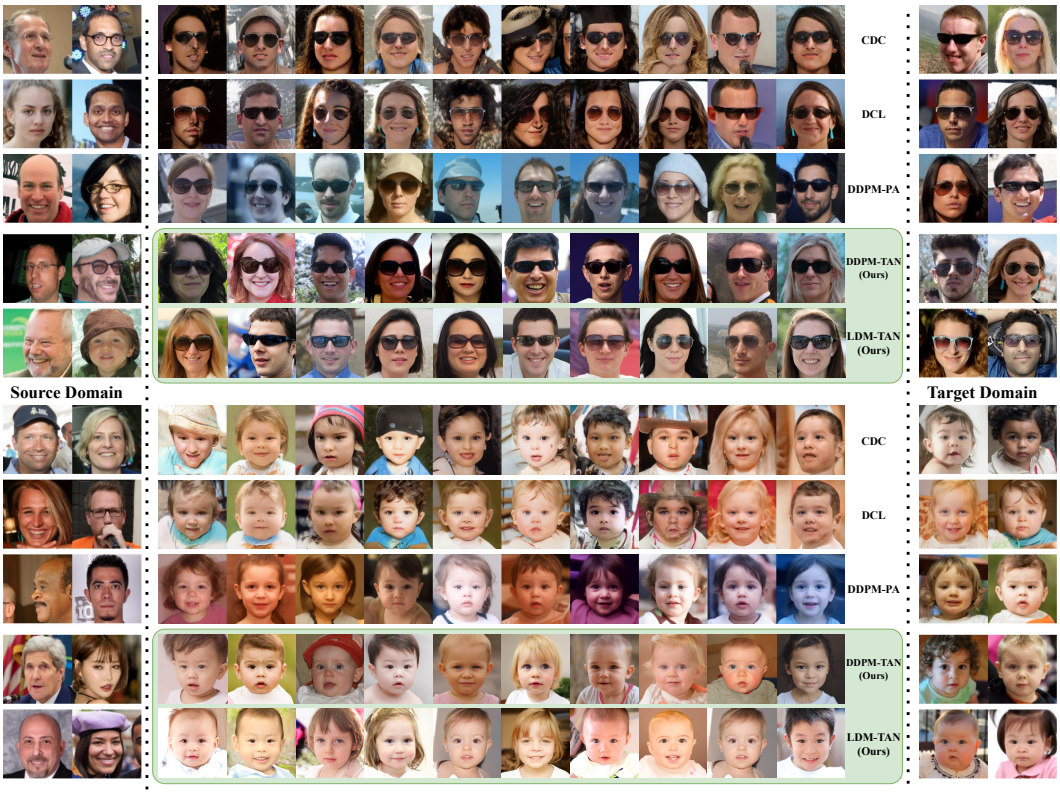


Figure 4: The 10-shot image generation samples on FFHQ → Sunglasses and FFHQ → Babies.

Figure 5: This figure shows our ablation study with all models trained for 300 iterations on a 10-shot sunglasses dataset measured with FID (↓)

B.4 Ablation Analysis

Figure 5 shows an ablation study using identical noise for image synthesis. Fine-tuning only the adaptor layer (2nd row) nearly matches the FID scores of full model fine-tuning (38.65 vs. 41.88). DPMs-TAN, with and without adversarial noise selection, uses an extra adaptor layer for efficiency, focusing on the last three rows for analysis.

The study illustrates successful model transfer to sunglasses across all methods, with TAN producing more detailed images. The similarity-guided method (3rd row) generates images with people wearing sunglasses, outperforming the traditional approach (2nd row). The effectiveness of TAN's adaptive noise selection is evident in the progressive transformation of faces to wearing sunglasses, demonstrating the method's ability to enhance the transfer process. FID scores decrease from 41.88 in direct adaptation to 26.41 with similarity-guided training, and further to 20.66 with DPMs-TAN, showcasing significant improvements in image quality.

B.5 Additional Ablation Studies

In this subsection, we study the effects of some key hyperparameters, including γ for the similarity-guided training, ω for the adversarial noise selection, and the count of training iterations. All experiments are conducted using a pre-trained LDM, and for evaluation purposes, we generate 1000 and 10000 images to compute the Intra-LPIPS and FID metrics, respectively.

Effects of Similarity-guided Training Scale γ. Table 4 shows the changes in FID (↓) and Intra-LPIPS (↑) scores for FFHQ → Sunglasses as γ (in Equation equation 7) increases. Initially, the FID score decrease, as the generated images gradually become closer to the target domain. At $\gamma = 5$, the FID reaches its lowest value of 18.13. Beyond this point, the FID score increases as the generated images become too similar to the target images or become random noise as failed case, leading to lower diversity and fidelity. The Intra-LPIPS score consistently decreases with gamma increasing, which further supports the idea that larger γ values lead to overfitting with the target image. Therefore, we select $\gamma = 5$ as a trade-off.

Table 4: This shows the change in FID (↓) and Intra-LPIPS (↑)kan results for FFHQ → Sunglasses as the γ value increases.

γ	FID (↓)	Intra-LPIPS (↑)
1	20.75	0.641 ± 0.014
3	18.86	0.627 ± 0.013
5	**18.13**	**0.613** ± 0.011
7	24.12	0.603 ± 0.017
9	29.48	0.592 ± 0.017

Effects of Adversarial Noise Selection Scale ω. As shown in Table 5, the FID (↓) and Intra-LPIPS (↑) scores for FFHQ → Sunglasses vary with an increase in the omega (ω) value (from Equation equation 8). Initially, the FID score decreases as the generated images gradually grow closer to the target image. When $\omega = 0.02$, the FID reaches its lowest value of 18.13. Beyond this point, the FID score increases because the synthesized images become too similar to the target image, which lowers diversity. The Intra-LPIPS score consistently decreases as ω increases, further supporting that larger ω values lead to overfitting with the target image. We also note that the results are relatively stable when ω is between 0.1 and 0.3. As such, we choose $\omega = 0.02$ as a balance between fidelity and diversity.

Table 5: This shows the change in FID (lower is better) and Intra-LPIPS (higher is better) results for FFHQ → Sunglasses as the ω value increases.

ω	FID (↓)	Intra-LPIPS (↑)
0.01	18.42	0.616 ± 0.020
0.02	**18.13**	**0.613** ± 0.011
0.03	18.42	0.613 ± 0.016
0.04	19.11	0.614 ± 0.013
0.05	19.48	0.623 ± 0.015

Effects of Training Iteration. As illustrated in Table 6, the FID (↓) and Intra-LPIPS (↑) for FFHQ → Sunglasses vary as training iterations increase. Initially, the FID value drops significantly as the generated image gradually resembles the target image, reaching its lowest at 18.13 with 300 training iterations. After this point, the FID score stabilizes after around 400 iterations as the synthesized images closely mirror the target image. The Intra-LPIPS score steadily decreases with an increase in iterations up to 400, further suggesting that a higher number of iterations can lead to overfitting to the target image. Therefore, we select 300 as an optimal number of training iterations, offering a balance between image quality and diversity.

Table 6: This shows the change in FID (lower is better) and Intra-LPIPS (higher is better) results for FFHQ → Sunglasses as the number of training iterations increases.

Iteration	FID (↓)	Intra-LPIPS (↑)
0	111.32	0.650 ± 0.071
50	93.82	0.666 ± 0.020
100	58.27	0.666 ± 0.015
150	31.08	0.654 ± 0.017
200	19.51	0.635 ± 0.014
250	18.34	0.624 ± 0.011
300	**18.13**	**0.613** ± 0.011
350	21.17	0.604 ± 0.016
400	21.17	0.608 ± 0.019

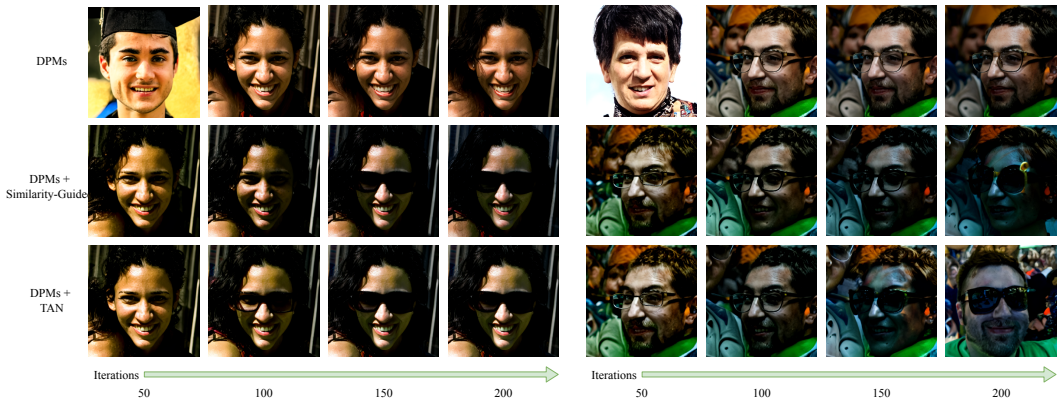


Figure 6: This figure shows our ablation study with all models trained for in different iterations on a 10-shot sunglasses dataset: the first line - baseline (direct fine-tuning model), second line - DPMs-TAN w/o A (only using similarity-guided training), and third line - DPMs-TAN (our method).

Quantitative Evaluation of Different Iteration. As shown in Figure 6, the first row demonstrate that the orangial train the DPMs with limited iterations is hard to get a successfully transfer. The second raw shows that training with our similarity-guide method can boost the convergence to the taget domain. The third rows shows that training further with adversrial noise can even more faster converge. As shown the 150 iteration of right pictures, compared with the training only with similarity-guide (2nd row) TAN can get the face with sunglasses image.

GPU Memory. Table 7 illustrates the GPU memory usage for each module in batch size 1, comparing scenarios with and without the use of an adaptor. It reveals that our module results in only a slight increase in GPU memory consumption.

B.6 Anonymous User Study

We carried out an additional anonymous user study to assess the qualitative performance of our method in comparison to DDPM-PA. In this study, participants were shown three sets of images

Method	DPMs	DPMs+SG	DPMs+AN	DPMs+TAN
Without Adaptor (MB)	17086	17130	17100	17188
With Adaptor (MB)	6010	6030	6022	6080

Table 7: This table displays the GPU memory consumption for each module, comparing scenarios with and without the use of the adaptor.

from each dataset, featuring DDPM-PA, our method (DDPM+TAN), and images from the target domain. For each set, we displayed five images from each method or the target image, as illustrated in our main paper. To maintain anonymity and neutrality, we labeled the methods as A/B instead of using the actual method names (PA and TAN). We recruited volunteers through an anonymous online platform for this study. During the study, participants were tasked with choosing the set of images (labeled as A or B, corresponding to PA or TAN) that they believed demonstrated higher quality and a closer resemblance to the target image set.

Of the 60 participants, a significant 73.35% favored our method (DDPM+TAN), indicating that it produced images of superior quality and more effectively captured the intricate types of target domains, as shown in Table 3. While this experiment did not comprehensively account for factors such as the participants' gender, age, regional background and others, the results nonetheless suggest that our images possess better visual quality to a notable extent.