EXPLORING FEW-SHOT IMAGE GENERATION WITH MINIMIZED RISK OF OVERFITTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot image generation (FSIG) using deep generative models (DGMs) presents a significant challenge in accurately estimating the distribution of the target domain with extremely limited samples. Recent work has addressed the problem using a transfer learning approach, *i.e.* fine-tuning, leveraging a DGM that pre-trained on a large-scale source domain dataset, and then adapting it to the target domain with very limited samples. However, despite various proposed regularization techniques, existing frameworks lack a systematic mechanism to analyze the degree of overfitting, relying primarily on empirical validation without rigorous theoretical grounding. We present Few-Shot Diffusion-regularized Representation Learning (FS-DRL), an innovative approach designed to minimize the risk of over-fitting while preserving distribution consistency in target image adaptation. Our method is distinct from conventional methods in two aspects: First, instead of fine-tuning, FS-DRL employs a novel scalable Invariant Guidance Matrix (IGM) during the diffusion process, which acts as a regularizer in the feature space of the model. This IGM is designed to have the same dimensionality as the target images, effectively constraining its capacity and encouraging it to learn a low-dimensional manifold that captures the essential structure of the target domain. Second, our method introduces a controllable parameter called sharing degree, which determines how many target images correspond to each IGM, enabling a fine-grained balance between overfitting risk and model flexibility, thus providing a quantifiable mechanism to analyze and mitigate overfitting. Extensive experiments demonstrate that our approach effectively mitigates overfitting, enabling efficient and robust few-shot learning across diverse domains.

1 INTRODUCTION

In recent years, Deep Generative Models (DGMs) have achieved remarkable breakthroughs in the generation of high-quality and diverse samples across various domains (Higgins et al., 2016; Karras et al., 2019; Song et al., 2020b; Ruiz et al., 2023). However, reliance on extensive data presents a significant challenge in scenarios where data is scarce (Abdollahzadeh et al., 2023). To address this issue, Few-Shot Image Generation (FSIG) methods (Wang et al., 2018; Zhao et al., 2022) have emerged, aiming to generate diverse images with limited training samples.

Most FSIG methods rely on fine-tuning a DGM, typically a generative adversarial network (GAN) (Goodfellow et al., 2014), which pretrained on a larger and "similar" dataset (Ojha et al., 2021; Zhu et al., 2022; Zhao et al., 2022; 2023). However, this fine-tuning process, which involves adjusting the generator $p_\theta(z)$ to minimize the loss in the target domain $\mathcal{Y}$, $\min_\theta \mathbb{E}_{(z\sim\mathcal{N}(0,I),y\sim\mathcal{Y})}\left[\mathcal{L}(p_\theta(z),y)\right]$, often leads to overfitting, visual artifacts, and catastrophic forgetting (Saito et al., 2017; Radford et al., 2015; Kirkpatrick et al., 2017) when only a few samples are available.

More recently, Diffusion Models (DMs) (Ho et al., 2020; Song et al., 2020b) have demonstrated remarkable success, surpassing GANs in image generation (Dhariwal & Nichol, 2021). Their inherent scalability and more stable training process allow DMs to be trained on larger datasets, resulting in superior generalization capabilities. This makes them particularly adept at tasks that require fine-grained detail manipulation, such as text-to-image translation (Saharia et al., 2022; Ramesh et al., 2021) and intricate image editing (Meng et al., 2021). Given these strengths, it is attractive to consider adapting DMs for FSIG, potentially offering superior solutions to existing GAN-dominated methods.

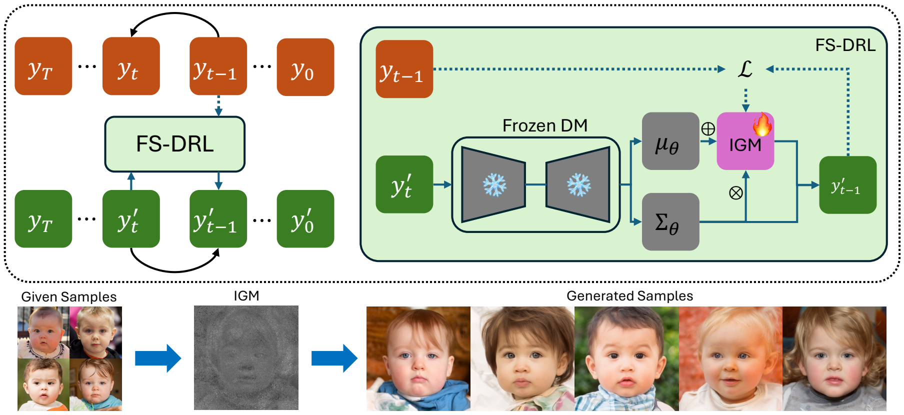

Figure 1: An illustration of FS-DRL, demonstrating how our method overcomes overfitting during IGM training, along with visual showcase. The dotted arrow (top) is used only during training.

However, directly applying current FSIG techniques such as regularization (Li et al., 2020; Ojha et al., 2021) and modulation (Zhao et al., 2022) to DMs proves challenging. The significantly larger number of parameters in DMs and their iterative nature not only fail to address the problems faced by GANs but may exacerbate overfitting and catastrophic forgetting issues (Abdollahzadeh et al., 2023). Consequently, we define the research question as follows: How can we adapt the pre-trained diffusion model to the target domain while minimizing the risk of overfitting?

To address this question, we present Few-Shot Diffusion-Regularized Representation Learning (FS-DRL), as shown in Fig. 1. Our method consists of three main contributions:

Firstly, we introduce a novel framework to adapt a pretrained DM to a specific domain. Unlike other approaches that attempt to modify the generator (Wang et al., 2018; Ojha et al., 2021; Zhao et al., 2023), our method is designed to "influence" the generation process. Specifically, given a target domain $\mathcal{Y}$, our method converts the unconditional generation process to a conditional one, and at the diffusion time t, the objective is thus $\min_\theta \mathbb{E}_{(q(y_t|y)),y\sim\mathcal{Y}}\left[\mathcal{L}(p_\theta(y_t|\mathcal{Y}), y)\right]$. We find that introducing a non-adaptive module, which we call the Invariant Gradient Matrix (IGM), is sufficient to achieve our objective by guiding the generation process.

Secondly, we theoretically demonstrated that this IGM is essentially equivalent to a "simplest" classifier in classifier-guided diffusion model (Song et al., 2020b). The weights can be seen as an "attention matrix", which determines the amount of "attention" different regions of the state should receive for a specific domain. Furthermore, we introduce a **Scalable** property for IGM, which allows flexible control over granularity. This scalability impacts the trade-off between generalization and specificity. Defining an IGM for multiple images provides high generalization with low overfitting risk, while a single IGM per image offers high specificity but increases overfitting risk.

Thirdly, we propose two optimization techniques that significantly enhance the performance of our method in Few-Shot Image Generation (FSIG). The introduction of percentile gradient clipping and simplified loss function allows our approach to achieve comparable results to state-of-the-art methods, with particularly notable improvements in mode coverage. Additionally, we conducted experiments on further parameter reduction, exploring the trade-offs between model complexity and performance.

We summarize the structure of the paper as follows. In Sec. 3.1, we provide a preliminary introduction to diffusion models and formalize the notion we used in this paper. We then introduce the details of our proposed method FS-DRL (Sec. 3.2) with theoretical analysis (Sec. 3.3) and two optimization strategies (Sec. 3.4. In Sec. 4, we demonstrate the effectiveness of our proposed method through empirical comparisons with the baseline, and a comprehensive component analysis.

2 Related Works

Few-Shot Image Generation Conventional approaches typically apply fine-tuning a Generative Model pre-trained on a large dataset of a similar domain (Bartunov & Vetrov, 2018; Wang et al., 2018; Clouâtre & Demers, 2019). However, full model fine-tuning typically leads to mode collapse

(Hu et al., 2023). To mitigate this, various selective fine-tuning techniques have been proposed. These include updating only part of the model, *e.g.* freezing the discriminator of GAN (Noguchi & Harada, 2019; Mo et al., 2020), preserving crucial pretrained weights identified by the modulation method and Fisher Information (Li et al., 2020; Zhao et al., 2022; 2023), and maintaining structural similarity between source and target domain distributions (Ojha et al., 2021; Xiao et al., 2022; Hu et al., 2023). GenDA (Mondal et al., 2022) first utilize the representation learning method for FSIG, however, their method is limited to StyleGAN (Karras et al., 2019) as it requires a "short" explicit latent code. CRDI (Cao & Gong, 2024) is the most similar work to ours. However, we showed that their framework can be regarded as a special case of ours with the highest degree of overfitting in Sec. 3.3.

DM for Representation Learning There are three main approaches which are close to our proposed method: (1) Diffusion Models with AutoEncoder (VAE) (Kingma & Welling, 2013), this approaches including D2C (Sinha et al., 2021), Diff-AE (Preechakul et al., 2022), DiTi (Yue et al., 2024) *et al*., which also be able to generate given only a few samples ($\geq$ 100), however, these methods require to train a Latent DMs from scratch to adapt a pre-trained VAE, which cause significant computational resources and cannot be applied to varies pre-trained diffusion model. (2) Text-to-Image Diffusion Model (DM), because of the high scalability of DMs, many LMMs such as DALL-E (Ramesh et al., 2021) and Stable Diffusion (Rombach et al., 2022) are also applied to FSIG task. However, existing multimodal foundation models have limited capacity for generating images of unseen categories in inferring. Although methods such as DreamBooth (Ruiz et al., 2023) can generate samples from a few shots, they are limited to adapting at the subject level. (3) Diffusion Inversion, which can be further decomposed into two methods, training-free method including SDEdit (Meng et al., 2021), Edict (Wallace et al., 2023) *et al.* and training-required method including Textual Inversion (Gal et al., 2022), MCG (Chung et al., 2022) *et al*., these methods are mainly for Image Editing task which requires deterministic inversion, hence not suitable for FSIG as the diversity is a key point.

3 METHODOLOGY

3.1 PRELIMINARIES

Diffusion Model Denoising Diffusion Probabilistic Model (Ho et al., 2020) (DDPM) is a latent variable model that learns to sample from a distribution by learning to iteratively denoise samples. The forward process $q(x_{0:T})$ adds noise to the sample x_0 as

$$q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)=\mathcal{N}\left(\mathbf{x}_t ; \sqrt{1-\beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}\right) \tag{1}$$

where β_t is pre-defined to control the variance schedule. Song et al. (2020b) and Ho et al. (2020) shown that the reverse process can be converted to a generative model by sampling $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and transforming incrementally into a data manifold as $p_\theta\left(\mathbf{x}_{0:T}\right)=p\left(x_T\right) \prod_{t=1}^T p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$, where

$$p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)=\mathcal{N}\left(\mathbf{x}_{t-1} ; \mu_\theta\left(\mathbf{x}_t, t\right), \Sigma_\theta\left(\mathbf{x}_t, t\right)\right) \tag{2}$$

Here μ_θ and Σ_θ are the outputs of a neural network. Furthermore, by using the reparameterization trick and Tweddie's formula (Stein, 1981), we can get two equivalent interpretations

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_t, t\right)=\frac{1}{\sqrt{\alpha_t}} \boldsymbol{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t} \sqrt{\alpha_t}} \boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_t, t\right)=\frac{1}{\sqrt{\alpha_t}} \boldsymbol{x}_t+\frac{1-\alpha_t}{\sqrt{\alpha_t}} \boldsymbol{s}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_t, t\right) \tag{3}$$

where α_t is mean coefficient defined as $1-\beta_t$, $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_t, t\right)$ and $\boldsymbol{s}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_t, t\right)$ are noise network and score network, respectively. See Luo (2022) and Song et al. (2020b) for complete deviation.

3.2 FEW-SHOT DIFFUSION-REGULARIZED REPRESENTATION

Definition 1. *(Target Domain Adaptation) Given a diffusion model trained on a source domain dataset $\mathcal{X}$, we say that the diffusion model is adapted to target domain $\mathcal{Y}$ with degree η at t when $\mathbb{E}_{x_0 \in \mathcal{X}, y_0 \in \mathcal{Y}}\left[\mathcal{M}(x_0, y_0, t)\right] \geq \eta$, where domain adaptation measure $\mathcal{M}(x_0, y_0, t)$ is defined as:*

$$\mathcal{M}(x_0, y_0, t):=\frac{1}{2}\left(\underset{q(\mathbf{x}_t \mid \mathbf{x}_0)}{\mathbb{E}}[\underbrace{\mathbb{I}\left(\left|\hat{\mathbf{x}}_0-\mathbf{x}_0\right|>\delta\right)}_{adaptation}]+\underset{q(\mathbf{y}_t \mid \mathbf{y}_0)}{\mathbb{E}}[\underbrace{\mathbb{I}\left(\left|\hat{\mathbf{y}}_0-\mathbf{y}_0\right|<\delta\right)}_{reconstruction}]\right) \tag{4}$$

where $\hat{\mathbf{x}}_0=p_\theta(\mathbf{x}_{t:T})$, $\hat{\mathbf{y}}_0=p_\theta(\mathbf{y}_{t:T})$, indicator function $\mathbb{I}(\cdot)$ and a given threshold δ.

Specifically, $\mathcal{M}(x_0, y_0, t)$ measures the target domain adaptation degree by assessing how well a noised sample $\mathbf{y}_t$ obtained from $\mathbf{y}_0$ is reconstructed and how likely a noised sample $\mathbf{x}_t$ obtained from $\mathbf{x}_0$ is falsely reconstructed. In the context of the FSIG task, the source domain and target domain differ in attribute, we can assume that, initially, the target domain adaptation degree is close to 0, as the model is trained solely on the source domain. To increase the adaptation degree and enable effective generation in the target domain, we apply the conditioning mechanism for diffusion models.

Few-Shot Image Generation can be considered as a fine-grained conditional generating. Specifically, a conditional generative model can be formulated as $p_t(\mathbf{x}_t \mid \mathbf{y})$, where $\mathbf{y}$ is the condition (given samples in FSIG task). Per Bayes' theorem, $p_t(\mathbf{x}_t \mid \mathbf{y}) \propto p_t(\mathbf{x}_t)p_t(\mathbf{y} \mid \mathbf{x}_t)$. Expressing this relationship as a score function, a score-based conditional diffusion model is described as:

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t \mid \mathbf{y}) = \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{y} \mid \mathbf{x}_t) \tag{5}$$

where $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ and $\nabla_{\mathbf{x}_t} \log p(\mathbf{y} \mid \mathbf{x}_t)$ are respectively the scores of an unconditional DM and a time-dependent intermediate state (x_t) classifier. However, the distribution of x_t at different timestep of diffusion model is different, therefore raising the difficulty of training the classifier. To mitigate overfitting under few-shot, instead of choosing classifier with a simple structure, we propose replacing the time-dependent intermediate state classifier with a non-adaptive Invariant Gradient Matrix $\mathbf{G}(t)$. This matrix captures the essential characteristics of the target domain at each timestep t, without relying on the current state $\mathbf{x}_t$. Incorporating $\mathbf{G}(t)$ into the score function (Eq. 5), we obtain:

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t \mid \mathbf{y}) = \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) + \mathbf{G}(t) \tag{6}$$

The Invariant Gradient Matrix (IGM) $\mathbf{G}(t)$ guides the sampling process towards the target domain, effectively capturing essential domain characteristics under few-shot setting while avoiding overfitting. The training loss associated with our definition of Target Domain Adaptation is defined as:

$$\mathcal{L}_{DA} = \mathop{\mathbb{E}}_{t, \mathbf{x}_0 \in \mathcal{X}, \mathbf{y}_0 \in \mathcal{Y}} \left[\left| \mathbf{y}_0 - \hat{p}_\theta \left(\sqrt{\bar{\alpha}_t} \mathbf{y}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t \right) \right| - \left| \mathbf{x}_0 - \hat{p}_\theta \left(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t \right) \right| \right] \tag{7}$$

where $\hat{p}_\theta$ is the pretrained diffusion model with our IGM, $|\cdot|$ denotes a distance metric.

3.3 Theoretical Analysis

We shall now provide the theoretical justification of our proposed method.

IGM Fundamentals Without loss of generality, let us consider the case at time t. To simplify the notation, we denote $\mathbf{c} = \mathbf{G}(t)$. According to Eq. 5 and 6, we have $\mathbf{c} = \nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x}_t)$, solving this differential equation yields $p(\mathbf{y} \mid \mathbf{x}_t) \propto \exp(\mathbf{c} \cdot \mathbf{x}_t)$. This equation defines a pixel-wise linear regression model followed by a softmax activation function, where each pixel of the intermediate state $\mathbf{x}_t$ is weighted by the corresponding element of the IGM. Intuitively, the IGM functions as an attention mechanism that determines how much "attention" or "importance" should be assigned to different regions of $\mathbf{x}_t$, conditioned on a specific target domain $\mathcal{Y}$. See Fig. 1 for an IGM visualization and Section C.1 for further explanation and more visual examples of IGM.

Overfitting Mitigation Strategy From a pixel-wise perspective, if each image of the target domain is assigned a unique IGM, it may lead to overfitting as the model can memorize the specific pixel. However, when an IGM is shared across multiple images, it effectively becomes a linear regression model fitting multiple data points, promoting better generalization. To balance model expressiveness and generalization, we introduce the IGM Sharing Degree, γ, representing the number of images that share an IGM. As γ increases from 1, the model shifts from potential overfitting toward better generalization, allowing for fine-tuned performance across diverse datasets. However, excessively high γ values can lead to underfitting. We provide an in-depth analysis of this trade-off in Sec. 4.1.

Theoretical Foundation of Domain Adaptation with IGM We develop a theoretical framework for domain adaptation in diffusion models, showing how our Invariant Gradient Matrix (IGM) guides the generative process from source domain to target domains towards the desired distribution.

Theorem 1. *Let $\mathbf{x}$ be a random variable following a normal distribution with mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$. If the conditional probability $p(\mathbf{y} \mid \mathbf{x})$ has the form $p(\mathbf{y} \mid \mathbf{x}) \propto \exp(c \cdot \mathbf{x})$, where c is a constant, then the conditional probability $p(\mathbf{x} \mid \mathbf{y})$ is also a normal distribution, and its posterior density is given by (See Section C.2 for the proof):*

$$p(x|y) = \frac{1}{p(y)\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - (\mu + c\sigma^2))^2}{2\sigma^2} + \frac{c^2\sigma^2}{2} + c\mu \right) \tag{8}$$

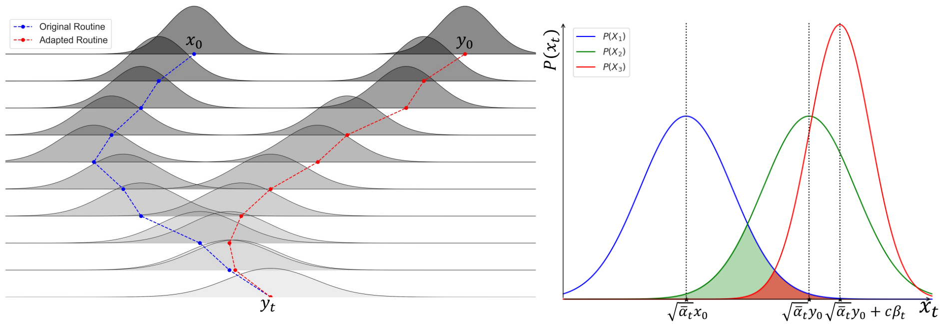

Figure 2: **Left**: Left: A density ridgeline plot showing an 1-D example of our method, transforming a standard normal distribution to a target distribution through an adapted diffusion process. **Right**: Zooming in a specific step from the left plot, the PDFs of x_t (blue), y_t (green) and adapted target domain sample y'_t (red) are shown. The adapted version reduces the overlapping area (green $\rightarrow$ red).

Remark 1. *According to Theorem 1, the conditional probability $p(\mathbf{x} \mid \mathbf{y})$ differs from the original distribution $p(\mathbf{x})$ in the following aspects:*

1. *Mean shift: The mean of the conditional probability shifts from the original mean $\boldsymbol{\mu}$ to $\mu + c\sigma^2$. This implies that the center of the distribution moves in the direction of c, and the distance of the shift is determined by the magnitude of $\boldsymbol{\sigma}$.*
2. *Scaled distribution height: The distribution is vertically scaled at each point by a factor of $\frac{1}{p(y)} \exp\left(\frac{c^2\sigma^2}{2} + c\mu\right)$, based on the observed data and the original hyper-parameters.*

For any samples $x_0 \in \mathcal{X}$ and $y_0 \in \mathcal{Y}$, Eq. 1 defines a forward process in which x_t and y_t progressively approach $\mathcal{N}(\mathbf{0}, \mathbf{I})$. This process ensures that samples from different domains converge to a common Gaussian distribution. The shared endpoint guarantees an overlap between the distributions of x_t and y_t at certain timesteps, despite the model not being trained on the target domain. Conversely, the reverse process starts from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and aims to recover the training samples. The Fokker-Planck equation (Risken, 1996) describes the evolution of probability density during this process:

$$\frac{\partial p(x,t)}{\partial t} = -\nabla_x \cdot (p(x,t)\nabla_x \log p(x,t)) + \frac{1}{2}\nabla_x^2 p(x,t) \tag{9}$$

The score function $\nabla_x \log p(x,t)$ learned by the model primarily captures the distribution of the source domain $\mathcal{X}$. Consequently, during the reverse process, this source domain-biased score function influences both x_t and y_t, causing the generated samples to gravitate towards the source domain distribution, even if y_t has already deviated from its intended trajectory. Intuitively, the learned probability flow acts as a “force” pulling samples towards the center of the source domain $\mathcal{X}$. Our proposed Invariant Gradient Matrix acts as a “counterforce”, steering the reverse process towards the target domain while mitigating influence from the source domain. A visual illustration is shown in Fig. 2. For more theoretical analysis from the probability flow point of view refer to Section C.3.

3.4 Optimization

In Section 3.3, we theoretically analyzed the feasibility of our proposed method. While leveraging a model trained on a source domain that closely resembles the target domain somewhat reduces the complexity of the task, employing a non-adaptive gradient matrix to generate out-of-distribution images still poses significant challenges. Therefore, in this section, we introduce two optimization strategies to further enhance the performance and generalization capability in the target domain.

Percentile Gradient Clipping The gradient matrix $\mathbf{G}(t)$ may contain gradient values $g_{i,j}(t)$ at certain pixels that represent noise or weakly correlated information between the source domain $\mathcal{X}$ and the target domain $\mathcal{Y}$. Accordingly, we introduce Percentile Gradient Clipping (PGC) as:

$$\hat{g}_{i,j}(t) = g_{i,j}(t) \cdot (|g_{i,j}(t)| \geq Q(\mathbf{G}(t), \rho)) \tag{10}$$

where $Q(\mathbf{G}(t), \rho)$ represents the ρ-th percentile of the gradient matrix $\mathbf{G}(t)$. PGC removes smaller gradients that are more likely to represent noise or weak correlations, while retaining stronger gradients potentially more informative for target domain adaptation. From an information-theoretic perspective, this process increases the ratio of effective information $\frac{I(\mathbf{G}(t);\mathcal{Y})}{H(\mathbf{G}(t))}$ in $\mathbf{G}(t)$. Here, $I(\mathbf{G}(t);\mathcal{Y})$ represents the mutual information between $\mathbf{G}(t)$ and $\mathcal{Y}$ and $H(\mathbf{G}(t))$ denotes the entropy of $\mathbf{G}(t)$, quantifying its informational uncertainty. Enhancing this ratio enables $\mathbf{G}(t)$ to more effectively capture common features across different domains (Ganin et al., 2016), potentially leading to better generalization in the target domain. For more detail and theoretical analysis see Section C.4.

Simplified Loss Function In Section 3.2, according to the definition of Target Domain Adaptation, we can express the domain adaptation loss function as Eq. 7, which aims to encourage the model to reverse intermediate states to the target domain $\mathcal{Y}$ instead of the source domain $\mathcal{X}$. However, experimental results suggest that this approach may suppress useful knowledge learned by the model in the source domain. Considering that the goal of FSIG is to select a source domain $\mathcal{X}$ that is close to the target domain $\mathcal{Y}$, we can simplify the loss function by emphasizing reconstruction ability:

$$\mathcal{L}_{DA} = \mathop{\mathbb{E}}_{\mathbf{y}_0 \in \mathcal{Y}} \left[\left| \mathbf{y}_0 - \hat{p}_\theta \left(\sqrt{\bar{\alpha}_t}\mathbf{y}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t \right) \right| \right] \tag{11}$$

This simplified loss function allows the model to retain useful knowledge learned from the source domain while adapting to the target domain. Intuitively, by minimizing the reconstruction error of target domain samples, the model naturally gravitates towards the target domain while preserving relevant information from the source domain to the greatest extent possible.

4 Experiments

Datasets and Baseline Following previous work (Wang et al., 2018; Li et al., 2020; Ojha et al., 2021), we used Flickr Faces HQ (FFHQ) (Karras et al., 2019) as the source domain datasets for all quantitative analysis, LSUN (Yu et al., 2015) and FFHQ for qualitative analysis. We applied our method to adapt to the following common target domains for comparisons to existing FSIG methods: FFHQ-Babies (Ojha et al., 2021), FFHQ-Sunglasses (Ojha et al., 2021), MetFaces (Karras et al., 2020), portrait paintings from the artistic faces dataset (Yaniv et al., 2019). We select three FSIG methods as baseline, including RICK (SOTA method) (Zhao et al., 2023), GenDA (SOTA representation learning method for GAN) (Mondal et al., 2022), CRDI (SOTA representation learning method for DM) (Cao & Gong, 2024). More methods comparison results are given in Section F.

Metrics We compute two commonly used metrics in FSIG, FID (Fréchet inception distance) (Heusel et al., 2017) and Intra-LPIPS (Intra-cluster pairwise Learned Perceptual Image Patch Similarity) Ojha et al. (2021), to quantitatively assess the quality and diversity of generated samples with respect to the target domain, respectively. We also calculate MC-SSIM (Mode Coverage Structural Similarity Index Measure) (Cao & Gong, 2024) which quantify the mode coverage for complex domain.

Implementation Details We used Guided Diffusion (Dhariwal & Nichol, 2021) framework from OpenAI and pretrained weight from Segmentation DDPM (Baranchuk et al., 2021). We utilized DDIM (Song et al., 2020a) with 25 inference steps to improve the efficiency while training. Model training is performed with 256 x 256 resolution and batch size 10 on a single A100/H100 GPU.

4.1 Sharing Degree: Balancing Generalization and Specificity

To validate the theoretical analysis presented in Sec. 3.3 regarding the impact of the IGM sharing degree on overfitting, we conducted experiments across three commonly used target domains in FSIG, Babies, Sunglasses and MetFaces. We applied our method for each domain at three different timestep periods $[t_s, t_e]$ during the diffusion process, varying the degree of IGM sharing. We evaluated the generated images using FID scores; the results are shown in Fig. 3. The IGM sharing degree, γ, ranges from 1 (one IGM per one image) to 10 (one IGM per ten images). We additionally fitted an Exponential Moving Average (EMA) curve (green line) to each graph to highlight the overall trend.

It can be observed that, for the target domain Babies and Sunglasses, the EMA of FID shows varying degrees of the U-shaped curve as the IGM sharing degree increases from 1 to 10. When $\gamma = 1$, the model exhibits the highest degree of overfitting, resulting in images generated with low diversity. As shown in Fig. 4a (middle), some modifications are concentrated on facial expressions without altering personal identity. As γ increases, the FID (↓) decreases, reaching a minimum at an optimal sharing degree. This optimum balances specific image feature capture and generalizable pattern learning of a

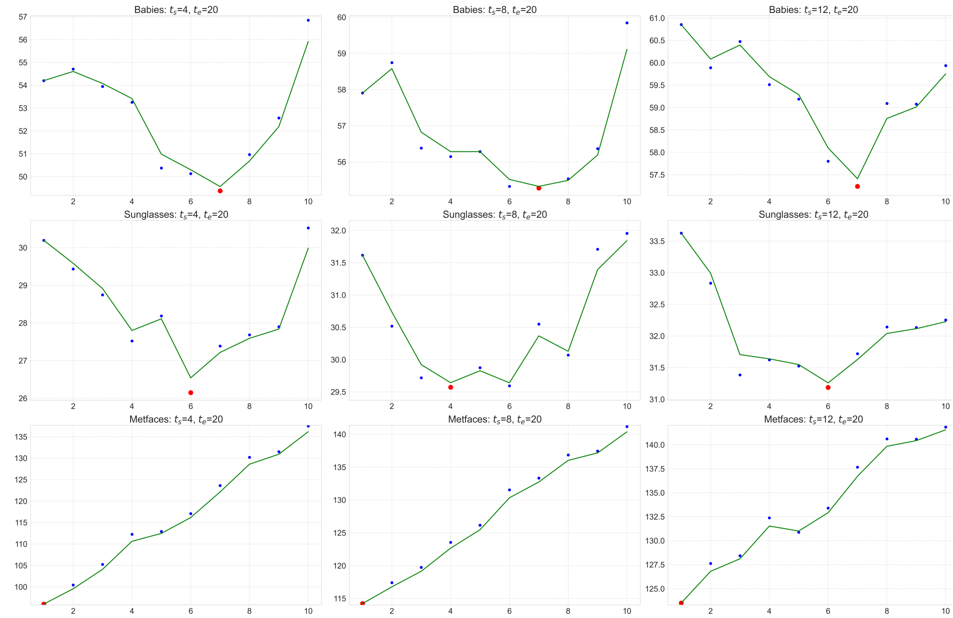

Figure 3: FID (↓) values across different IGM sharing degrees (γ) for three target domains in FSIG: Babies, Sunglasses, and MetFaces. Each subplot represents a different domain, where the x-axis denotes the IGM sharing degree γ, ranging from 1 (one IGM per one image) to 10 (one IGM per ten images), and the y-axis shows the corresponding FID score. An Exponential Moving Average curve (green line) illustrates the trend, and the lowest FID(↓) score is marked with a red dot.

target domain. Further increasing γ to 10 leads to FID (↓) increase, indicating underfitting. Generated images include source domain samples due to insufficient fitting capacity. IGMs fail to fully adapt the source model to the target domain, as the orange boxed samples in Fig. 4a (right). This phenomenon illustrates the trade-off between model capacity and generalization in IGM-guided DMs.

In contrast, the MetFaces target domain exhibits a distinct pattern. When $\gamma = 1$, generated samples closely match the target domain style but lack diversity. As the sharing degree increases to 10, the generated samples predominantly resemble the source domain, with only slight characteristics of the target domain (Fig. 4a second row, right). This behavior differs from Babies and Sunglasses, where intermediate sharing degrees yield optimal results. For MetFaces, the significant disparity from the source domain exposes the limitations of IGMs in bridging large domain gaps, resulting in effective target domain capture only at lower sharing degrees (We provide a detailed analysis in Sec. 4.3 and visualization in Section C.1). This finding highlights the importance of selecting an appropriate source domain that shares sufficient similarities with the target domain in FSIG tasks.

4.2 Main Results on FSIG

Building upon the insights from our analysis of IGM sharing degree, we now apply our method to real-world Few-Shot Image Generation (FSIG) experiments. In this section, we present a comparative evaluation of our approach against current state-of-the-art (SOTA) methods; the quantitative results are shown in Tab. 1. To demonstrate the robustness of our method, we further present the experimental results with sharing degrees of 10 (FS-DRL-10) and 5 (FS-DRL-5). These configurations utilize one-tenth and one-half of the parameters employed in the CRDI (Cao & Gong, 2024), respectively.

As seen in Tab. 1, FS-DRL significantly improves the performance of representation learning method in FSIG. However, in Babies and MetFaces, a gap remains compared to fine-tuning methods in terms of FID. Consistent with the findings of Cao & Gong (2024), we observe that while fine-tuning approaches achieve better performance on evaluation metrics, they tend to produce samples with certain visual artifacts. In contrast, representation learning methods generate "cleaner" samples, but with reduced diversity. See Fig. 4b for visual examples. However, FID score failed to capture these differences, as in Fig. 3 (first and second rows), FID scores at $\gamma = 1$ and 10 are comparable.

Figure 4: **a**: Impact of IGM sharing degree (FS-DRL-γ) on generated image quality and diversity, highlighting source domain leakage (orange) and low diversity (blue) (First row: Babies, Second row: MetFaces). **b**: Visual examples of our method alongside four other high-performance methods on Sunglasses (RL: Representation Learning and FT: Fine-Tuning). **c**: Mode coverage comparison across GenDA, RICK and our method. For each row, the leftmost image is from the MetFaces target domain, followed by the most similar (SSIM) generated images. Please zoom in for more details.

Table 1: Comparing FID (↓) Scores and MC-SSIM (↑) (for MetFaces only) between our methods and the baselines (Mean $\pm$ Std.). FS-DRL-γ represents our method with a sharing degree γ, and FS-DRL-opt denotes the optimized result. RL and FT represent Representation Learning and Fine-Tuning, respectively. Best in **bold** and the second best in **underline with bold**.

		Babies	Sunglasses	MetFaces	
Method	**Type**	FID ↓	FID↓	FID↓	MC-SSIM ↑*
GenDA	RL	63.31 $\pm$ 0.05	35.64 $\pm$ 0.15	104.48 $\pm$ 0.58	0.33 $\pm$ 0.03
RICK	FT	**39.39** $\pm$ 0.09	25.22 $\pm$ 0.11	**48.53** $\pm$ 0.34	0.41 $\pm$ 2e-3
CRDI	RL	48.52 $\pm$ 0.28	24.62 $\pm$ 0.18	94.86 $\pm$ 0.72	0.62 $\pm$ 5e-3
FS-DRL-10	RL	56.96 $\pm$ 0.31	31.69 $\pm$ 0.25	110.54 $\pm$ 0.50	0.57 $\pm$ 0.01
FS-DRL-5	RL	43.73 $\pm$ 0.29	**22.69** $\pm$ 0.16	88.36 $\pm$ 0.52	**0.64** $\pm$ 7e-3
FS-DRL-opt	RL	**41.95** $\pm$ 0.22	**21.93** $\pm$ 0.16	**77.17** $\pm$ 0.43	**0.70** $\pm$ 2e-3

*Calculated using 5000 samples for improved stability compared to prior work.

This indicates the limitation of FID score in distinguishing between source domain leakage and low diversity issues, as it measures both quality and diversity using feature space distances.

This limitation is particularly evident in complex domains like MetFaces (given samples in Fig.4a second row, left). While fine-tuning methods achieve lower FID (↓) scores, they capture only a limited subset of styles with prominent artifacts. Our approach, despite higher FID (↓) scores, achieve superior mode coverage and sample quality. To better quantify this aspect, we employ the MC-SSIM metric (Tab.1 last column), which shows that our method outperforms others in preserving target domain styles. Fig.4c provides qualitative results of this advantage. These findings underscore the importance of using complementary metrics for comprehensive model evaluation in FSIG tasks and highlight the strength of our approach in maintaining target domain styles.

Table 4: Comparison of model performance under 10-shot, 5-shot, and 1-shot with GenDA and CRDI, evaluated based on generation quality using the FID score (↓). Best in **Bold**.

	1-shot		5-shot		10-shot	
Methods	**Babies**	**Sunglasses**	**Babies**	**Sunglasses**	**Babies**	**Sunglasses**
GenDA	105.13	83.70	65.47	45.44	62.14	35.64
CRDI	100.85	74.60	55.87	31.35	48.52	24.62
Ours	**95.90**	**60.99**	**48.27**	**28.45**	**41.95**	**21.93**

4.3 Further Analysis and Discussion

Table 2: Comparisons of model performance with different ρ, evaluated by the FID (↓). Best in **Bold**.

ρ-th	0	20	40	60	80
Babies	45.70	44.56	42.40	**41.95**	43.53
Sunglasses	22.46	22.08	**21.93**	22.55	25.90
MetFaces	78.31	**77.17**	79.38	81.66	88.19

Effective of Percentile Gradient Clipping Tab. 2 demonstrates the impact of Percentile Gradient Clipping (PGC) across three target domains. The results show a U-shaped trend in FID scores, indicating the presence of noise in the IGM that can be effectively removed using PGC. However, excessive clipping eliminates informative gradients, degrading results. For Babies and Sunglasses, performance improves significantly with high percentile clipping (40th-60th), which suggests that IGM for these domains is inherently sparse. Conversely, MetFaces performs optimally at a lower percentile (20th), implying a denser IGM that requires more gradient information preservation; see the visualization and in-depth analysis in Section C.1. These divergent behaviors highlight IGM adaptability to domain complexity, motivating further exploration of domain-specific parameter optimization techniques.

Table 3: Comparisons of model performance and parameter count when further decrease number of parameter using Upsampling and LRMA, evaluated by the FID (↓). Best in **Bold**.

	Upsampling		LRMA		Original
# **Params**	m=64 $12K$	m=128 $49K$	r=64 $37K$	r=128 $82K$	n=256 $196K$
Babies	58.45	54.53	45.96	43.21	**41.95**
Sunglasses	33.16	30.62	40.21	39.89	**21.93**
MetFaces	100.46	88.21	133.42	131.49	**77.17**

Further Decrease Number of Parameter To explore the possibility of further reducing the number of parameters in our Invariant Gradient Matrix (IGM), we investigated two additional approaches: Upsampling and Low-Rank Matrix Approximation (LRMA). For Upsampling, we initialize a low-resolution gradient matrix $\mathbf{G}_{low}(t) \in \mathbb{R}^{m \times m}$, where $m < n$, with n being the dimensionality of the input samples. During the training and sampling process, we upsample $\mathbf{G}_{low}(t)$ to the original resolution using bilinear interpolation. For LRMA, we assume that $\mathbf{G}(t) = \mathbf{U}(t)\mathbf{\Sigma}(t)\mathbf{V}(t)^T$ is an anti-symmetric matrix, where $\mathbf{U}(t) \in \mathbb{R}^{n \times r}, \mathbf{\Sigma}(t) \in \mathbb{R}^{r \times r}, \mathbf{V}(t) \in \mathbb{R}^{n \times r}$, with $r < n$. The results are shown in Tab. 3. These results indicate that, while IGM exhibits some sparsity, simple parameter reduction methods may not effectively capture its full information content. LRMA shows more promise, particularly on certain datasets, but requires further refinement to achieve performance comparable to that of the original method across diverse datasets.

From Few-Shot to One-Shot To evaluate the performance of our method in more extreme scenarios, we designed experiments under 5-shot and 1-shot settings. In these cases, conventional models face an increased risk of overfitting. However, our approach, leveraging the adjustable sharing degree γ, demonstrates significant advantages. As shown in Tab. 4, our method significantly outperforms GenDA (Mondal et al., 2022) and CRDI Cao & Gong (2024) under both 5-shot and 1-shot scenarios, highlighting its effectiveness in extreme few-shot conditions.

5 Conclusion

We present a novel representation learning framework for Few-Shot Image Generation, featuring a tunable parameter to explicitly mitigate overfitting while adapting a specific domain. Our method achieves competitive SOTA performance while surpassing representation learning-based approaches using only half of the parameters. By focusing on the diffusion process, our approach is compatible with all diffusion models, offering a versatile and efficient solution for Few-Shot Image Generation.

REPRODUCIBILITY STATEMENT

To ensure that the proposed work is reproducible, we have included a pseudocode for training (Algo. 1) and sampling (Algo. 2). We have an explicit section (Sec. 4) with implementation details. We have also clearly mentioned evaluation details in Section .E. Complete code will be released upon acceptance.

REFERENCES

Milad Abdollahzadeh, Touba Malekzadeh, Christopher TH Teo, Keshigeyan Chandrasegaran, Guimeng Liu, and Ngai-Man Cheung. A survey on generative modeling with limited data, few shots, and zero shot. *arXiv preprint arXiv:2307.14397*, 2023.

Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

Sergey Bartunov and Dmitry Vetrov. Few-shot generative modelling with generative matching networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 670–678. PMLR, 2018.

Yu Cao and Shaogang Gong. Few-shot image generation by conditional relaxing diffusion inversion. *arXiv preprint arXiv:2407.07249*, 2024.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 35:25683–25696, 2022.

Louis Clouâtre and Marc Demers. Figr: Few-shot image generation with reptile. *arXiv preprint arXiv:1901.02199*, 2019.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Teng Hu, Jiangning Zhang, Liang Liu, Ran Yi, Siqi Kou, Haokun Zhu, Xu Chen, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Phasic content fusing diffusion model with directional distribution consistency for few-shot model adaption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2406–2415, 2023.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

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

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

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

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

Calvin Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. *arXiv preprint arXiv:2002.10964*, 2020.

Arnab Kumar Mondal, Piyush Tiwary, Parag Singla, and AP Prathosh. Few-shot cross-domain image generation via inference-time latent-code learning. In *The Eleventh International Conference on Learning Representations*, 2022.

Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2750–2758, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10743–10752, 2021.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10619–10629, 2022.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

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

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

H Risken. The fokker-planck equation, 1996.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *Proceedings of the IEEE international conference on computer vision*, pp. 2830–2839, 2017.

Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2c: Diffusion-decoding models for few-shot conditional generation. *Advances in Neural Information Processing Systems*, 34: 12533–12548, 2021.

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

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pp. 1135–1151, 1981.

Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22532–22541, 2023.

Yaxing Wang, Chenshen Wu, Luis Herranz, Joost Van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 218–234, 2018.

Jiayu Xiao, Liang Li, Chaofei Wang, Zheng-Jun Zha, and Qingming Huang. Few shot generative model adaption via relaxed spatial structural alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11204–11213, 2022.

Jordan Yaniv, Yael Newman, and Ariel Shamir. The face of art: landmark detection and geometric style in portraits. *ACM Transactions on graphics (TOG)*, 38(4):1–15, 2019.

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.

Zhongqi Yue, Jiankun Wang, Qianru Sun, Lei Ji, Eric I Chang, Hanwang Zhang, et al. Exploring diffusion time-steps for unsupervised representation learning. *arXiv preprint arXiv:2401.11430*, 2024.

Yunqing Zhao, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-Man Man Cheung. Few-shot image generation via adaptation-aware kernel modulation. *Advances in Neural Information Processing Systems*, 35:19427–19440, 2022.

Yunqing Zhao, Chao Du, Milad Abdollahzadeh, Tianyu Pang, Min Lin, Shuicheng Yan, and Ngai-Man Cheung. Exploring incompatible knowledge transfer in few-shot image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7380–7391, 2023.

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

A Appendix

This is the appendix for "Exploring Few-Shot Image Generation With Minimized Risk of Overfitting". Tab. 5 summarizes the abbreviations and symbols used in the paper.

This appendix is organized as follows:

- Section B discusses the limitation and broader impact of our work.
- Section C gives the full proof of our Theorem with additional explanation.
- Section D presents additional details of our approach.
- Section E presents additional details of the FSIG evaluation metric.
- Section F presents additional quantitative and qualitative results.

Table 5: List of abbreviations and symbols used in the paper

Abbreviation/Symbol	Meaning
Abbreviation	
Sec. A.B	Section in the main paper
Section. A.B	Section in the Appendix
FSIG	Few-Shot Image Generation
DM	Diffusion Model
DDPM	Denoising Diffusion Probabilistic Model
IGM	Invariant Gradient Matrix
LRMA	Low-Rank Matrix Approximation
Symbol in Theory	
$\mathcal{X}$	Source Domain
$\mathcal{Y}$	Target Domain
$G(t)$	Invariant Gradient Matrix
$g_{i,j}(t)$	(i,j)-th element of Invariant Gradient Matrix $G(t)$
$\mathcal{M}(\cdot)$	Domain adaptation measure
Symbol in Algorithm	
x_0	Original source domain sample
x_t	Noisy original source sample after t forward step
y_0	Target domain sample
y_t	Noisy target sample after t forward step
$q(\cdot)$	Distribution in the encoding process
$p_\theta(\cdot)$	Distribution in the θ-parameterized decoding process
ρ	Percentile of percentile gradient clipping
$\hat{p}_\theta$	Pretrained diffusion model with our IGM
θ	Parameter of U-Net
$\hat{x}_0$	Reconstructed source domain sample x_0
$\hat{y}_0$	Reconstructed target domain sample y_0
T	Total time-steps
$\beta_1, \ldots, \beta_T$	Variance schedule
α_t	$1 - \beta_t$
$\bar{\alpha}_t$	$\prod_{s=1}^{t} \alpha_s$

B Limitation and Broader Impact

Limitation Although our method effectively balances specificity and generalization, its performance degrades when the disparity between the source and target domains is substantial, such as MetFaces (Karras et al., 2020). In such cases, overfitting tends to outperform underfitting (Fig. 3). A potential solution involves incorporating Large Multi-modality Models (LMMs) like CLIP (Radford et al., 2021) to constrain style more effectively, allowing the Invariant Gradient Matrix to preserve more non-style information. We avoided using CLIP to minimize target domain exposure, as LMMs may have been trained on these samples. However, if this constraint can be relaxed, integrating

LMMs could enhance our method's robustness across diverse domains. Future work will explore this integration while maintaining data privacy.

Broader Impact Although our method outperforms state-of-the-art (SOTA) approaches in various comparisons, our research is not centered on topping leaderboards but rather on exploring the limits of FSIG while "fundamentally" avoiding overfitting. It is worth noting that while diffusion models have made impressive progress in recent years, surpassing GANs in most fields, they are rarely used in Few-Shot Image Generation (FSIG) tasks. This is primarily because most FSIG methods rely on fine-tuning, and diffusion models, despite being trained on the same datasets, have more parameters, making them seemingly "unsuitable" for FSIG tasks.

However, on the one hand, the training data for large models continues to expand rapidly and is becoming crucial in many real-world applications. On the other hand, although large models can generate highly realistic images, they still underperform on most user-defined real-world subjects. This gap requires FSIG methods that can align with the capabilities of these large models. Our method presents a novel attempt toward this goal, showing promising initial progress.

C Proof and Additional Theoretical Analysis

C.1 Additional Analysis of Equivalent Classifier

Consider the gradient of the log-conditional probability:

$$\nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x}_t) = \mathbf{c} \tag{12}$$

This differential equation can be solved to obtain the form of $p(\mathbf{y} \mid \mathbf{x}_t)$. Integrating both sides with respect to $\mathbf{x}$:

$$\int \nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x}_t) \cdot d\mathbf{x} = \int \mathbf{c} \cdot d\mathbf{x} \tag{13}$$

yield

$$\log p(\mathbf{y} \mid \mathbf{x}_t) = \mathbf{c} \cdot \mathbf{x}_t + K \tag{14}$$

where K is an integration constant. Exponentiating both sides:

$$p(\mathbf{y} \mid \mathbf{x}_t) = \exp(\mathbf{c} \cdot \mathbf{x}_t + K) = \exp(K) \cdot \exp(\mathbf{c} \cdot \mathbf{x}_t) \tag{15}$$

Let $Z = \exp(K)$, which serves as a normalization constant. Thus:

$$p(\mathbf{y} \mid \mathbf{x}_t) = Z \cdot \exp(\mathbf{c} \cdot \mathbf{x}_t) \tag{16}$$

This exponential form aligns with the softmax mechanism, where $\mathbf{c}$ acts as an attention matrix, determining the "attention" or "importance" of different regions in the state space given $\mathbf{y}$.

Invariant Gradient Matrix Visualization To validate our theoretical analysis, we visualized the Invariant Gradient Matrices (IGMs) at different diffusion timesteps for three target domains: Babies, Sunglasses, and MetFaces (Fig. C.1). Notably, for Babies and Sunglasses domains, the IGMs exhibit significant sparsity, aligning with our analysis in Sec.4.3. In contrast, the IGM for MetFaces contains more intricate details, likely capturing additional information such as style variations. This increased complexity in the MetFaces IGM correlates with the observed reduction in diversity, as the model focuses on preserving more domain-specific features.

C.2 Proof of Theorem 1 and Remark 1

Let x be a random variable following a normal distribution, $\mathcal{N}(\mu, \sigma)$, i.e.,

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{17}$$

Assume that the conditional probability $p(y|x)$ has the form:

$$p(y|x) = \exp(cx) \cdot \text{const} \tag{18}$$

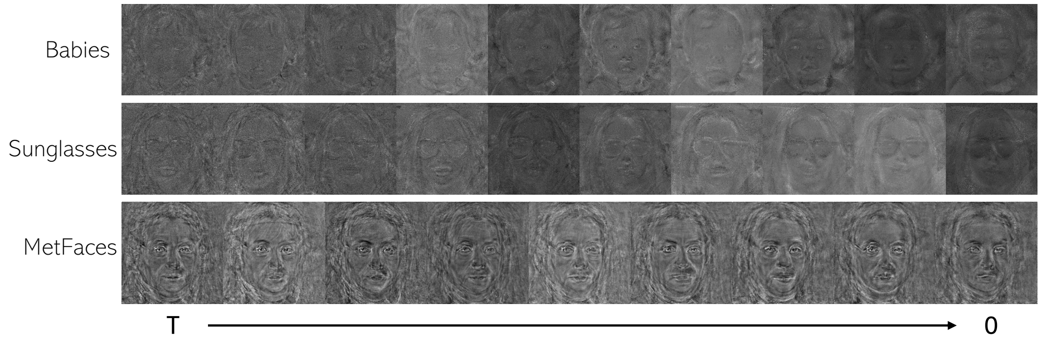

Figure 5: Visualization of Invariant Gradient Matrices (IGMs) across three target domains: Babies, Sunglasses, and MetFaces. Each row represents the IGM at different diffusion timesteps for the corresponding domain.

where c is an invariant variable (IGM in our case). Applying Bayes' theorem, we obtain (the constant term from $p(y|x)$ is absorbed into $p(y)$):

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{\exp(cx)}{p(y)\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{19}$$

Combining the exponential terms, we have:

$$p(x|y) = \frac{1}{p(y)\sqrt{2\pi\sigma^2}} \exp\left(cx - \frac{(x-\mu)^2}{2\sigma^2}\right) \tag{20}$$

By completing the square, we can rewrite the expression as:

$$p(x|y) = \frac{1}{p(y)\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-(\mu+c\sigma^2))^2}{2\sigma^2} + \frac{c^2\sigma^2}{2} + c\mu\right) \tag{21}$$

This expression shows that $p(x|y)$ is also a normal distribution with mean $\mu + c\sigma^2$ and variance σ^2, where the normalization constant is given by:

$$\frac{1}{p(y)} \exp\left(\frac{c^2\sigma^2}{2} + c\mu\right) \tag{22}$$

For a d-dimensional case where all dimensions are independent, we can treat each dimension separately and combine the results. The mean of each dimension will be updated as $\mu_i + c_i\sigma_i^2$, where i is the dimension index. The variances remain unchanged. The overall normalization constant will be the product of the normalization constants for each dimension. $\square$

C.3 Theoretical Analysis of Probability Flow Correction

Consider a diffusion model with probability density function (PDF) $p(x,t)$ for its data distribution, where x represents the data and t represents the time step of the diffusion process. The probability flow vector field $v(x,t)$ satisfies the modified Fokker-Planck equation with a diffusion coefficient $g(t)$:

$$\frac{\partial p(x,t)}{\partial t} = -\nabla_x \cdot (p(x,t)v(x,t)) + \frac{1}{2}\nabla_x \cdot \left(g(t)^2 \nabla_x p(x,t)\right). \tag{23}$$

The first term $-\nabla_x \cdot (p(x,t)v(x,t))$ represents the drift induced by the vector field $v(x,t)$, while the second term $\frac{1}{2}\nabla_x \cdot \left(g(t)^2 \nabla_x p(x,t)\right)$ accounts for diffusion, with $g(t)$ ($\sqrt{\beta_t}$ in DDPM) as the time-dependent diffusion coefficient.

To improve the alignment of the model's probability flow with the target domain, we introduce a correction term $\delta v(x,t)$:

$$\hat{v}(x,t) = v(x,t) + \delta v(x,t), \tag{24}$$

where $\delta v(x,t)$ is learned from an underfitted classifier at the intermediate state t. This correction term can be represented as:

$$\delta v(x,t) = \mathbb{E}_{\theta_c \sim p(\theta_c|x)}[f(x,\theta_c)], \tag{25}$$

where θ_c represents the classifier parameters, $p(\theta_c|x)$ is the posterior distribution given the data x, and $f(x, \theta_c)$ maps these parameters to a correction in the probability flow. This correction aims to capture the discrepancy between the current model state and the target domain.

By introducing the correction, the modified vector field $\hat{v}(x, t)$ adjusts the dynamics of the diffusion process, resulting in the corrected Fokker-Planck equation:

$$\frac{\partial p(x,t)}{\partial t} = -\nabla_x \cdot (p(x,t)\hat{v}(x,t)) + \frac{1}{2}\nabla_x \cdot \left(g(t)^2 \nabla_x p(x,t)\right). \tag{26}$$

Proof (Informal) To analyze the effect of the correction $\delta v(x, t)$, we expand the divergence term in the corrected Fokker-Planck equation:

$$\begin{aligned}
\frac{\partial p(x,t)}{\partial t} &= -\nabla_x \cdot (p(x,t)\hat{v}(x,t)) + \frac{1}{2}\nabla_x \cdot \left(g(t)^2 \nabla_x p(x,t)\right) \\
&= -\nabla_x \cdot (p(x,t)(v(x,t) + \delta v(x,t))) + \frac{1}{2}\nabla_x \cdot \left(g(t)^2 \nabla_x p(x,t)\right) \\
&= -\nabla_x \cdot (p(x,t)v(x,t)) - \nabla_x \cdot (p(x,t)\delta v(x,t)) + \frac{1}{2}\nabla_x \cdot \left(g(t)^2 \nabla_x p(x,t)\right).
\end{aligned} \tag{27}$$

The term $-\nabla_x \cdot (p(x,t)v(x,t)) + \frac{1}{2}\nabla_x \cdot \left(g(t)^2 \nabla_x p(x,t)\right)$ corresponds to the original diffusion model, while the new term $-\nabla_x \cdot (p(x,t)\delta v(x,t))$ introduces a correction based on the classifier. This correction guides the probability flow to better match the target distribution. □

In summary, by modifying the probability flow vector field to $\hat{v}(x, t)$, we adjust the generative process to produce samples that more closely align with the target data distribution, enhancing both the quality and diversity of the generated samples.

C.4 Theoretical analysis of Percentile Gradient Clipping

Given a gradient matrix $\mathbf{G}(t)$ containing gradient information between the source domain $\mathcal{X}$ and the target domain $\mathcal{Y}$, let $Q(\mathbf{G}(t), \rho)$ denote the ρ-th percentile of $\mathbf{G}(t)$. Define the gradient clipping operation $\mathcal{T}$ as follows:

$$\mathcal{T}(\mathbf{G}(t))_{i,j} = \begin{cases} 0, & \text{if } |g_{i,j}(t)| < Q(\mathbf{G}(t), p) \\ g_{i,j}(t), & \text{otherwise} \end{cases} \tag{28}$$

where $g_{i,j}(t)$ denotes the (i, j)-th element of $\mathbf{G}(t)$. Then, the gradient clipping operation $\mathcal{T}$ satisfies the following inequality:

$$\frac{I(\mathcal{T}(\mathbf{G}(t)); \mathcal{Y})}{H(\mathcal{T}(\mathbf{G}(t)))} \geq \frac{I(\mathbf{G}(t); \mathcal{Y})}{H(\mathbf{G}(t))} \tag{29}$$

where $I(\cdot;\cdot)$ denotes the mutual information and $H(\cdot)$ denotes the entropy. In other words, the gradient clipping operation $\mathcal{T}$ increases the ratio of effective information, enabling the clipped gradient matrix $\mathcal{T}(\mathbf{G}(t))$ to capture the characteristics of the target domain $\mathcal{Y}$ more effectively.

Proof (Informal) The gradient clipping operation $\mathcal{T}$ sets the elements of $\mathbf{G}(t)$ with smaller magnitudes to zero. This is equivalent to removing the gradient information that has a relatively weak influence on the target domain $\mathcal{Y}$. Since elements with smaller magnitudes are assumed to contribute less to mutual information $I(\mathbf{G}(t); \mathcal{Y})$, their removal has a limited impact on the overall mutual information between the gradient matrix and the target domain. At the same time, removing this information reduces the entropy $H(\mathbf{G}(t))$ of $\mathbf{G}(t)$, since it reduces the overall noise and randomness in the gradient matrix.

Specifically, let $g_{i,j}(t)$ denote the (i, j)-th element of $\mathbf{G}(t)$. The clipping threshold $Q(\mathbf{G}(t), \rho)$ is selected such that elements below this threshold contribute minimally to the mutual information $I(\mathbf{G}(t); \mathcal{Y})$. Hence, we have:

$$I(\mathcal{T}(\mathbf{G}(t)); \mathcal{Y}) \approx I(\mathbf{G}(t); \mathcal{Y})$$

At the same time, setting these elements to zero reduces the entropy $H(\mathbf{G}(t))$, as the sparsity of $\mathcal{T}(\mathbf{G}(t))$ increases and the overall uncertainty within the gradient matrix is reduced. This reduction in entropy is significant, since the clipped elements are removed entirely, resulting in:

$$H(\mathcal{T}(\mathbf{G}(t))) < H(\mathbf{G}(t))$$

Therefore, the ratio of mutual information to entropy increases after clipping:

$$\frac{I(\mathcal{T}(\mathbf{G}(t));\mathcal{Y})}{H(\mathcal{T}(\mathbf{G}(t)))} > \frac{I(\mathbf{G}(t);\mathcal{Y})}{H(\mathbf{G}(t))}$$

In essence, the gradient clipping operation $\mathcal{T}$ preserves the information that is relevant to the target domain $\mathcal{Y}$ while reducing the entropy of the gradient matrix. This increases the relative effectiveness of the retained information, allowing $\mathcal{T}(\mathbf{G}(t))$ to more effectively capture the characteristics of the target domain. □

D Additional Detail for Approach

Training Algorithm Algo. 2 shows the training pseudocode when $\gamma = 10$. When $\gamma < 10$, we randomly create a mapping function to distribute the images such that each IGM may correspond to multiple images. Specifically, as γ decreases, we aim to evenly distribute the images among the available IGMs. When γ eventually reduces to one, it results in a single IGM corresponding to all images. This mapping approach ensures that the images are distributed fairly and shared as evenly as possible across varying γ.

Algorithm 1 FS-DRL - Training Pseudo-code

```
Input: Target Domain 𝒴 = {y^0, y^1, ..., y^{n-1}} (n=10), start point t_s, end point t_e, Randomly
  Initialized IGM G_θ(t), a Frozen Noise Network (DM) ε_θ and Learning Rate ν.
while not converge do
    Sample: t uniformly from [t_s, ..., t_e]
    for i, y^i in enumerate(𝒴) do
        Given y^i_{t-1} ← sample from sqrt(ᾱ_{t-1}) y^i_0 + sqrt(1 - ᾱ_{t-1}) ε,   ε ~ 𝒩(0, I)
        ε̂ ← ε_θ(y^i_t) - sqrt(1 - ᾱ_t) G_θ(t, i)
        ŷ^i_0 ← (y^i_t - sqrt(1 - ᾱ_t) ε̂) / sqrt(ᾱ_t)
        G_θ(t, i) ← G_θ(t, i) - ν ∇_{G_θ(t,i)} ℒ|y^i_0 - ŷ^i_0|
return G_θ
```

Sampling Algorithm We show the sampling pseudocode in Algo. 2.

Algorithm 2 FS-DRL - Sampling Code

```
Input: Target Domain 𝒴 = {y^0, y^1, ..., y^{n-1}} (n=10), start point t_s, end point t_e, Proposed
  IGM G_θ(t), a Frozen Noise Network (DM) ε_θ and a mask (Percentile Gradient Clipping).
Sample y_0 randomly from 𝒴, i from [0, ..., n-1], Set t ← t_e
y_t ← sample from sqrt(ᾱ_t) y_0 + sqrt(1 - ᾱ_t) ε,   ε ~ 𝒩(0, I)
for t in reversed(range(t_e + 1)) do
    if t < t_s then
        G_θ(t, i) ← 0
    ε̂ ← ε_θ(y_t) - sqrt(1 - ᾱ_t)(G_θ(t, i) ⊙ mask)
    ŷ_0 ← (y_t - sqrt(1 - ᾱ_t) ε̂) / sqrt(ᾱ_t)
    y_{t-1} ← sample from sqrt(ᾱ_i) ŷ_0 + sqrt(1 - ᾱ_i) ε,   ε ~ 𝒩(0, I)
return y_0
```

E Additional Detail for Evaluation

Implemented Intra-LPIPS Algorithm As most implementations of Intra-LPIPS skip empty clusters when calculating the average, reducing the number of comparisons (e.g., from 10 to only 3), misrepresenting true diversity, we modify the implementation as Algo. 3 (modified parts in red).

Implemented MC-SSIM Algorithm For pseudocode of MC-SSIM please refer to Algo. 4. Note that in Tab. 1, MC-SSIM was calculated using 5000 samples for improved stability, which may lead to disparities with prior work.

Algorithm 3 Calculate Intra-LPIPS within clusters

Input:
1. Generated images $X = x_1, \dots, x_b$
2. Real image dataloader L
3. Number of images per cluster m

Output: Average Intra-LPIPS within clusters

Step 1. Initialize empty clusters $C_i = \emptyset$ for $i \in 0, \dots, 9$
for $i = 1, \dots, b$ **do**
  Initialize distances $D = []$
  for real image r in L **do**
    $d = \texttt{LPIPS}(x_i, r)$ ▷ Compute LPIPS distance
    D.append(d)
  $j = \arg\min_j D$ ▷ Index of closest cluster
  C_j.append(i) ▷ Assign x_i to cluster C_j

Step 2. Restrict clusters to size m
for $i = 0, \dots, 9$ **do**
  $C_i = C_i[1 : m]$

Step 3. Compute pairwise Intra-LPIPS within each cluster
Initialize distances $D = []$
for $i = 0, \dots, 9$ **do**
  Initialize temp distances $T = [0]$ ▷ Initialize T with $[0]$ instead of an empty list.
  for $j = 1, \dots, |C_i|$ **do**
    for $k = j + 1, \dots, |C_i|$ **do**
      $d = \texttt{LPIPS}(x_{C_i[j]}, x_{C_i[k]})$ ▷ Pairwise LPIPS
      T.append(d)
  D.append(mean(T)) ▷ Average pairwise distance per cluster
return mean(D)

Algorithm 4 Calculate MC-SSIM

Input:
1. Target Domain Y
2. Synthesis images I
3. Number of top scores k

Output: Average Top-K SSIM for each reference image
Initialize dictionary $D = []$ ▷ To store average SSIM per reference
for reference image x in Y **do**
  Initialize list $S = []$ ▷ To store SSIM scores
  for image i in I **do**
    $score = \texttt{SSIM}(x, i)$ ▷ Compute SSIM
    S.append($score$)
  Sort S in descending order
  $T = S[1 : k]$ ▷ Top-K scores
  if T is not empty **then**
    $avg = \text{mean}(T)$
  else
    $avg = 0$
  $D[x] = avg$ ▷ Store average SSIM for x
return mean(D)

F Additional Experiment Results

F.1 Additional Quantitative Evaluations

Extended Results To extend the results presented in Tab. 1, a more comprehensive comparison with additional methods, including TGAN Wang et al. (2018), TGAN+ADA (Karras et al., 2020), BSA Noguchi & Harada (2019), FreezeD Mo et al. (2020), EWC Li et al. (2020), CDC Ojha et al. (2021), RSSA Xiao et al. (2022), DDPM-PA Zhu et al. (2022) AdAM Zhao et al. (2022), is shown in Tab. 7.

Diversity Quantitative Analysis Tab. 6 presents Intra-LPIPS results. While our method not always achieve the highest scores, it is crucial to note that Intra-LPIPS has limitations in assessing true diversity. Visual artifacts can inflate this metric, potentially rewarding methods that produce diverse but low-quality outputs. Our approach prioritizes balancing diversity with fidelity to the target domain, which may not be fully captured by Intra-LPIPS alone. For a more comprehensive evaluation of generation quality, qualitative results provide additional insight (Babies: Fig. 7 and MetFaces: Fig. 8, RICK generated samples come from CRDI (Cao & Gong, 2024)).

Table 6: Comparisons Intra-LPIPS (↑) Scores between our methods and the baseline methods. Best in **bold** and the second best in **underline with bold**.

Domains	FreezeD	RSSA	RICK	GenDA	CRDI	Ours
Babies	0.51	0.50	**0.60**	0.48	0.52	**<u>0.53</u>**
MetFaces	0.21	0.15	0.37	0.35	**0.41**	**0.41**

Table 7: (Extended Tab. 1) FID (↓) Scores for more baseline methods. FT represents Fine-Tuning.

Method	Type	Babies	Sunglasses	MetFaces
TGAN	FT	104.79	55.61	76.81
TGAN+ADA	FT	101.58	53.64	75.82
BSA	FT	140.34	76.12	–
FreezeD	FT	110.92	51.29	73.33
EWC	FT	87.41	59.73	62.67
CDC	FT	74.39	42.13	65.45
RSSA	FT	75.67	44.35	72.63
DDPM-PA	FT	48.92	34.75	–
AdAM	FT	48.83	28.03	51.34

F.2 More Domain Adaptation

To validate the performance of our method beyond the face-related domains, we performed experiments on various visual categories, including FFHQ to Otto (Yaniv et al., 2019), Church (Yu et al., 2015) to Haunted House Ojha et al. (2021), and Church to Van Gogh's house Ojha et al. (2021) adaptations. Qualitative results in Fig. 6 demonstrate consistent performance across these varied domain pairs.

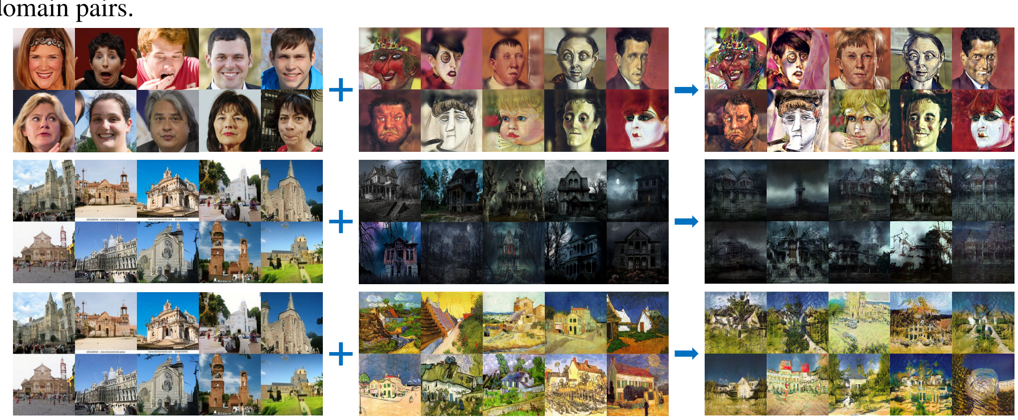

Figure 6: Adapting FFHQ → Otto (first row), Church → Haunted House (second row) and Church → Van Gogh's house (third row). First column: source domain, second column: target domain, third column: generated samples

Figure 7: Qualitative comparison with RICK (state-of-the-art) on Target Domain Babies.

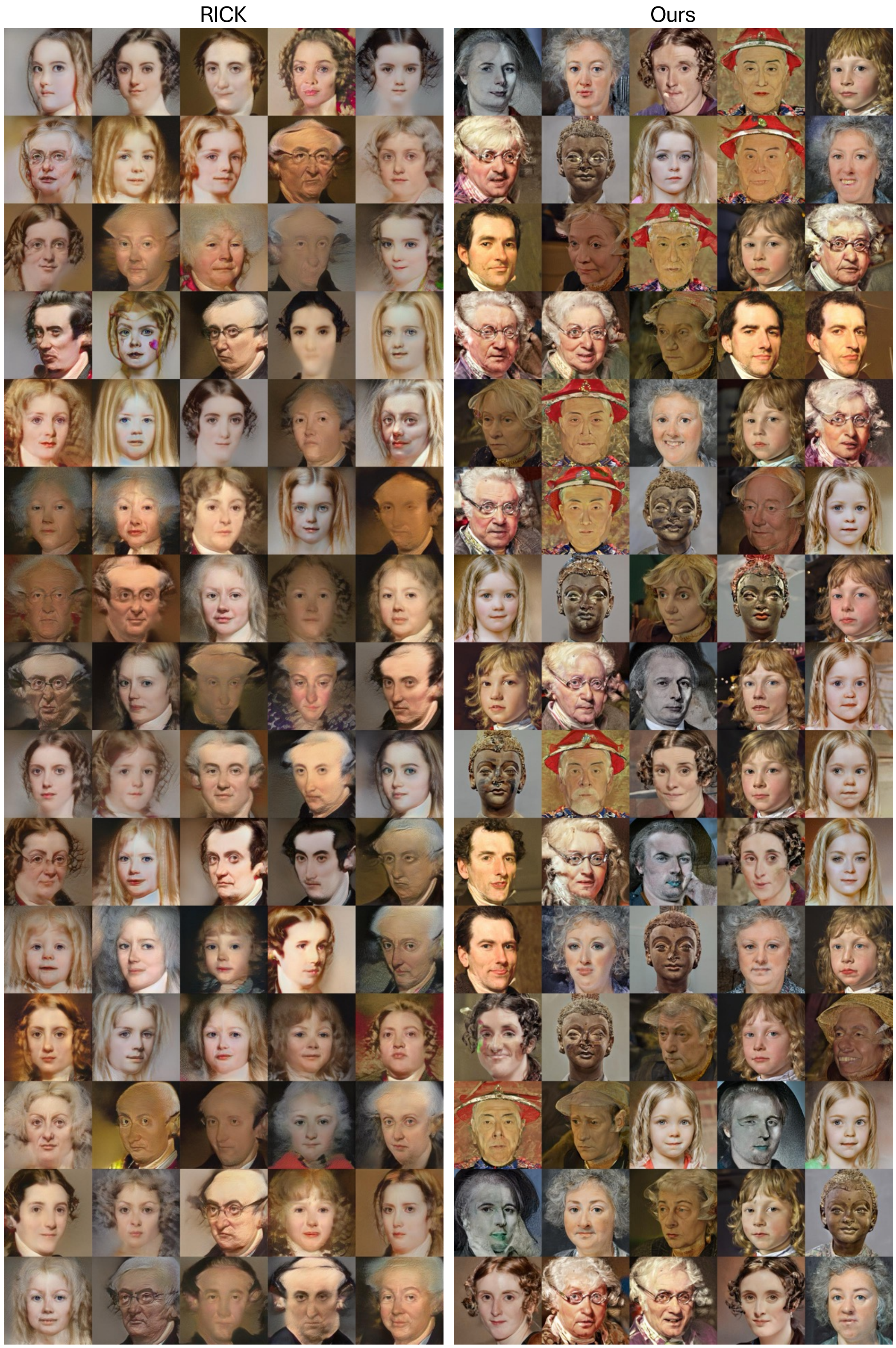

Figure 8: Qualitative comparison with RICK (state-of-the-art) on Target Domain MetFaces.