SUPPLEMENTARY MATERIAL

We provide a detailed supplementary to help readers further understand our work and make this paper more convincing. The supplementary materials are organized as follows:

- Appendix A: **DDPMs Trained from Scratch**
  A detailed illustration of DDPMs trained from scratch on limited data. Experiment setups, training details, qualitative and quantitative evaluation are provided.
- Appendix B: **More Details of Employed Losses**
  Introduction of the variational lower bound loss (Kingma et al., 2021) and prior preservation loss (Ruiz et al., 2023) proposed in prior works and employed by DomainStudio.
- Appendix C: **Evaluation Metrics**
  Detailed explanations of the metrics used in the quantitative evaluation of DomainStudio.
- Appendix D: **Additional Quantitative Evaluation**
  The quantitative evaluation of DomainStudio compared with baselines under a series of unconditional and text-to-image generation setups, as supplements to Sec. 4.2.
- Appendix E: **Additional Ablation Analysis**
  The ablation analysis of each component in DomainStudio, as supplements to Sec. 4.3.
- Appendix F: **Limitations and Societal Impact**
  The limitations and societal impact of DomainStudio.
- Appendix G: **Personalization of DomainStudio**
  The personalization of DomainStudio achieves domain-driven and subject-driven image generation at the same time using two sets of reference data. Methods and visualized samples are provided.
- Appendix H: **More Details of Implementation**
  The implementation of DomainStudio and baselines is introduced in detail.
- Appendix I: **Unconditional Source Models**
  The training details, visualized samples, and quantitative evaluation of the source models trained on FFHQ (Karras et al., 2020b) and LSUN Church (Yu et al., 2015).
- Appendix J: **Inspiration of DomainStudio**
  The inspiration of DomainStudio design is discussed.
- Appendix K: **DDPM Adaptation Process Analysis**
  Visualized samples across different training iterations to show the domain adaptation process qualitatively.
- Appendix L: **Additional Comparison with Related Works**
  Comparison between DomainStudio and other related works.
- Appendix M: **Additional Visualized Samples**
  More visualized samples are shown, including the few-shot datasets used in this paper and the visualized results of DomainStudio under unconditional and text-to-image generation setups.
- Appendix N: **Computational Cost**
  The computational cost of DomainStudio compared with DDPM-based baselines.

Reproducibility: See the code provided in the submitted compressed file.

A DDPMs TRAINED FROM SCRATCH

We make the first attempt to evaluate the performance of DDPMs trained from scratch as data become scarce. We first train DDPMs on small-scale datasets containing various numbers of images from scratch. We analyze generation diversity to study when do DDPMs overfit as training samples decrease. We sample 10, 100, and 1000 images from FFHQ-babies (Babies), FFHQ-sunglasses (Sunglasses) (Ojha et al., 2021), and LSUN Church (Yu et al., 2015) respectively as small-scale training datasets. The image resolution of all the datasets is set as 256×256. We follow the model setups in prior works (Nichol & Dhariwal, 2021; Dhariwal & Nichol, 2021) used for LSUN 256^2 (Yu et al., 2015) and use a learning rate of 1e-4 and a batch size of 48.

In our experiments, the smaller datasets are included in the larger datasets. For example, the 1000-shot Sunglasses datasets include all the images in 100-shot and 10-shot Sunglasses. Similarly, all the images in 10-shot Sunglasses are included in 100-shot Sunglasses as well. We train DDPMs for 40K iterations (about 20 hours on $\times 8$ NVIDIA RTX A6000 GPUs) on datasets containing 10 or 100 images. While for datasets containing 1000 images, DDPMs are trained for 60K iterations (about 30 hours on $\times 8$ NVIDIA RTX A6000 GPUs).

Qualitative Evaluation Compared with the generated images shown in Fig. 10, it can be seen that DDPMs trained from scratch need enough training samples to synthesize diverse results and avoid replicating the training samples. They overfit and tend to replicate training samples when datasets are limited to 10 or 100 images. Since some training samples are flipped in the training process as a step of data augmentation, we can also find some generated images symmetric to the training samples. For datasets containing 1000 images, DDPMs can generate diverse samples following similar distributions of training samples instead of replicating them. The overfitting problem is relatively alleviated. However, the generated samples are coarse and lack high-frequency details compared with training samples.

Quantitative Evaluation LPIPS (Zhang et al., 2018a) is proposed to evaluate the perceptual distances (Johnson et al., 2016) between images. We propose a Nearest-LPIPS metric based on LPIPS to evaluate the generation diversity of DDPMs trained on small-scale datasets. More specifically, we first generate 1000 images randomly and find the most similar training sample having the lowest LPIPS distance to each generated sample. Nearest-LPIPS is defined as the LPIPS distances between generated samples and the most similar training samples in correspondence averaged over all the generated samples. If a generative model reproduces the training samples exactly, the Nearest-LPIPS metric will have a score of zero. Larger Nearest-LPIPS values indicate lower replication rates and greater diversity relative to training samples.

We provide the Nearest-LPIPS results of DDPMs trained from scratch on small-scale datasets in the top part of Table 3. For datasets containing 10 or 100 images, we have lower Nearest-LPIPS values. While for datasets containing 1000 images, we get measurably improved Nearest-LPIPS values. To avoid the influence of generated images symmetric to training samples, we flip all the training samples as supplements to the original datasets and recalculate the Nearest-LPIPS metric. The results are listed in the bottom part of Table 3. With the addition of flipped training samples, we find apparently lower Nearest-LPIPS values for datasets containing 10 or 100 images. However, we get almost the same Nearest-LPIPS results for DDPMs trained on larger datasets containing 1000 images, indicating that these models can generate diverse samples different from the original or symmetric training samples.

Number of Samples	Babies	Sunglasses	Church
10	0.2875	0.3030	0.3136
100	0.3152	0.3310	0.3327
1000	**0.4658**	**0.4819**	**0.5707**
10 (+ flip)	0.1206	0.1217	0.0445
100 (+ flip)	0.1556	0.1297	0.1177
1000 (+ flip)	**0.4611**	**0.4726**	**0.5625**

Table 3: Nearest-LPIPS (↑) results of DDPMs trained from scratch on several small-scale datasets.

Figure 10: Samples produced by DDPMs trained from scratch on small-scale datasets, including Babies, Sunglasses, and LSUN Church containing 10, 100, and 1000 images.

To sum up, it becomes harder for DDPMs to learn the representations of datasets as training data become scarce. When trained on limited data from scratch, DDPMs fail to match target data distributions exactly and cannot produce high-quality and diverse samples.

B More Details of Employed Losses

Variational Lower Bound Loss ($\mathcal{L}_{vlb}$) In Ho et al. (2020), the variance $\Sigma_\theta(x_t, t)$ is fixed as a constant $\sigma_t^2\mathbf{I}$, where $\sigma_t^2 = \beta_t$ and is not learned. The network is only trained to learn the model mean $\mu_\theta(x_t, t)$ through predicting noises with $\epsilon_\theta(x_t, t)$. Following works (Kingma et al., 2021) propose to optimize the variational lower bound (VLB) and guide the learning of $\Sigma_\theta(x_t, t)$ with an additional optimization term L_{vlb} as follows:

$$L_{vlb} := L_0 + L_1 + ... + L_{T-1} + L_T, \tag{17}$$
$$L_0 := -log\ p_\theta(x_0|x_1), \tag{18}$$
$$L_{t-1} := D_{KL}(q(x_{t-1}|x_t, x_0)\ ||\ p_\theta(x_{t-1}|x_t)), \tag{19}$$
$$L_T := D_{KL}(q(x_T|x_0)\ ||\ p(x_T)). \tag{20}$$

We set the weight of $\mathcal{L}_{vlb}$ as 0.001 to avoid it from overwhelming the other losses.

Prior Preservation Loss ($\mathcal{L}_{pr}$) DreamBooth (Ruiz et al., 2023) generates source samples x^{pr} with randomly sampled Gaussian noises and the source text condition c_{sou} using the pre-trained text-to-image model. Then the pre-trained encoder $\mathcal{E}$ is employed to compress x^{pr} to latent codes z^{pr}. DreamBooth proposes a class-specific prior preservation loss as follows to relieve overfitting for subject-driven generation by preserving the information of source samples:

$$\mathcal{L}_{pr} = \mathbb{E}_{t, z_t^{pr}, c_{sou}, \epsilon^{pr}} ||\epsilon_{sou}(z_t^{pr}, c_{sou}) - \epsilon_{ada}(z_t^{pr}, c_{sou})||^2. \tag{21}$$

DomainStudio employs the prior preservation loss $\mathcal{L}_{pr}$ to maintain the source samples produced by adapted models during the few-shot fine-tuning process. We follow DreamBooth to set its weight as 1 for fair comparison.

Datasets	FFHQ → Babies	FFHQ → Sunglasses	FFHQ → Raphael's paintings
TGAN (Wang et al., 2018)	0.510 ± 0.026	0.550 ± 0.021	0.533 ± 0.023
TGAN+ADA (Karras et al., 2020a)	0.546 ± 0.033	0.571 ± 0.034	0.546 ± 0.037
FreezeD (Mo et al., 2020)	0.535 ± 0.021	0.558 ± 0.024	0.537 ± 0.026
MineGAN (Wang et al., 2020)	0.514 ± 0.034	0.570 ± 0.020	0.559 ± 0.031
EWC (Li et al., 2020)	0.560 ± 0.019	0.550 ± 0.014	0.541 ± 0.023
CDC (Ojha et al., 2021)	0.583 ± 0.014	0.581 ± 0.011	0.564 ± 0.010
DCL (Zhao et al., 2022b)	0.579 ± 0.018	0.574 ± 0.007	0.558 ± 0.033
AdAM (Zhao et al., 2022a)	0.573 ± 0.016	0.559 ± 0.017	0.551 ± 0.033
RICK (Zhao et al., 2023)	0.589 ± 0.010	0.591 ± 0.030	0.582 ± 0.028
Fine-tuned DDPMs	0.513 ± 0.026	0.527 ± 0.024	0.466 ± 0.018
DomainStudio (ours)	$\mathbf{0.599 \pm 0.024}$	$\mathbf{0.604 \pm 0.014}$	$\mathbf{0.594 \pm 0.022}$
Datasets	FFHQ → Sketches	LSUN Church → Haunted houses	LSUN Church → Landscape drawings
TGAN (Wang et al., 2018)	0.394 ± 0.023	0.585 ± 0.007	0.601 ± 0.030
TGAN+ADA (Karras et al., 2020a)	0.427 ± 0.022	0.615 ± 0.018	0.643 ± 0.060
FreezeD (Mo et al., 2020)	0.406 ± 0.017	0.558 ± 0.019	0.597 ± 0.032
MineGAN (Wang et al., 2020)	0.407 ± 0.020	0.586 ± 0.041	0.614 ± 0.027
EWC (Li et al., 2020)	0.430 ± 0.018	0.579 ± 0.035	0.596 ± 0.052
CDC (Ojha et al., 2021)	0.454 ± 0.017	0.620 ± 0.029	0.674 ± 0.024
DCL (Zhao et al., 2022b)	0.461 ± 0.021	0.616 ± 0.043	0.626 ± 0.021
AdAM (Zhao et al., 2022a)	0.424 ± 0.018	0.584 ± 0.031	0.694 ± 0.026
RICK (Zhao et al., 2023)	0.443 ± 0.025	0.622 ± 0.021	0.694 ± 0.031
Fine-tuned DDPMs	0.473 ± 0.022	0.590 ± 0.045	0.666 ± 0.044
DomainStudio (ours)	$\mathbf{0.495 \pm 0.024}$	$\mathbf{0.628 \pm 0.029}$	$\mathbf{0.715 \pm 0.034}$

Table 4: Intra-LPIPS (↑) results of DDPM-based approaches and GAN-based baselines on 10-shot unconditional image generation tasks adapted from the source datasets FFHQ and LSUN Church. Standard deviations are computed across 10 clusters (the same number as training samples). DomainStudio outperforms modern GAN-based approaches and achieves state-of-the-art performance in generation diversity.

C Evaluation Metrics

We follow CDC (Ojha et al., 2021) to use Intra-LPIPS for generation diversity evaluation. To be more specific, we generate 1000 images and assign them to one of the training samples with the lowest LPIPS (Zhang et al., 2018a) distance. Intra-LPIPS is defined as the average pairwise LPIPS distances within members of the same cluster averaged over all the clusters. If a model exactly replicates training samples, its Intra-LPIPS will have a score of zero. Larger Intra-LPIPS values correspond to greater generation diversity.

FID (Heusel et al., 2017) is widely used to evaluate the generation quality of generative models by computing the distribution distances between generated samples and datasets. However, FID would become unstable and unreliable when it comes to datasets containing a few samples (e.g., 10-shot datasets used in this paper). Therefore, we provide FID evaluation using relatively richer datasets including Sunglasses and Babies, which contain 2500 and 2700 images for unconditional image generation.

Given a text prompt like "a [V] volcano" in representation of adapted samples, we use the text prompt "a volcano" to compute CLIP-Text to evaluate the subject preservation in domain-driven generation. Apart from CLIP-text, we add a CLIP-Image metric to measure the domain consistency of DomainStudio on T2I generation. CLIP-Image is defined as the average pairwise cosine similarity between the CLIP embeddings of training and generated samples. CLIP-Image may be unbiased when the model is overfitting. For example, if a model exactly replicates training samples, its CLIP-Image will have the highest score of 1. We provide CLIP-Image results as reference.

The noise inputs are fixed for DDPM-based and GAN-based approaches respectively to synthesize samples for fair comparison of generation quality and diversity.

Method	TGAN	TGAN+ADA	FreezeD	MineGAN	EWC	CDC	DCL	AdAM	RICK	Ours
Babies	104.79	102.58	110.92	98.23	87.41	74.39	52.56	48.43	**39.39**	48.92
Sunglasses	55.61	53.64	51.29	68.91	59.73	42.13	38.01	28.03	**25.22**	34.75

Table 5: FID (↓) results of DomainStudio compared with GAN-based baselines under unconditional adaptation from FFHQ to 10-shot Babies and Sunglasses.

Datasets	Van Gogh houses	Wrecked trains	Ink painting volcanoes
Metrics	CLIP-Text		
LoRA (Hu et al., 2021)	0.269 ± 0.012	0.199 ± 0.018	0.292 ± 0.018
Textual Inversion (Gal et al., 2022)	0.259 ± 0.011	0.243 ± 0.024	0.244 ± 0.019
DreamBooth (Ruiz et al., 2023)	0.262 ± 0.035	0.267 ± 0.013	0.275 ± 0.020
DomainStudio (ours)	$\mathbf{0.276 \pm 0.028}$	$\mathbf{0.271 \pm 0.041}$	$\mathbf{0.301 \pm 0.024}$
Metrics	CLIP-Image		
LoRA (Hu et al., 2021)	0.773 ± 0.032	0.689 ± 0.069	0.668 ± 0.062
Textual Inversion (Gal et al., 2022)	0.763 ± 0.022	$\mathbf{0.737 \pm 0.035}$	0.658 ± 0.023
DreamBooth (Ruiz et al., 2023)	0.569 ± 0.039	0.557 ± 0.011	0.600 ± 0.086
DomainStudio (ours)	$\mathbf{0.789 \pm 0.024}$	0.600 ± 0.068	$\mathbf{0.676 \pm 0.091}$

Table 6: CLIP-Text (↑) and CLIP-Image results of DomainStudio compared with LoRA, Textual Inversion, and DreamBooth on text-to-image generation tasks. DomainStudio outperforms baselines on text alignment.

D Additional Quantitative Evaluation

We add earlier baselines in this section for more complete quantitative evaluation, including unconditional GAN-based methods TGAN (Wang et al., 2018), TGAN+ADA (Karras et al., 2020a), FreezeD (Mo et al., 2020), MineGAN (Wang et al., 2020), EWC (Li et al., 2020), and T2I method LoRA (Hu et al., 2021) based on Stable Diffusion (Rombach et al., 2022).

Unconditional Image Generation We provide the Intra-LPIPS results of DomainStudio under a series of 10-shot adaptation setups in Table 4. DomainStudio realizes a superior improvement of Intra-LPIPS compared with directly fine-tuned DDPMs. Besides, DomainStudio outperforms state-of-the-art GAN-based approaches under all the employed adaptation setups, indicating its strong capability of maintaining generation diversity.

As shown by the FID results in Table 5, DomainStudio performs better on learning target distributions from limited data than most prior GAN-based approaches. Despite its outstanding FID results, RICK still fails to avoid generating unnatural deformation and blurs like prior GAN-based methods. DomainStudio achieves better visual effects, as shown in Fig. 24 and 26. We only provide the FID results on Babies and Sunglasses since we have no access to enough samples to support stable and reliable FID evaluation for other datasets.

T2I Generation We report the CLIP-based metrics of DomainStudio compared with LoRA (Hu et al., 2021), Textual Inversion (Gal et al., 2022), and DreamBooth (Ruiz et al., 2023) in Table 6. DomainStudio achieves better results of CLIP-Text than baselines, indicating its ability to synthesize images consistent with text prompts while adapting to target domains. As for CLIP-Image results, DomainStudio also outperforms baselines on several benchmarks. **Textual Inversion achieves the best image alignment on 10-shot Wrecked trains since it overfits to the few-shot car samples instead of synthesizing train samples consistent with the text prompt.**

In Table 7, we provide Intra-LPIPS results of DomainStudio and baselines to evaluate the generation diversity. DomainStudio achieves state-of-the-art performance when generating adapted samples sharing the same category of subjects with training samples like Van Gogh houses and Wrecked cars. **Although Textual Inversion and LoRA achieve better generation diversity in terms of Intra-LPIPS on adapted samples like Watercolor pandas and temples, it fails to produce samples sharing styles with training samples and containing subjects consistent with text prompts, as shown in Fig. 7, 8, 29, and 32.**

Datasets	Van Gogh houses	Watercolor pandas	Watercolor temples
LoRA (Hu et al., 2021)	0.578 ± 0.029	0.606 ± 0.018	0.602 ± 0.019
Textual Inversion (Gal et al., 2022)	0.480 ± 0.235	$\mathbf{0.744 \pm 0.031}$	$\mathbf{0.763 \pm 0.033}$
DreamBooth (Ruiz et al., 2023)	0.558 ± 0.009	0.450 ± 0.099	0.553 ± 0.082
DomainStudio (ours)	$\mathbf{0.588 \pm 0.012}$	0.519 ± 0.014	0.544 ± 0.010
Datasets	Wrecked cars	Wrecked houses	Ink painting volcanoes
LoRA (Hu et al., 2021)	0.593 ± 0.011	0.606 ± 0.014	0.580 ± 0.053
Textual Inversion (Gal et al., 2022)	0.612 ± 0.024	0.624 ± 0.015	$\mathbf{0.648 \pm 0.038}$
DreamBooth (Ruiz et al., 2023)	0.534 ± 0.027	0.601 ± 0.034	0.535 ± 0.049
DomainStudio (ours)	$\mathbf{0.636 \pm 0.012}$	$\mathbf{0.628 \pm 0.017}$	0.633 ± 0.029

Table 7: Intra-LPIPS (↑) results of DomainStudio compared with LoRA, Textual Inversion, and DreamBooth on T2I generation tasks.

E Additional Ablation Analysis

We provide detailed ablation analysis of the weight coefficients of $\mathcal{L}_{img}$, $\mathcal{L}_{hf}$, and $\mathcal{L}_{hfmse}$ using 10-shot FFHQ → Babies (unconditional) as an example. Intra-LPIPS and FID are employed for quantitative evaluation.

We first ablate λ_2, the weight coefficient of $\mathcal{L}_{img}$. We adapt the source model to 10-shot Babies without $\mathcal{L}_{hf}$ and $\mathcal{L}_{hfmse}$. The quantitative results are listed in Table 8. Corresponding generated samples are shown in Fig. 11. When λ_2 is set as 0.0, the directly fine-tuned model produces coarse results lacking high-frequency details and diversity. With an appropriate choice of λ_2, the adapted model achieves greater generation diversity and better learning of target distributions under the guidance of $\mathcal{L}_{img}$. Too large values of λ_2 make $\mathcal{L}_{img}$ overwhelm $\mathcal{L}_{simple}$ and prevent the adapted model from learning target distributions, leading to degraded generation quality and diversity. The adapted model with λ_2 value of 2.5 gets unnatural generated samples even if it achieves the best FID result. We recommend λ_2 ranging from 0.1 to 1.0 for the unconditional adaptation setups used in our paper based on a comprehensive consideration of the qualitative and quantitative evaluation.

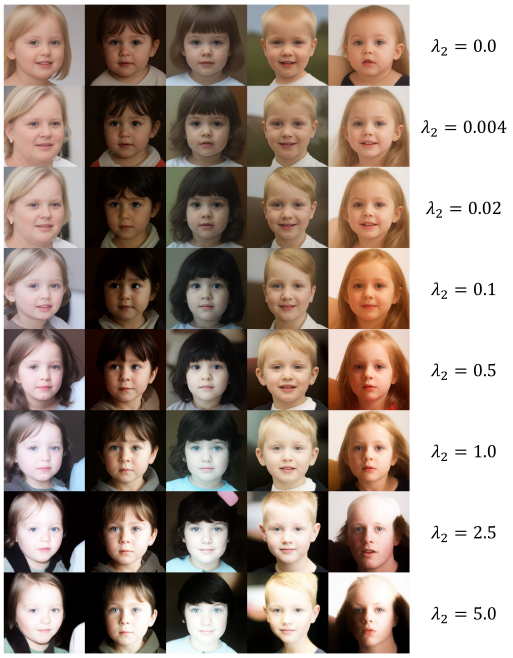

Figure 11: Visualized ablations of λ_2, the weight coefficient of $\mathcal{L}_{img}$ on 10-shot FFHQ → Babies.

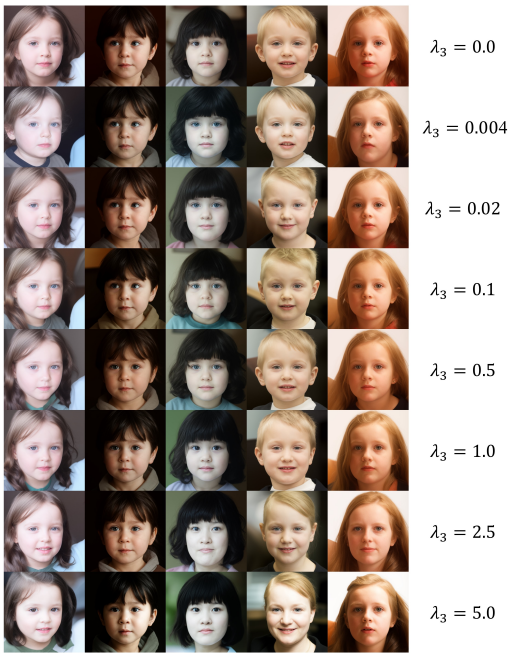

Figure 12: Visualized ablations of λ_3, the weight coefficient of $\mathcal{L}_{hf}$ on 10-shot FFHQ → Babies.

λ_2	Intra-LPIPS (↑)	FID (↓)
0.0	0.520 ± 0.026	114.95
0.004	0.531 ± 0.031	92.87
0.02	0.544 ± 0.026	85.11
0.1	0.558 ± 0.033	75.17
0.5	$\mathbf{0.572 \pm 0.027}$	71.77
1.0	0.560 ± 0.034	74.68
2.5	0.543 ± 0.038	**64.08**
5.0	0.537 ± 0.028	69.18

Table 8: Intra-LPIPS (↑) and FID (↓) results of adapted models trained on 10-shot FFHQ → Babies with different λ_2, the weight coefficient of $\mathcal{L}_{img}$.

λ_3	Intra-LPIPS (↑)	FID (↓)
0.0	0.572 ± 0.027	71.77
0.004	0.576 ± 0.034	**66.48**
0.02	0.581 ± 0.045	72.67
0.1	0.589 ± 0.047	70.75
0.5	$\mathbf{0.592 \pm 0.031}$	70.40
1.0	0.583 ± 0.032	68.06
2.5	0.577 ± 0.032	71.69
5.0	0.591 ± 0.031	71.20

Table 9: Intra-LPIPS (↑) and FID (↓) results of adapted models trained on 10-shot FFHQ → Babies with different λ_3, the weight coefficient of $\mathcal{L}_{hf}$.

Next, we ablate λ_3, the weight coefficient of $\mathcal{L}_{hf}$ with λ_2 set as 0.5. The quantitative results are listed in Table 9. Corresponding generated samples are shown in Fig. 12. $\mathcal{L}_{hf}$ guides adapted models to keep diverse high-frequency details learned from source samples for more realistic results. $\mathcal{L}_{hf}$ helps the adapted model enhance details like clothes and hairstyles and achieves better FID and Intra-LPIPS, indicating improved quality and diversity. Too large values of λ_3 make the adapted model pay too much attention to high-frequency components and fail to produce realistic results following the target distributions. We recommend λ_3 ranging from 0.1 to 1.0 for the unconditional adaptation setups used in our paper.

Finally, we ablate λ_4, the weight coefficient of $\mathcal{L}_{hfmse}$, with λ_2 and λ_3 set as 0.5. The quantitative results are listed in Table 10. Corresponding generated samples are shown in Fig. 13. $\mathcal{L}_{hfmse}$ guides the adapted model to learn more high-frequency details from limited training data. Appropriate choice of λ_4 helps the adapted model generate diverse results containing rich details. Besides, the full DomainStudio approach achieves state-of-the-art results of FID and Intra-LPIPS on 10-shot FFHQ → Babies (see Table 4 and 5). Similar to λ_2 and λ_3, too large values of λ_4 lead to unreasonable results deviating from the target distributions. We recommend λ_4 ranging from 0.01 to 0.08 for the unconditional adaptation setups in this paper. Results in Fig. 11, 12, and 13 are synthesized from fixed noise inputs.

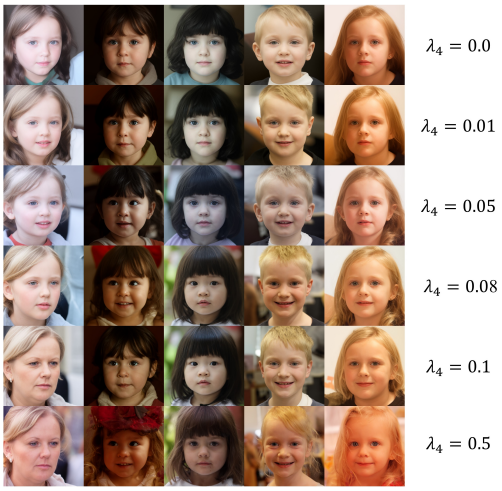

Figure 13: Visualized ablations of λ_4, the weight coefficient of $\mathcal{L}_{hfmse}$ on 10-shot FFHQ → Babies.

λ_4	Intra-LPIPS (↑)	FID (↓)
0.0	0.592 ± 0.031	70.40
0.01	0.594 ± 0.038	66.31
0.05	0.599 ± 0.024	**48.92**
0.08	0.607 ± 0.025	55.88
0.1	0.603 ± 0.031	59.28
0.5	$\mathbf{0.612 \pm 0.023}$	70.26

Table 10: Intra-LPIPS (↑) and FID (↓) results of adapted models trained on 10-shot FFHQ → Babies with different λ_4, the weight coefficient of $\mathcal{L}_{hfmse}$.

In addition, we add the visualized ablations of DomainStudio on T2I generation using houses in the ink painting style as an example in Fig. 9. Without relative distances preservation and high-frequency details enhancement, DomainStudio degrades to DreamBooth (Ruiz et al., 2023), which is designed to preserve key features of the subjects in training samples. As a result, it overfits and fails to achieve domain-driven generation. DomainStudio without high-frequency details enhancement applies pairwise similarity loss to relieve overfitting and guide adapted models to learn the knowledge of target domains while preserving source subjects corresponding to text prompts. The

full DomainStudio approach adds high-frequency details enhancement and preserves more details learned from source and training samples.

F LIMITATIONS AND SOCIETAL IMPACT

Limitations Despite the compelling results of our approach, it still has some limitations. All the datasets used for unconditional image generation in this paper share the resolution of 256×256. The experiments of DomainStudio are conducted on NVIDIA RTX A6000 GPUs (48 GB memory of each). However, the batch size on each GPU is still limited to 3. Therefore, it is challenging to expand our approach to larger image resolution. We will work on more lightweight few-shot image generation approaches for unconditional DDPMs. Despite that, the datasets used in this paper have larger resolution than many unconditional DDPM-based works (Giannone et al., 2022; Nichol & Dhariwal, 2021; Austin et al., 2021; Chen et al., 2023; Kingma et al., 2021; Zhang et al., 2022) which use datasets with resolution 32×32 and 64×64. For T2I generation, DomainStudio based on Stable Diffusion (Rombach et al., 2022) can synthesize images with super-resolution (512×512 or 1024×1024).

Besides, DomainStudio is trained on individual categories of subjects separately in this paper. In our experiments, we find that the adapted T2I models can also generate samples in target domains with other different subjects. Taking Fig. 7 as an example, the adapted models trained with the text prompt "A car in the [V] style" can produce some high-quality samples with other text prompts like "A temple in the [V] style." It indicates that the adapted T2I models generalize the concept of "the [V] style" across different subjects. However, we find that the generation quality of different subjects is not stable enough. Therefore, we still recommend independently training adapted models for the target subject to achieve more stable generation quality. We perceive the stable generalization of the domains learned from few-shot data across diverse subjects as future work. In addition, we recommend users to adjust the hyperparameters of the proposed optimization losses to achieve compelling results for different target domains.

Furthermore, this paper implements conditional DomainStudio based on T2I diffusion models. We will consider realizing conditional DomainStudio using varieties of prompts based on GLIGEN (Li et al., 2023) and ControlNet (Zhang & Agrawala, 2023) in future work.

Societal Impact DomainStudio proposed in this work could be applied to provide additional data for corner cases needed by downstream tasks and improve the efficiency of artistic creation by synthesizing images containing diverse subjects and sharing similar styles with training samples. We recognize that DomainStudio has potential risks of being misused to imitate existing works without permission since it only needs a few samples as training data.

G PERSONALIZATION OF DOMAINSTUDIO

DomainStudio is designed to realize domain-driven generation, which differs from modern subject-driven approaches like DreamBooth (Ruiz et al., 2023) and Textual Inversion (Gal et al., 2022). In this section, we further explore the personalization of DomainStudio to satisfy both domain-driven and subject-driven requests. Given two sets of images as reference for the target subject and domain respectively, we combine the proposed DomainStudio with DreamBooth to personalize domain-driven image generation.

The overview of the personalization of DomainStudio is illustrated in Fig. 14. Taking a personalized cat in the watercolor style as an example, we use text prompts: "a cat", "a [V] cat", and "a [V] cat in the [S] style" corresponding to the source samples, personalized subject, and personalized subject in the target domain.

We denote the encoded text prompts of source samples, personalized subjects, and personalized subjects in target domains as c_{sou}, c_{sub}, and c_{dom}. We have the reconstruction loss for the domain reference samples $x_0 \sim q(x_0)$ and subject reference images $x_1 \sim q(x_1)$ as follows:

$$\mathcal{L}_{simple}^{dom} = \mathbb{E}_{t,z_t^0,c_{dom},\epsilon^0}||\epsilon_{ada}(z_t^0, t, c_{dom}) - \epsilon^0||^2, \tag{22}$$

$$\mathcal{L}_{simple}^{sub} = \mathbb{E}_{t,z_t^1,c_{sub},\epsilon^1}||\epsilon_{ada}(z_t^1, t, c_{sub}) - \epsilon^1||^2, \tag{23}$$

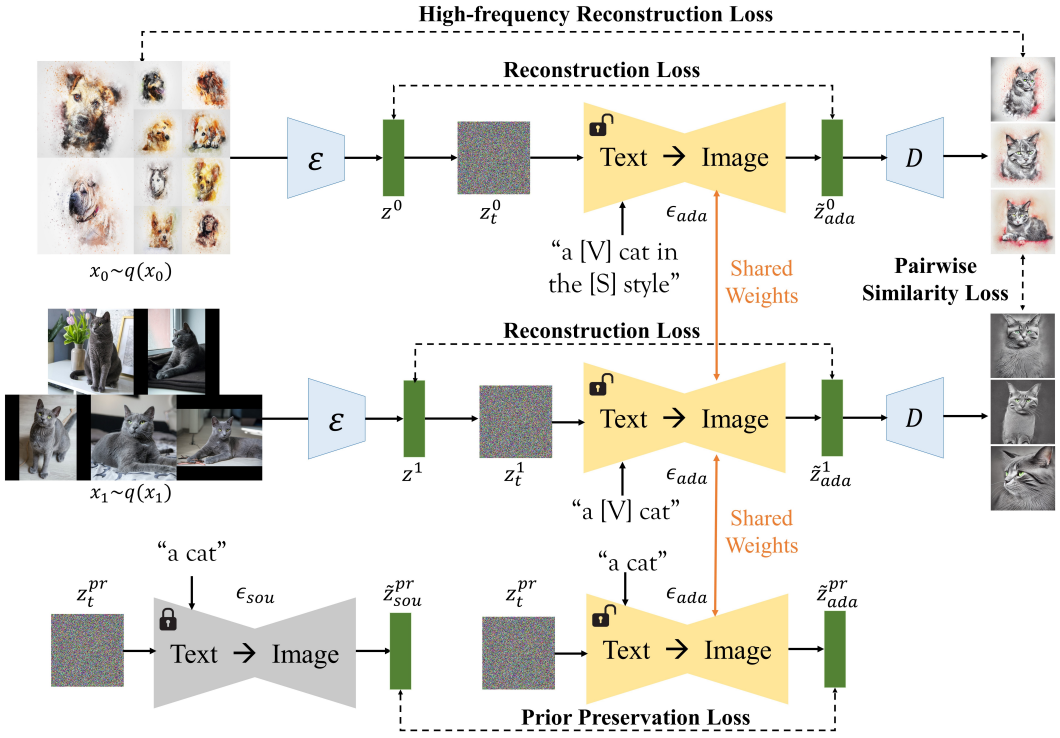

Figure 14: **Overview of the personalization of DomainStudio.** We combine DomainStudio with DreamBooth to achieve personalized domain-driven image generation.

where z_t^0 and ϵ_0 represent the noised compressed latent codes of domain reference samples and corresponding noises, z_t^1 and ϵ_1 represent the noised compressed latent codes of subject reference samples and corresponding noises, as shown in Fig. 14.

The pairwise similarity loss is computed between personalized subjects and personalized subjects in target domains. We build probability distributions using batches of denoised latent codes of personalized subjects $\left\{\tilde{z}_{ada}^{1,n}\right\}_{n=0}^{N}$ and denoised latent codes of personalized subjects in target domains $\left\{\tilde{z}_{ada}^{0,n}\right\}_{n=0}^{N}$ as shown in Eq. 24 and 25 and get the image-level pairwise similarity loss as shown in Eq. 26.

$$p_i^{dom} = sfm(\left\{sim(D(\tilde{z}_{ada}^{0,i}), D(\tilde{z}_{ada}^{0,j})\right\}_{\forall i \neq j}), \tag{24}$$

$$p_i^{sub} = sfm(\left\{sim(D(\tilde{z}_{ada}^{1,i}), D(\tilde{z}_{ada}^{1,j}))\right\}_{\forall i \neq j}), \tag{25}$$

$$\mathcal{L}_{img}^{per} = \mathbb{E}_{t,z_t^0,z_t^1,\epsilon^0,\epsilon^1} \sum_i D_{KL}(p_i^{dom} || p_i^{sub}). \tag{26}$$

Similarly, the probability distributions and pairwise similarity loss for high-frequency components are defined as Eq. 27, 28, and 29. The high-frequency reconstruction loss between personalized subjects and personalized subjects in target domains is defined as Eq. 30.

$$pf_i^{dom} = sfm(\left\{sim(hf(D(\tilde{z}_{ada}^{0,i})), hf(D(\tilde{z}_{ada}^{0,j})))\right\}_{\forall i \neq j}), \tag{27}$$

$$pf_i^{sub} = sfm(\left\{sim(hf(D(\tilde{z}_{ada}^{1,i})), hf(D(\tilde{z}_{ada}^{1,j})))\right\}_{\forall i \neq j}), \tag{28}$$

$$\mathcal{L}_{hf}^{per} = \mathbb{E}_{t,z_t^0,z_t^1,\epsilon^0,\epsilon^1} \sum_i D_{KL}(pf_i^{dom} || pf_i^{sub}), \tag{29}$$

$$\mathcal{L}_{hfmse}^{per} = \mathbb{E}_{t,z_t^0.z_t^1,\epsilon^0,\epsilon^1} ||hf(D(\tilde{z}_{ada}^0)) - hf(x_0)||^2. \tag{30}$$

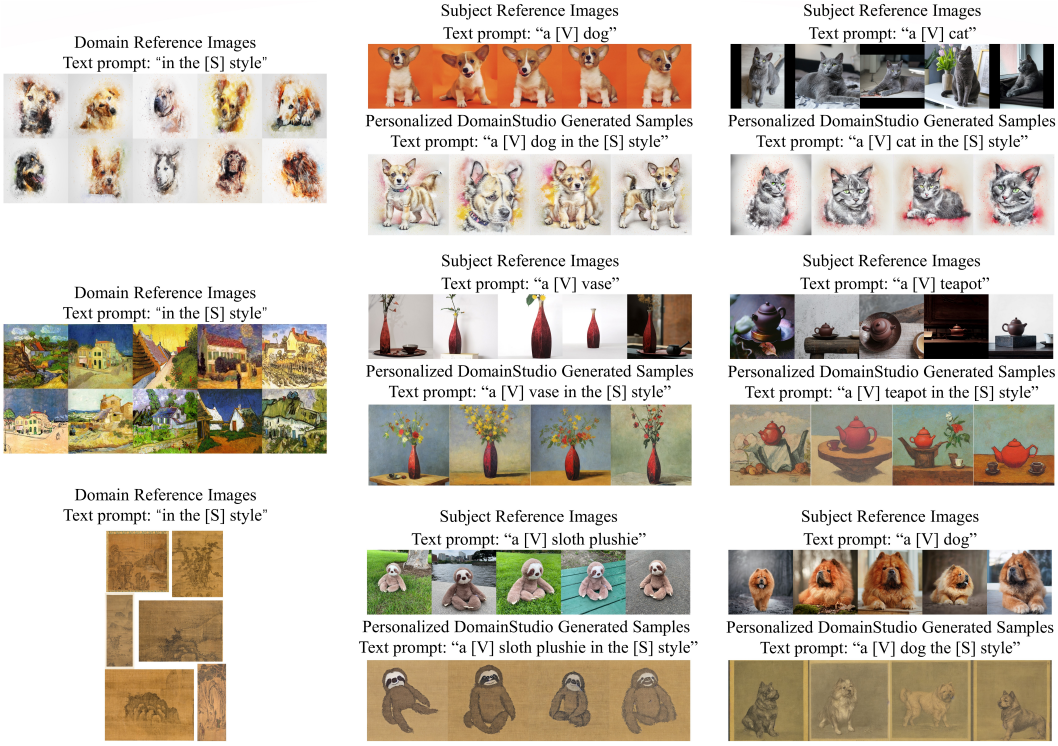

Figure 15: Samples produced by the personalization of DomainStudio using different sets of images as reference.

The overall optimization target of personalized DomainStudio can be expressed as:

$$\mathcal{L}^{per} = \mathcal{L}^{dom}_{simple} + \mathcal{L}^{sub}_{simple} + \lambda_{12}\mathcal{L}_{pr} + \lambda_2\mathcal{L}^{per}_{img} + \lambda_3\mathcal{L}^{per}_{hf} + \lambda_4\mathcal{L}^{per}_{hfmse}. \tag{31}$$

We empirically find that the setups of hyperparameters used for T2I adaptation setups (see Sec 3) also work well for the personalized DomainStudio.

We provide several personalized domain-driven generation samples containing diverse subjects and styles in Fig. 15. Our approach successfully adapts the personalized subject to target domains under the guidance of few-shot reference images. For instance, we adapt the reference dog and cat to the watercolor style (first row of Fig. 15). Besides, we synthesize the reference vase and teapot in Van Gogh's style using 10-shot Van Gogh houses as domain reference (second row of Fig. 15). The reference sloth plushie and dog are adapted to the ink painting style (third row of Fig. 15).

H More Details of Implementation

H.1 GAN-based Baselines

We employ several GAN-based few-shot image generation approaches as baselines for comparison with the proposed DomainStudio approach. Here we provide more details of these baselines. We implement all these approaches based on the same codebase of StyleGAN2 (Karras et al., 2020b). The source models are fine-tuned directly on the few-shot training datasets to realize TGAN (Wang et al., 2018). TGAN+ADA applies ADA (Karras et al., 2020a) augmentation method to the TGAN baseline. For FreezeD (Mo et al., 2020), the first 4 high-resolution layers of the discriminator are frozen following the ablation analysis provided in their work. The results of MineGAN (Wang et al., 2020), CDC (Ojha et al., 2021), AdAM (Zhao et al., 2022a), and RICK (Zhao et al., 2023) are produced through their official implementation. As for EWC (Li et al., 2020) and DCL (Zhao et al., 2022b), we implement these approaches following formulas and parameters in their papers since there is no official implementation. These GAN-based approaches are designed for generators (Wang et al., 2020; Li et al., 2020; Ojha et al., 2021; Zhao et al., 2022b;a; 2023) and discriminators

(Karras et al., 2020a; Mo et al., 2020; Zhao et al., 2022b; 2023) specially and cannot be expanded to DDPMs directly.

H.2 HAAR WAVELET TRANSFORMATION

Haar wavelet transformation contains four kernels including LL^T, LH^T, HL^T, HH^T, where L and H represent the low and high pass filters, respectively:

$$L^T = \frac{1}{\sqrt{2}}[1, 1], \quad H^T = \frac{1}{\sqrt{2}}[-1, 1]. \tag{32}$$

Fig. 16 visualizes several examples of Haar wavelet transformation. The low-frequency components LL contain the fundamental structures of images. High-frequency components including LH, HL, and HH contain rich details like contours and edges in images.

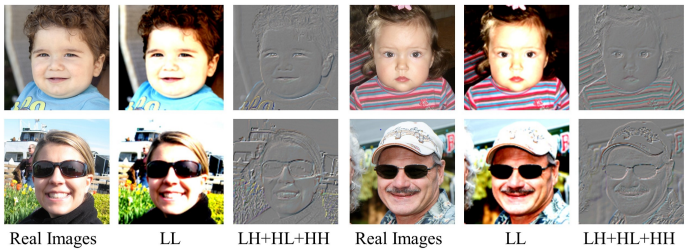

Figure 16: Visualization of the low and high-frequency components obtained with Haar wavelet transformation using images from Babies and Sunglasses as examples. LL represents the low-frequency components, and LH+HL+HH represents the sum of the high-frequency components.

H.3 UNCONDITIONAL DDPMS

We follow the model setups of DDPMs used in prior works (Nichol & Dhariwal, 2021) for LSUN 256^2 (Yu et al., 2015) datasets. All the DDPM-based models used in this paper are implemented based on the same codebase (Nichol & Dhariwal, 2021; Dhariwal & Nichol, 2021) and share the same model structure for fair comparison under different adaptation setups and optimization targets. All the source and training datasets are modified to the resolution of 256×256. The adapted models of DomainStudio are trained for 3K-5K iterations with a batch size of 24 on $\times 8$ NVIDIA RTX A6000 GPUs. We use a max diffusion step T of 1000 and a dropout rate of 0.1. The models are trained to learn the variance with $\mathcal{L}_{vlb}$. The Adam optimizer (Kingma & Ba, 2014) is employed to update the trainable parameters. We set the learning rate as 0.001 and apply the linear noise addition schedule. Besides, we use half-precision (FP16) binary floating-point format to save memory and make it possible to use a larger batch size in our experiments (batch size 6 for directly fine-tuned DDPMs and batch size 3 for DomainStudio per NVIDIA RTX A6000 GPU). All the results produced by DDPM-based models in this paper follow the sampling process proposed in Ho et al. (2020) (about 21 hours needed to generate 1000 samples on a single NVIDIA RTX A6000 GPU) without any fast sampling methods (Song et al., 2021; Zhang et al., 2022; Lu et al., 2022a;b; Zhang & Chen, 2022; Karras et al., 2022). The weight coefficient λ_2, λ_3, and λ_4 are set as 0.5, 0.5, 0.05 for the quantitative evaluation results of DomainStudio listed in Table 4 and 5.

H.4 T2I DDPMS

The adapted models of DomainStudio are trained for 1200-1500 iterations with a batch size of 4 on a single NVIDIA RTX A6000 GPU. We follow DreamBooth (Ruiz et al., 2023) to set the learning rates of DomainStudio ranging from 1e-6 to 5e-6. LoRA (Hu et al., 2021) uses the learning rate of 1e-4 and trains adapted models for 500 iterations. Experiments of DreamBooth and DomainStudio share the same hyperparameters in training for fair comparison. Textual Inversion (Gal et al., 2022) sets the learning rate as 5e-4 and trains text prompts for 2K-3K iterations. The image resolution used for training is 256×256. When predicting the original images with the predicted noises, we can choose to follow Eq. 3 or to use DDIM (Zhang et al., 2022) sampling method to generate samples for computing pairwise similarity losses and high-frequency reconstruction losses. DDIM sampling needs more computational cost and achieves higher-quality samples.

Samples of the source DDPM pre-trained on FFHQ 256^2

Samples of the source DDPM pre-trained on LSUN Church 256^2

Figure 17: Samples produced by unconditional DDPMs trained on FFHQ 256^2 (Karras et al., 2020b) (300K iterations) and LSUN Church 256^2 (Yu et al., 2015) (250K iterations).

I Unconditional Source Models

We train DDPMs on FFHQ 256^2 (Karras et al., 2020b) and LSUN Church 256^2 (Yu et al., 2015) from scratch for 300K iterations and 250K iterations as source models for DDPM adaptation, which cost 5 days and 22 hours, 4 days and 22 hours on $\times 8$ NVIDIA RTX A6000 GPUs, respectively. Samples produced by these two source models can be found in Fig. 17.

Models	FFHQ	LSUN Church
StyleGAN2	0.6619 ± 0.0581	0.7144 ± 0.0537
DDPM	$\mathbf{0.6631 \pm 0.0592}$	$\mathbf{0.7153 \pm 0.0513}$

Table 11: Average pairwise LPIPS (↑) results of 1000 samples produced by StyleGAN2 and DDPMs trained on FFHQ 256^2 and LSUN Church 256^2.

We randomly sample 1000 images with these two models to evaluate their generation diversity using the average pairwise LPIPS (Zhang et al., 2018a) metric, as shown in Table 11. For comparison, we also evaluate the generation diversity of the source StyleGAN2 (Karras et al., 2020b) models used by GAN-based baselines (Wang et al., 2018; Karras et al., 2020a; Mo et al., 2020; Wang et al., 2020; Li et al., 2020; Ojha et al., 2021; Zhao et al., 2022b). DDPMs trained on FFHQ 256^2 and LSUN Church 256^2 achieve generation diversity similar to the widely-used StyleGAN2 models.

Besides, we sample 5000 images to evaluate the generation quality of the source models using FID (Heusel et al., 2017). As shown in Table 12, DDPM-based source models achieve FID results similar to StyleGAN2 on the source datasets FFHQ 256^2 and LSUN Church 256^2.

Models	FFHQ	LSUN Church
StyleGAN2	7.71	8.09
DDPM	**7.00**	**6.06**

Table 12: FID (↓) results of StyleGAN2 and DDPMs trained on FFHQ 256^2 and LSUN Church 256^2.

J Inspiration of DomainStudio

J.1 Pairwise Similarity Loss

The proposed pairwise similarity loss designed for DDPMs is mainly inspired by the methods in contrastive learning (Oord et al., 2018; He et al., 2020; Chen et al., 2020) and CDC (Ojha et al., 2021), as discussed in Sec. 3.1.

It is worth noting that our approach is different from prior works, which contributes to the novelty of this work. GAN-based approaches depend on perceptual features in the generator and discriminator

to compute similarity and probability distributions. As for the proposed DomainStudio approach, the predicted input images $\tilde{x}_0$ calculated in terms of x_t and $\epsilon_\theta(x_t, t)$ (Equation 3) are applied in replacement of perceptual features used for GANs. Besides, the high-frequency components of $\tilde{x}_0$ are applied to pairwise similarity loss calculation for high-frequency details enhancement. DomainStudio directly uses image-level information to preserve the relative pairwise distances between adapted samples and during domain adaptation. Moreover, DomainStudio is compatible with both unconditional and T2I generation while prior GAN-based methods are totally unconditional.

We tried to use features in diffusion processes (Design A) and images of several diffusion steps (Design B) for pairwise similarity loss calculation. As shown in Table 13 (FID evaluation on FFHQ $\rightarrow$ Sunglasses, Intra-LPIPS evaluation on 10-shot FFHQ $\rightarrow$ Sunglasses), the proposed loss design using image-level information directly is simple, effective, inexpensive, and achieves the best quality and diversity. Here we do not include high-frequency details enhancement for fair comparison.

Method	FID (↓)	Intra-LPIPS (↑)	Time / 1K iterations (↓)
Ours	**37.92**	$\mathbf{0.59 \pm 0.02}$	***34min***
Design A	40.30	0.55 ± 0.03	$52min$
Design B	58.28	0.57 ± 0.06	$38min$

Table 13: Quantitative evaluation comparison between different designs for the pairwise similarity loss.

Image-level CDC on 10-shot FFHQ → Sunglasses

Image-level CDC on 10-shot FFHQ → Babies

Figure 18: Samples synthesized by CDC (Ojha et al., 2021) using image-level information on 10-shot FFHQ $\rightarrow$ Sunglasses and FFHQ $\rightarrow$ Babies.

As illustrated in Sec. 4, DomainStudio synthesizes more realistic images with fewer blurs and artifacts and achieves better generation diversity than current state-of-the-art GAN-based approaches (Ojha et al., 2021; Zhao et al., 2022b). We also try to use image-level information to replace the perceptual features for the GAN-based approach CDC (Ojha et al., 2021). However, we fail to avoid generating artifacts or achieve higher generation quality, as shown in Fig. 18. The proposed image-level pairwise similarity loss matches better with DDPMs than GANs.

J.2 High-frequency Reconstruction Loss

DDPMs learn target distributions mainly through mean values of predicted noises using the reweighted loss function (Equation 1). As a result, it is hard for DDPMs to learn high-frequency distributions from limited data, as shown in the smooth samples produced by models trained on limited data from scratch in Fig. 10. Therefore, we propose $\mathcal{L}_{hfmse}$ to strengthen the learning of high-frequency details from limited data during domain adaptation.

J.3 Prior Preservation Loss in DomainStudio

For unconditional image generation, we directly use samples produced from source models as reference to keep the diversity of adapted samples. For T2I generation, we employ the prior preservation loss proposed in DreamBooth (Ruiz et al., 2023) to avoid overfitting of source prompts (e.g., "a house"), based on which DomainStudio guides adapted models to maintain the diversity of subjects in adapted samples. In the training process, the original Stable Diffusion model (Rombach et al., 2022) is employed to generate source samples before fine-tuning adapted models. When fine-tuning adapted models, the original models are no longer needed. The prior preservation loss is equivalent to the reconstruction loss of source samples. Both source and adapted samples used for pairwise similarity losses (Eq. 7 and 12) computation are produced by adapted models with different text prompts. As a result, only adapted models are needed during the fine-tuning process, which saves GPU memory occupancy and improves training efficiency.

Figure 19: Samples produced by DomainStudio trained for different iterations on 10-shot FFHQ → Babies. All the visualized samples of different models are synthesized from fixed noise inputs.

J.4 Full Approach

Prior GAN-based approaches like CDC (Ojha et al., 2021) and DCL (Zhao et al., 2022b) aim to build a one-to-one correspondence between source and adapted samples. However, DomainStudio focuses on maintaining the distributions of subjects in source samples and generating realistic and diverse results following target distributions. Building one-to-one correspondences between source and adapted samples is not the first consideration of DomainStudio. As illustrated in Sec. 3, since we cannot build cross-domain correspondence with fixed noise inputs due to different conditions, we directly use randomly denoised samples to build probability distributions for T2I generation and also achieve diverse and high-quality results. Besides, the high-frequency reconstruction loss (Eq. 13 and 14) also influences the one-to-one correspondence between source and adapted samples.

K DDPM Adaptation Process Analysis

This paper concentrates on the challenging few-shot generation tasks. When fine-tuning pre-trained DDPMs on target domains using limited data directly, too many iterations lead to overfitting and seriously degraded diversity. Fine-tuned models trained for about 10K iterations almost exclusively focus on replicating the training samples. Therefore, we train the directly fine-tuned DDPMs for 3K-4K iterations to adapt source models to target domains and maintain diversity. However, the directly fine-tuned DDPMs still generate coarse samples lacking details.

In Fig. 19, we provide samples produced by DomainStudio trained for different iterations on 10-shot FFHQ → Babies. We apply fixed noise inputs to different models for comparison. As the iterations increase, the styles of the generated images become closer to the training samples. Images synthesized from the same noise inputs as Fig. 20 are included in red boxes. In addition, the detailed evaluation of cosine similarity is added in Fig. 21. The source samples are adapted to the target domain while keeping relatively stable cosine similarity. Compared with the directly fine-tuned DDPMs, DomainStudio has a stronger ability to maintain generation diversity and achieve realistic results containing rich details. Nonetheless, too many iterations still lead to the replication of training samples. Therefore, we recommend choosing suitable iterations for different adaptation setups (e.g., 4K-5K iterations for 10-shot FFHQ → Babies) to adapt the pre-trained models to target domains naturally and guarantee the high quality and great diversity of generated samples.

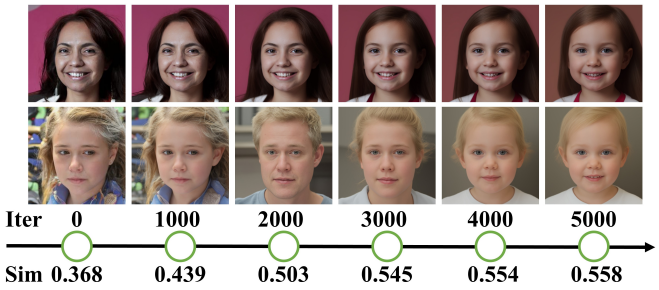

Figure 20: Two samples synthesized from fixed noise inputs by the directly fine-tuned DDPM on 10-shot FFHQ → Babies become more and more similar throughout training, as shown by the increasing cosine similarity computed with RGB values.

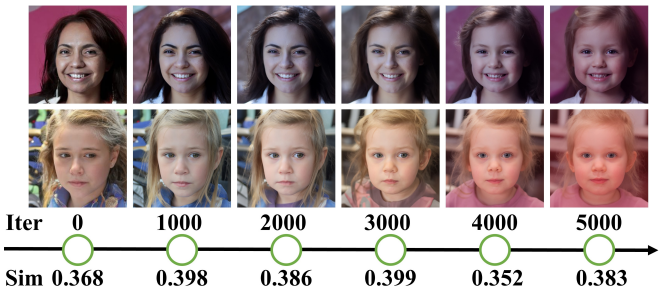

Figure 21: Two samples synthesized from fixed noise inputs by DomainStudio on 10-shot FFHQ → Babies. DomainStudio keeps the relative pairwise distances during domain adaptation and achieves diverse results containing high-frequency details.

L Additional Comparison with Related Works

Moon et al. (2022) investigates unconditional DDPMs fine-tuned with 800-1K images, which is a lot more than our work. In addition, we also explore DDPMs trained from scratch with extremely limited data.

ZADIS (Sohn et al., 2023b) and StyleDrop (Sohn et al., 2023a) are contemporary to this paper and share similar targets with the proposed DomainStudio approach. ZADIS is based on MaskGIT (Chang et al., 2022) and learns visual prompts for target domains/styles. In this way, ZADIS realizes compositional image synthesis with disentangled prompts for style and subjects. StyleDrop is based on MUSE (Chang et al., 2023) and synthesizes images with user-provided styles using reference images and descriptive style descriptors for training under the guidance of CLIP (Radford et al., 2021) scores and human feedback. DomainStudio is designed for DDPMs and compatible with typical unconditional DDPMs (Sohl-Dickstein et al., 2015; Ho et al., 2020) and modern large T2I models like Stable Diffusion (Rombach et al., 2022). DomainStudio aims to learn the domain knowledge from training samples, which may be artistic styles or properties like sunglasses. In addition, DomainStudio is also qualified for personalized domain-driven generation, as shown in Appendix G. In addition, Custom Diffusion (Kumari et al., 2023) also provides some examples of learning artistic styles. As shown in their paper, Custom Diffusion tends to combine the instances in style reference images with target instances mentioned in text prompts directly like Textual Inversion (Gal et al., 2022) and DreamBooth (Ruiz et al., 2023). When learning both styles and concepts, Custom Diffusion fails to preserve key features of learned concepts or adapt learned concepts to target styles naturally. The personalized DomainStudio approach achieves apparently better results.

M Additional Visualized Samples

Unconditional Image Generation We show all the 10-shot datasets used in this paper for unconditional few-shot image generation tasks in Fig. 22, including 4 target domains corresponding to

Figure 22: All the 10-shot datasets used for unconditional image generation, including 4 target domains corresponding to FFHQ and 2 target domains corresponding to LSUN Church.

Figure 23: Unconditional image generation samples comparison between DomainStudio and directly fine-tuned models.

the source datasets FFHQ (Karras et al., 2020b) and 2 target domains corresponding to the source datasets LSUN Church (Yu et al., 2015).

We visualize the samples of DomainStudio on 10-shot FFHQ $\rightarrow$ Babies, FFHQ $\rightarrow$ Sketches, and LSUN Church $\rightarrow$ Haunted houses in the bottom row of Fig. 23. DomainStudio produces more diverse samples containing richer high-frequency details than directly fine-tuned DDPMs. For example, DomainStudio generates babies with various detailed hairstyles and facial features.

Besides, we provide image generation samples of GAN-based baselines and DDPM-based approaches on 10-shot FFHQ $\rightarrow$ Sunglasses, FFHQ $\rightarrow$ Babies, FFHQ $\rightarrow$ Raphael's paintings, and LSUN Church $\rightarrow$ Landscape drawings in Fig. 24 26, 27, and 28 as supplements to Fig. 5. All the samples of GAN-based approaches are synthesized from fixed noise inputs (rows 1-9). Samples of the directly fine-tuned DDPM and DomainStudio are synthesized from fixed noise inputs as well (rows 10-11). DDPMs are more stable and less vulnerable to overfitting than GANs. Directly fine-tuned GANs easily overfit and tend to generate samples similar to training samples when

Figure 24: 10-shot unconditional image generation samples on FFHQ → Sunglasses.

Figure 25: 10-shot unconditional image generation samples of DomainStudio compared with AdAM and RICK on FFHQ → Sunglasses. Samples of AdAM and RICK are directly borrowed from the publications.

training data is limited (see samples of TGAN (Wang et al., 2018)). Directly fine-tuned DDPMs can still keep a measure of generation diversity under the same conditions. Besides, DDPM-based approaches relieve the generation of blurs and artifacts. However, directly fine-tuned DDPMs tend to produce too smooth results lacking high-frequency details and still face diversity degradation. DomainStudio generates more realistic results containing richer high-frequency details than GAN-based baselines under all these unconditional adaptation setups. We further provide comparison between our approach and visualized samples provided in AdAM (Zhao et al., 2022a) and RICK Zhao et al. (2023) in Fig. 25. We find incomplete structures of sunglasses and unnatural blurs and artifacts in the background and human faces in the samples of AdAM and RICK. DomainStudio avoids these problems and achieves better visual effects.

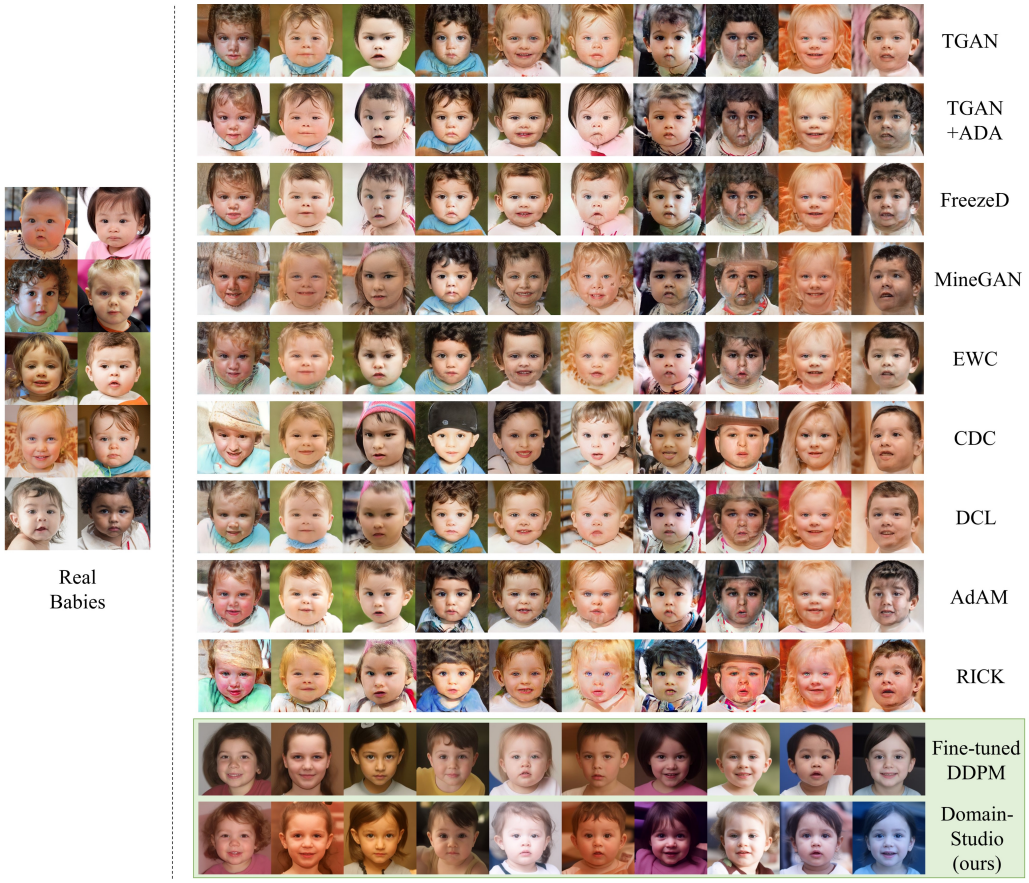

Figure 26: 10-shot unconditional image generation samples on FFHQ → Babies.

When the source and training samples are unrelated in unconditional image generation, DomainStudio is designed to preserve the subjects in source samples, leading to a target different from GAN-based few-shot image generation methods. As shown by the adaptation of LSUN Church → Landscape drawings in Fig. 28, DomainStudio preserves diverse church structures and adapts them to the style of landscape drawings. GAN-based baselines fail to adapt to the target domain naturally, resulting in low-quality samples full of blurs and artifacts.

T2I Generation As illustrated in Sec. 4, DomainStudio is capable of adapting the subjects prompted in text prompts to the style of few-shot training samples. However, baselines like DreamBooth (Ruiz et al., 2023) and Textual Inversion (Gal et al., 2022) fail to produce reasonable adapted samples. Similar phenomena can be found for baselines trained on 10-shot Wrecked cars and 4-shot Watercolor paintings, as shown in Fig. 29 and 30. Textual Inversion synthesizes car samples with text prompts of train or house. DreamBooth overfits and generates samples similar to few-shot data. It generates train and house samples containing wrecked cars instead of wrecked trains and houses like DomainStudio. In addition, we add "haunted" samples containing subjects including houses, temples, cars, and buses produced by adapted models trained through DomainStudio using 10-shot Haunted houses as training data in Fig. 31. We employ LoRA (Hu et al., 2021) as another baseline and provide the qualitative results in Fig. 32. LoRA also suffers from overfitting or underfitting in domain-driven generation like DreamBooth.

Fig. 33 shows the results of DomainStudio on T2I generation using a single image as training data. It's hard to define the target domain accurately with a single image. We recommend using 4-10 images to realize diverse, high-quality, and stable domain-driven T2I generation.

Figure 27: 10-shot unconditional image generation samples on FFHQ → Raphael's paintings.

N Computational Cost

Unconditional DDPMs The computational cost of unconditional DDPMs and DomainStudio approach are listed in Table 14. DomainStudio costs 24.14% more training time than the original unconditional DDPMs. DDPMs trained from scratch need about 40K iterations to achieve reasonable results, even if they can only replicate the training samples. DomainStudio utilizes models pre-trained on related source datasets to accelerate convergence (about 3K-5K iterations) and significantly improve generation quality and diversity. Compared with directly fine-tuned DDPMs, DomainStudio is not overly time-consuming and achieves more realistic results.

T2I DDPMs Here we only count the time of fine-tuning models. The computational cost of generating source samples is not included. The computational cost of LoRA Hu et al. (2021), Textual Inversion (Gal et al., 2022), DreamBooth (Ruiz et al., 2023), and DomainStudio on T2I generation are listed in Table 15. DomainStudio needs image-level information during training, while DreamBooth only needs latent-level computation. It makes DomainStudio more time-consuming. However, DomainStudio tackles a different task of domain-driven generation and achieves compelling results with acceptable computational cost. Building pairwise similarity losses based on the latent space may be a promising direction to accelerate the training of DomainStudio on T2I models.

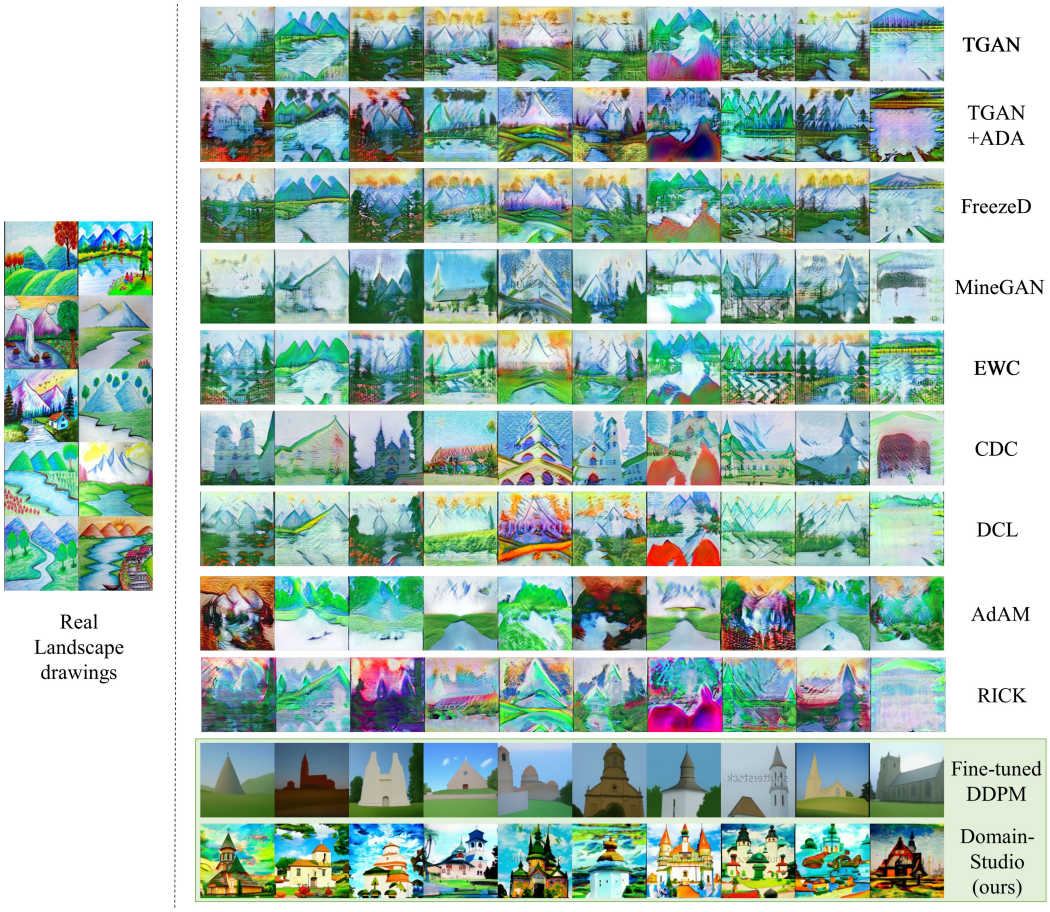

Figure 28: 10-shot unconditional image generation samples on LSUN Church → Landscape drawings.

Figure 29: Qualitative comparison of domain-driven T2I generation trained on 10-shot Wrecked cars.

Approaches	Time Cost
DDPMs	29 min
DomainStudio	36 min

Table 14: The time cost of directly fine-tuning and DomainStudio trained for 1K iterations on $\times 8$ NVIDIA RTX A6000 GPUs (image resolution: 256×256).

Approaches	Time Cost
LoRA (Hu et al., 2021)	4 min
Textual Inversion (Gal et al., 2022)	22 min
DreamBooth (Ruiz et al., 2023)	7 min
DomainStudio	15 min

Table 15: The time cost of Textual Inversion, DreamBooth, and DomainStudio trained for 1K iterations on a single NVIDIA RTX A6000 GPU (image resolution: 256×256).

Figure 30: Qualitative comparison of domain-driven T2I generation trained on 4-shot Watercolor paintings.

Figure 31: Visualized samples of DomainStudio trained on 10-shot Haunted houses.

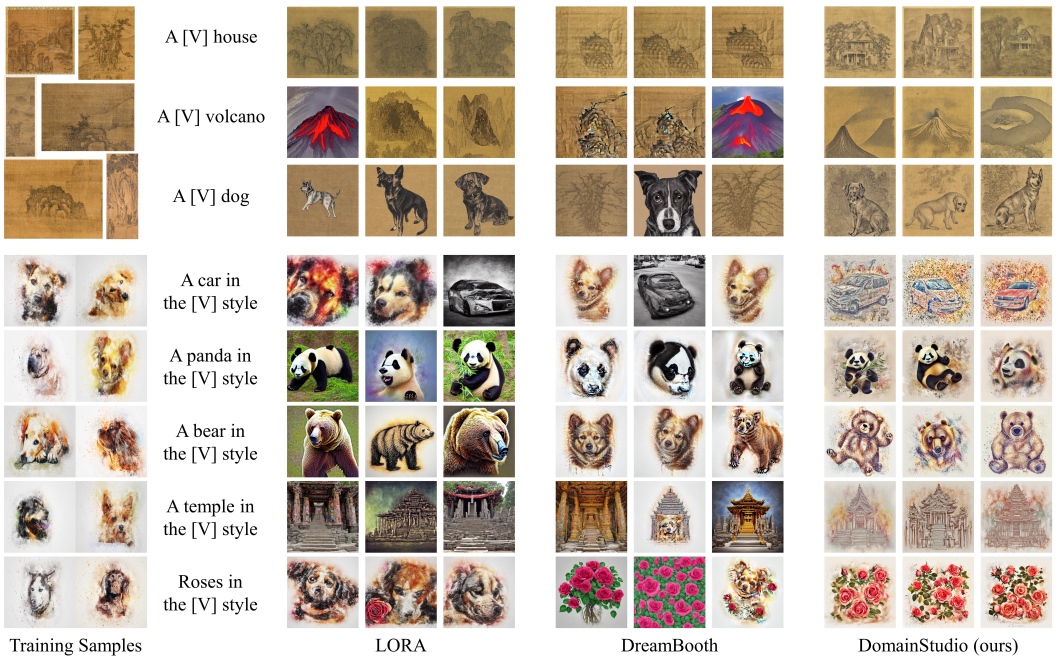

Figure 32: Qualitative comparison between LoRA, DreamBooth, and DomainStudio.

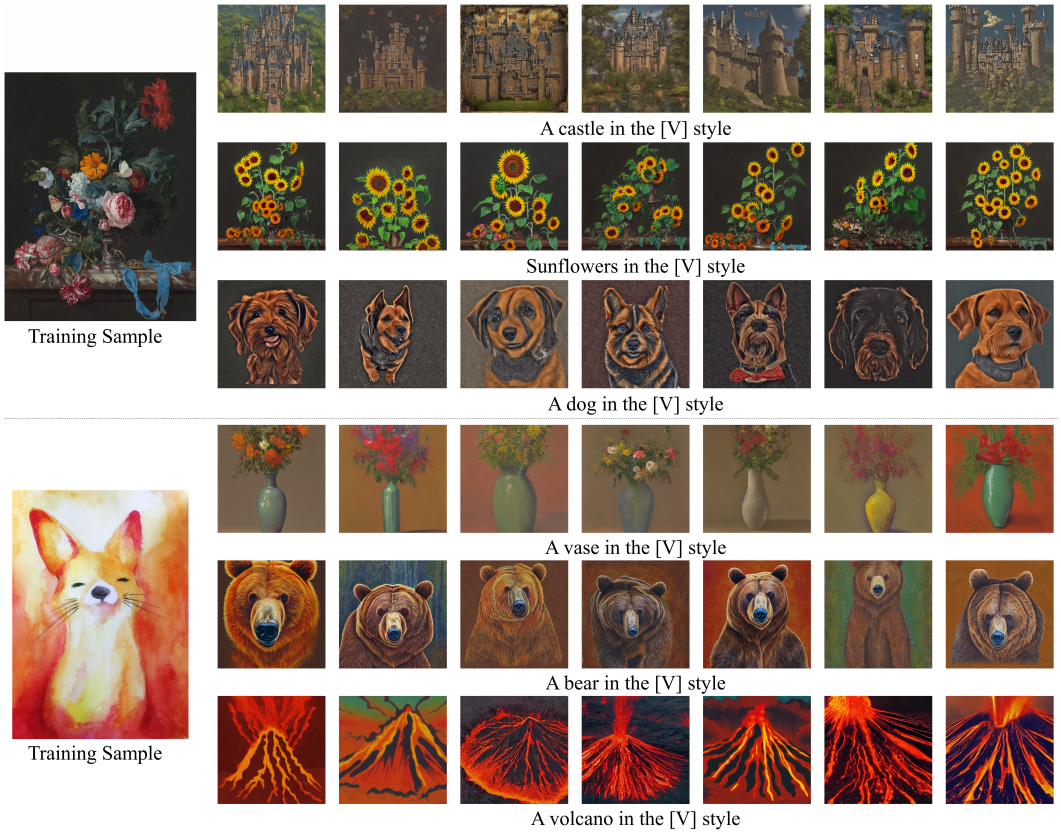

Figure 33: 1-shot T2I generation samples of DomainStudio using different source samples.