A EXPERIMENT

In this section, we present the hyper-parameters of shift adaptor layer, Intra-LPIPS evaluation results on two additional datasets for 10-shot transformations, FFHQ → Sketches and FFHQ → Amedeo's paintings. We also provide supplementary information detailing our selection process for the hyper-parameters.

Table 4: The Intra-LPIPS (↑) 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 → Sketches	FFHQ → Amedeo's paintings
TGAN	0.394±0.023	0.548±0.026
TGAN+ADA	0.427±0.022	0.560±0.019
EWC	0.430±0.018	0.594±0.028
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-TAN (Ours)	0.544±0.025	0.620±0.021

A.1 SHIFT ADAPTOR LAYER

For the l-th shift adaptor layer ψ, it can be expressed as: $\psi^l(x^{l-1}) = f(x^{l-1}W_{down})W_{up}$ (Houlsby et al., 2019). We project the input downward using W_{down}, transforming it from its original dimension $\mathbb{R}^{w\times h\times r}$ to a lower-dimensional space with a bottleneck dimension $\mathbb{R}^{\frac{w}{c}\times\frac{h}{c}\times d}$. Following this, we apply a non-linear activation function $f(\cdot)$ and execute an upward projection with W_{up}. For the DDPMs, we set the parameters $c = 4$ and $d = 8$, whereas we set $c = 2$ and $d = 8$ for the LDMs. To ensure the adapter layer outputs are initialized to zero, we set all the extra layer parameters to zero.

A.2 ADDITIONAL RESULTS

Quantitative Evaluation. As depicted in Table 4, 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 → Sketches and FFHQ → Amedeo's paintings. Especially, we achieve 0.544±0.025 for the FFHQ → Sketches, far more better than other methods.

Qualitative Evaluation. In Figure 6, we provide additional results for GAN-based and DDPM-based methods for the 10-shot FFHQ → Sunglasses and Babies task. 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.

A.3 HYPER-PARAMETERS

In this subsection, we delve into our process for selecting 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 images for the Intra-LPIPS evaluation and 10000 images for the FID.

Table 5: 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

Similarity-guided Training Scale γ. Table 5 shows the changes in FID (↓) and Intra-LPIPS (↑) scores for FFHQ → Sunglasses as the γ (in 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.

Adversarial Noise Selection Scale ω. As shown in Table 6, the FID (↓) and Intra-LPIPS (↑) scores for FFHQ → Sunglasses vary with an increase in the omega (ω) value (from 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 6: 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

Iteration. As illustrated in Table 7, 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 7: 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

A.4 QUANTITATIVE EVALUATION OF DIFFERENT ITERATION

As shown in Figure 5, 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

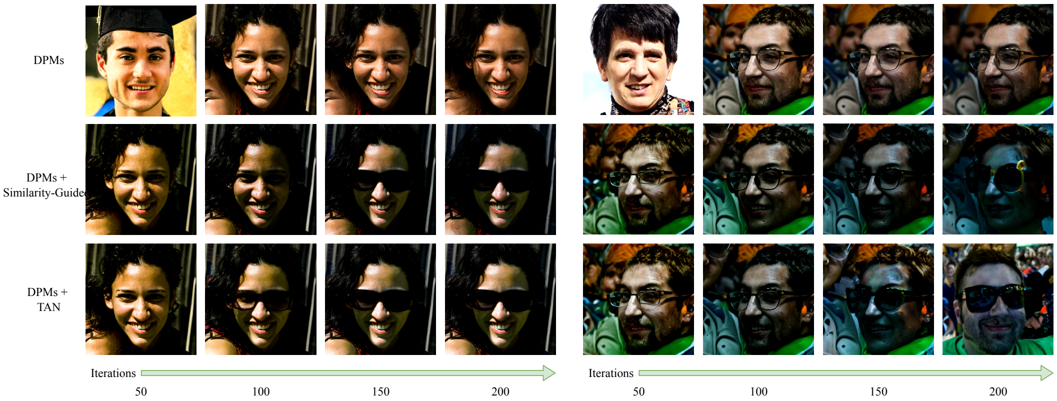

Figure 5: This figure shows our ablation study with all models trained for in different iterations on a 10-shot sunglasses dataset: the first line - baseline (direct fine-tuning model), second line - DPMs-TAN w/o A (only using similarity-guided training), and third line - DPMs-TAN (our method).

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

Table 8: This table displays the GPU memory consumption for each module, comparing scenarios with and without the use of the adaptor.

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.

A.5 GPU MEMORY

Table 8 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 PROOFS

B.1 SOURCE AND TARGET MODEL DISTANCE

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

$$\begin{aligned}
& \mathrm{D}_{\mathrm{KL}}\left(p_{\theta_{\mathcal{S}},\phi}(x_{t-1}^{\mathcal{S}}|x_t), p_{\theta_{\mathcal{T}},\phi}(x_{t-1}^{\mathcal{T}}|x_t)\right) \\
= \; & \mathrm{D}_{\mathrm{KL}}\left(p_{\theta_{(\mathcal{S},\mathcal{T})},\phi}(x_{t-1}|x_t, y=\mathcal{S}), p_{\theta_{(\mathcal{S},\mathcal{T})},\phi}(x_{t-1}|x_t, y=\mathcal{T})\right) \\
\approx \; & \mathrm{D}_{\mathrm{KL}}(\mathcal{N}(x_{t-1}; \mu_{\theta_{(\mathcal{S},\mathcal{T})}} + \sigma_t^2\gamma\nabla_{x_t}\log p_\phi(y=\mathcal{S}|x_t), \sigma_t^2\mathbf{I}), \mathcal{N}(x_{t-1}; \mu_{\theta_{(\mathcal{S},\mathcal{T})}} + \sigma_t^2\gamma\nabla_{x_t}\log p_\phi(y=\mathcal{T}|x_t), \sigma_t^2\mathbf{I})) \\
= \; & \mathbb{E}_{t,x_0,\epsilon}\left[\frac{1}{2\sigma_t^2}\left\|\mu_{\theta_{(\mathcal{S},\mathcal{T})}} + \sigma_t^2\gamma\nabla_{x_t}\log p_\phi(y=\mathcal{S}|x_t) - \mu_{\theta_{(\mathcal{S},\mathcal{T})}} - \sigma_t^2\gamma\nabla_{x_t}\log p_\phi(y=\mathcal{T}|x_t)\right\|^2\right] \\
= \; & \mathbb{E}_{t,x_0,\epsilon}\left[C_1\left\|\nabla_{x_t}\log p_\phi(y=\mathcal{S}|x_t) - \nabla_{x_t}\log p_\phi(y=\mathcal{T}|x_t)\right\|^2\right], \qquad (11)
\end{aligned}$$

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 11 is the same as Equation 5.

B.2 SIMILARITY-GUIDED LOSS

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

$$\begin{aligned}\mathbb{E}[-\log p_{\theta,\phi}(x_0|y=\mathcal{T})] \leq\ & \mathbb{E}_q\left[-\log \frac{p_{\theta,\phi}(x_{0:T}|y=\mathcal{T})}{q(x_{1:T}|x_0)}\right] \\ =\ & \mathbb{E}_q\left[-\log p(x_T) - \sum_{t\geq 1} \log \frac{p_{\theta,\phi}(x_{t-1}|x_t, y=\mathcal{T})}{q(x_t|x_{t-1})}\right] := L\,. \end{aligned} \tag{12}$$

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

$$q(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\right)\ . \tag{13}$$

Training efficiency can be obtained by optimizing random elements of L 12 using stochastic gradient descent. Further progress is made via variance reduction by rewriting L 12 with Equation 13 as Ho et al. (2020):

$$\begin{aligned} L =\ & \mathbb{E}_q[\underbrace{\mathrm{D}_{\mathrm{KL}}\left(q(x_T|x_0, p(x_T|y=\mathcal{T})\right)}_{L_T} + \sum_{t>1} \underbrace{\mathrm{D}_{\mathrm{KL}}\left(q(x_{t-1}|x_t, x_0), p_{\theta,\phi}(x_{t-1}|x_t, y=\mathcal{T})\right)}_{L_{t-1}} \\ & - \underbrace{\log p_{\theta,\phi}(x_0|x_1, y=\mathcal{T})}_{L_0}]\ . \end{aligned} \tag{14}$$

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

$$p_{\theta,\phi}(x_{t-1}|x_t, y) \approx \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t) + \sigma_t^2\gamma\nabla_{x_t}\log p_\phi(y|x_t), \sigma_t^2\mathbf{I}\right)\ . \tag{15}$$

The L_{t-1} with Equation 15 can be rewrited as:

$$\begin{aligned} L_{t-1} &:= \mathrm{D}_{\mathrm{KL}}\left(q(x_{t-1}|x_t, x_0), p_{\theta,\phi}(x_{t-1}|x_t, y=\mathcal{T})\right) \\ &= \mathbb{E}_q\left[\frac{1}{2\sigma_t^2}\left\|\tilde{\mu}_t(x_t, x_0) - \mu_t(x_t, x_0) - \sigma_t^2\gamma\nabla_{x_t}\log p_\phi(y|x_t)\right\|^2\right] \\ &= \mathbb{E}_{t,x_0,\epsilon}\left[C_2\left\|\epsilon_t - \epsilon_\theta(x_t, t) - \hat{\sigma}_t^2\gamma\nabla_{x_t}\log p_\phi(y=\mathcal{T}|x_t)\right\|^2\right]\ , \end{aligned} \tag{16}$$

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).

C LIMITATION

In this subsection, we acknowledge some limitations of our method. Given that our goal is to transfer the image from the source domain to the target domain, the images we synthesize will feature characteristics specific to the target domain, such as sunglasses as shown in Figure 4. This can potentially lead to inconsistency in the generated images and there is a risk of privacy leakage. For instance, the reflection in the sunglasses seen in the 3rd and 4th columns of the 4th row in Figure 4 is very similar to the one in the target image. This could potentially reveal sensitive information from the target domain, which is an issue that needs careful consideration in applying this method.

D SAMPLES

D.1 SUNGLASSES AND BABY

Figure 6 presents samples from GAN-based and DDPM-based methods for 10-shot FFHQ → Sunglasses (top) and FFHQ → Babies (bottom).

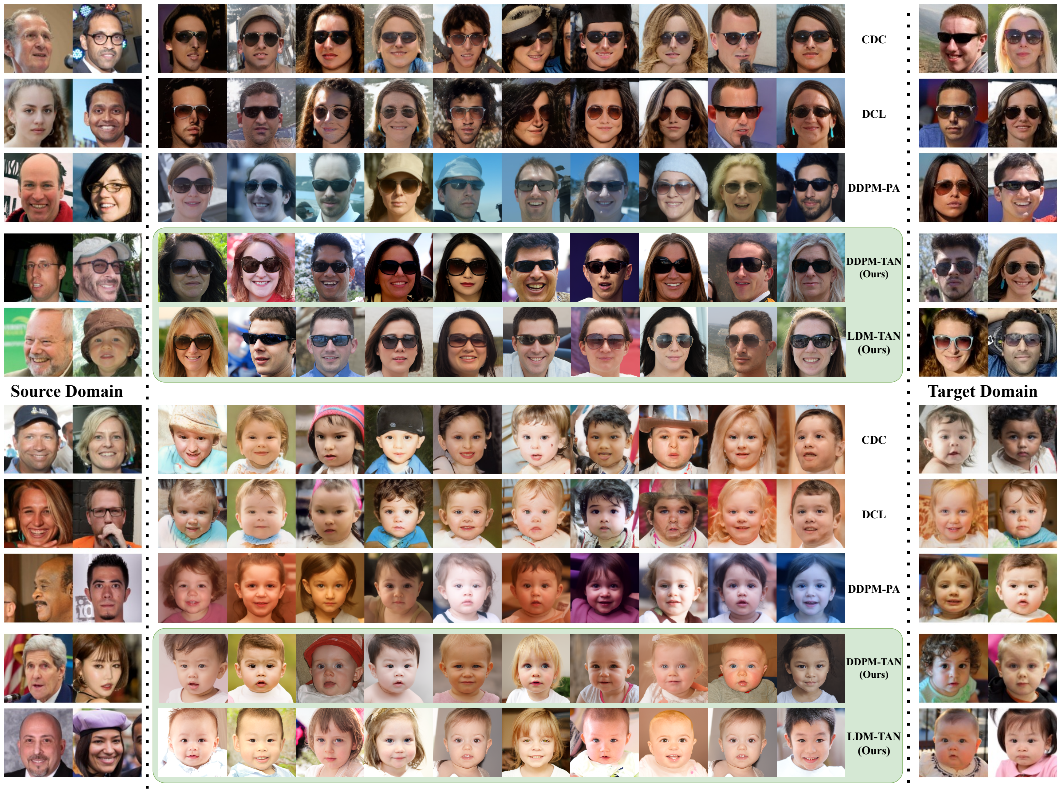

Figure 6: The 10-shot image generation samples on FFHQ → Sunglasses and FFHQ → Babies.

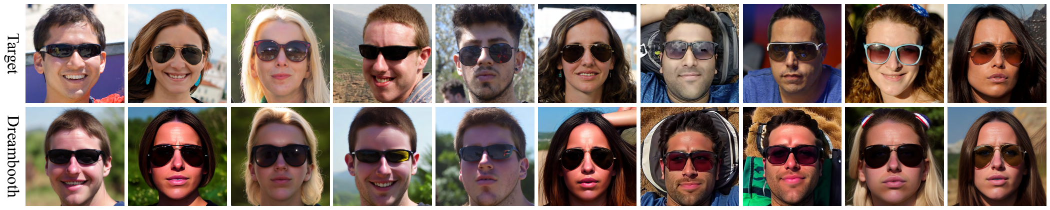

Figure 7: This is the generated image by dreambooth.

D.2 DREAMBOOTH

Figure 7 displays samples with the target image at the top and DreamBooth fine-tuned images at the bottom, as the generated images closely resemble the target images.