A. More Implementation Details

A.1. Training and Sampling Details

We present the training and sampling details of our SADM on different datasets in Tab. 6 for better reproducing our method.

Table 6. Training and sampling configurations in SADM.

	CIFAR-10		CelebA/FFHQ		ImageNet	
	Latent	Image	Latent	Image	Image	Latent
Training of SADM						
Based Diffusion Model	LSGM	EDM	LSGM	EDM	ADM	DiT
Sample Relation Measurement $\mathcal{R}$	cosine similarity	cosine similarity	cosine similarity	cosine similarity	cosine similarity	cosine similarity
Structural Distance Metric $\mathcal{D}$	L_2 distance	L_2 distance	L_2 distance	L_2 distance	L_2 distance	L_2 distance
Encoder Ψ_ϕ of Structure Discriminator	Inception V3	Inception V3	Inception V3	Inception V3	Inception V3	Inception V3
Round of Adversarial Training	2	2	3	3	4	4
Sampling of SADM						
SDE	LVP	WVE	LVP	WVE	LVP	LVP
Solver	PFODE	PFODE	PFODE	PFODE	DDPM	DDPM
Solver accuracy of $\mathbf{s}_{\boldsymbol{\theta}}$	1st-order	2nd-order	1st-order	2nd-order	1st-order	1st-order
Solver type of $\mathbf{s}_{\boldsymbol{\theta}}$	RK45	Heun	RK45	Heun	Euler (DDPM)	Euler (DDPM)
NFE	138	35	131	71	250	250
Classifier Guidance	✗	✗	✗	✗	✓	✓
w_t^{CG}	0	0	0	0	Adaptive	Adaptive

A.2. Datasets

Food101 [1]. This dataset contains 101 food categories, totaling 101,000 images. Each category includes 750 training images and 250 manually reviewed test images. The training images were kept intentionally uncleaned, preserving some degree of noise, primarily vivid colors and occasionally incorrect labels. All images have been adjusted to a maximum side length of 512 pixels.

SUN 397 [68]. The SUN benchmark database comprises 108,753 images labeled into 397 distinct categories. The quantities of images vary among the categories, however, each category is represented by a minimum of 100 images. These images are commonly used in scene understanding applications.

DF20M [48]. DF20 is a new fine-grained dataset and benchmark featuring highly accurate class labels based on the taxonomy of observations submitted to the Danish Fungal Atlas. The dataset has a well-defined class hierarchy and a rich observational metadata. It is characterized by a highly imbalanced long-tailed class distribution and a negligible error rate. Importantly, DF20 has no intersection with ImageNet, ensuring unbiased comparison of models fine-tuned from ImageNet checkpoints.

Caltech 101 [16]. The Caltech 101 dataset comprises photos of objects within 101 distinct categories, with roughly 40 to 800 images allocated to each category. The majority of the categories have around 50 images. Each image is approximately 300×200 pixels in size.

CUB-200-2011 [65]. CUB-200-2011 (Caltech-UCSD Birds-200-2011) is an expansion of the CUB-200 dataset by approximately doubling the number of images per category and adding new annotations for part locations. The dataset consists of 11,788 images divided into 200 categories.

ArtBench-10 [40]. ArtBench-10 is a class-balanced, standardized dataset comprising 60,000 high-quality images of artwork annotated with clean and precise labels. It offers several advantages over previous artwork datasets including balanced class distribution, high-quality images, and standardized data collection and pre-processing procedures. It contains 5,000 training images and 1,000 testing images per style.

Oxford Flowers [45]. The Oxford 102 Flowers Dataset contains high quality images of 102 commonly occurring flower categories in the United Kingdom. The number of images per category range between 40 and 258. This extensive dataset provides an excellent resource for various computer vision applications, especially those focused on flower recognition and classification.

Stanford Cars [34]. In the Stanford Cars dataset, there are 16,185 images that display 196 distinct classes of cars. These images are divided into a training and a testing set: 8,144 images for training and 8,041 images for testing. The distribution of samples among classes is almost balanced. Each class represents a specific make, model, and year combination, e.g., the 2012 Tesla Model S or the 2012 BMW M3 coupe.

B. Ablation Study

In the main text, we have conducted ablation study on our structural guidance and structure discriminator, and find both of them have a critical impact on the final model performance. In this section, we conduct more detailed ablation study on the designs in structure discriminator for better understanding of our model.

B.1. Encoder of Structure Discriminator

We here conduct ablation study on the encoder choice in our structure discriminator, and we compare with ResNet-18 and Transformer (ViT) architectures that are pre-trained on ImageNet in Fig. 7. In the ablation study, we evaluate the FID performance in three datasets with different encoders. From the results, we can find that Inception and ViT are both better than ResNet-18 because they are superior in capturing the visual semantics of images [7, 38, 39, 66], thus extracting more informative manifold structures. Overall, the encoder choice does not have an obvious impact on the model performance.

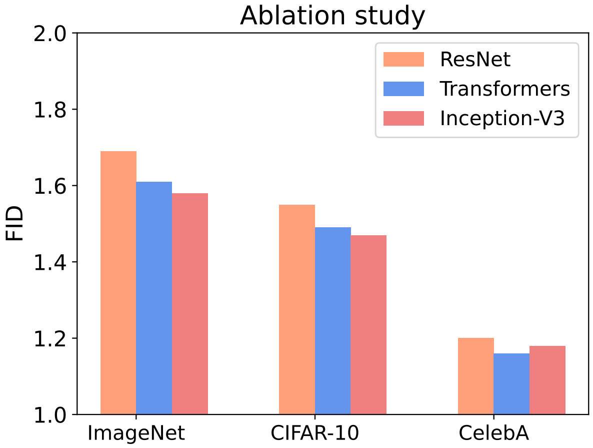


Figure 7. Ablation study on the encoder of structure discriminator in ImageNet, CIFAR-10, and CelebA datasets.

B.2. Metric of Structure Discriminator

In main text, we use cosine similarity for $\mathcal{R}$ and L_2 distance for $\mathcal{D}$. Here we conduct ablation study on the choice of these metrics, and put the results in Tab. 7. In the ablation study, we fix the $\mathcal{R}$ or $\mathcal{D}$ and change the other metric. We find that using cosine similarity and L_2 distance can achieve a similar result, and L_1 distance is slightly worse than other metrics. Overall, our model is robust to the choice of metrics.

B.3. Round of Adversarial Training

We further conduct ablation study on the rounds of our structure-guided adversarial training in Fig. 8. We find that in the initial round, the model performance can be significantly enhanced regarding FID score, demonstrating the effectiveness of our

Table 7. Ablation study on $\mathcal{R}$ and $\mathcal{D}$ in ImageNet 256×256.

Module \ Metric	L_1 distance	L_2 distance	cosine similarity
Sample Relation $\mathcal{R}$	1.65	**1.56**	1.58
Structural Distance $\mathcal{D}$	1.63	**1.58**	1.60

structure discriminator. After few rounds, the model performance tends to converge as the diffusion denoiser and structure discriminator in SADM have achieved a balance.

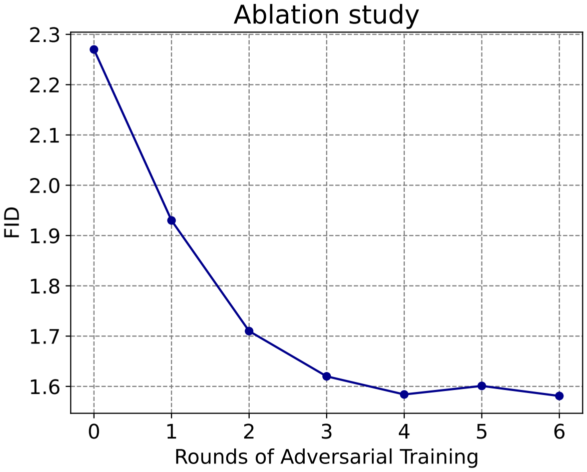
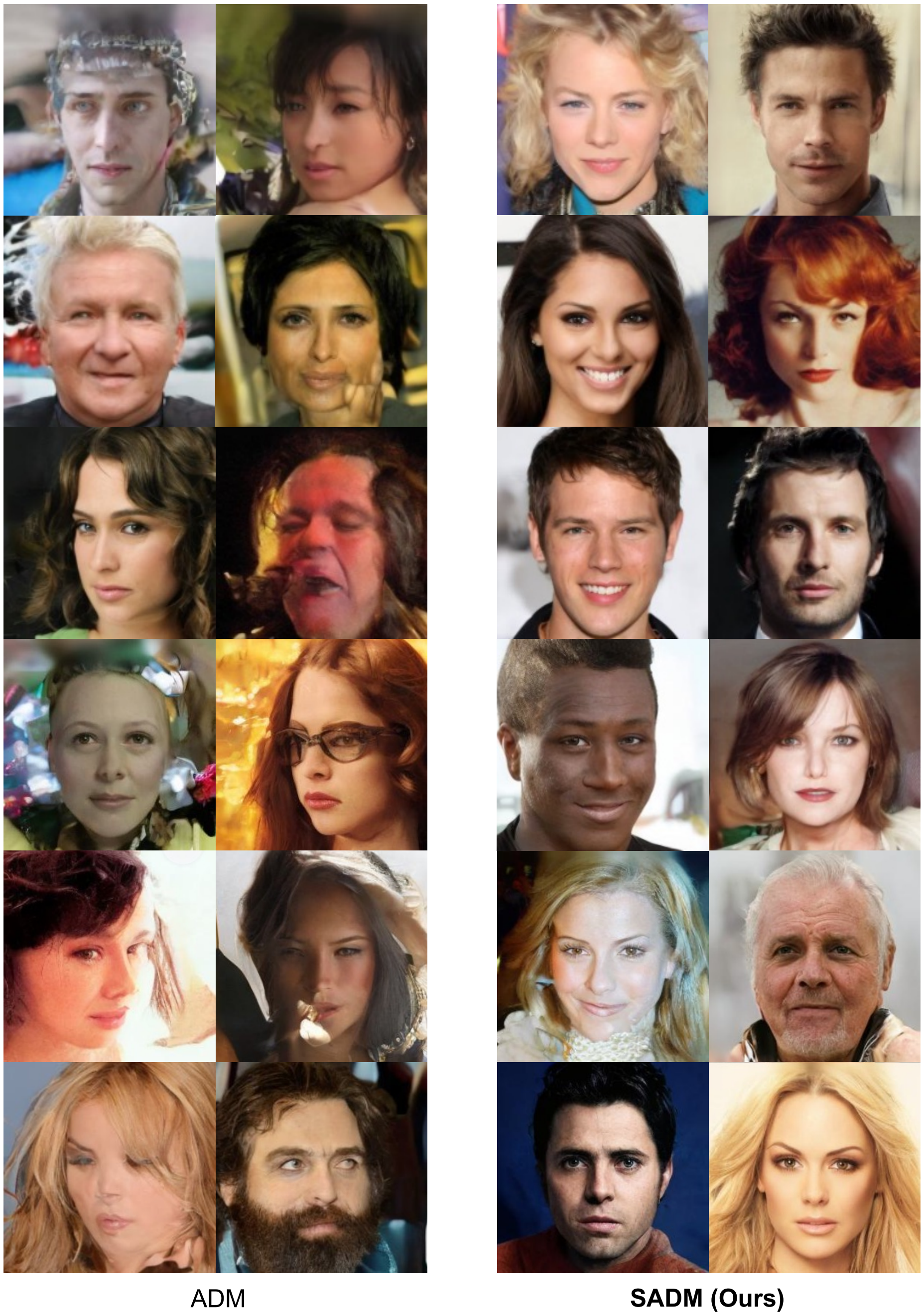


Figure 8. Ablation study on the round of our structure-guided adversarial training in ImageNet.

C. More Qualitative Comparisons

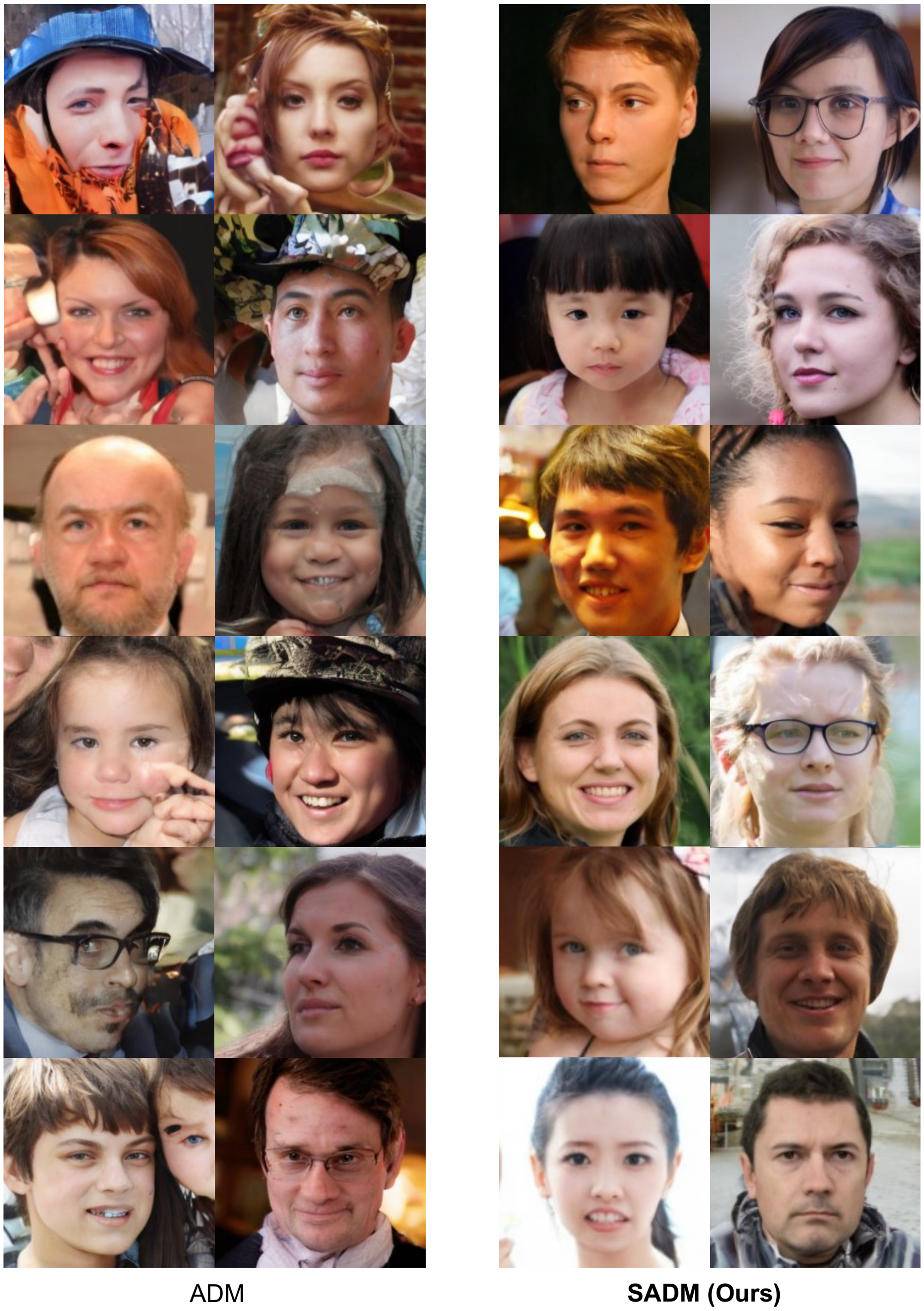
We here show more qualitative comparison results between our SADM and ADM [12]. Fig. 9 and Fig. 10 show the generated samples on CelebA and FFHQ datasets in unconditional image generation task, and Fig. 11 and Fig. 12 show the generated samples on CUB-200 and Oxford-Flowers datasets in cross-domain fine-tuning task. We observe that our SADM can comprehensively achieve improvements over previous diffusion models in fidelity and quality, demonstrating the superiority of our new training algorithm.



ADM

SADM (Ours)

Figure 9. Random generated samples of ADM [12] and our SADM on unconditional CelebA.



ADM

SADM (Ours)

Figure 10. Random generated samples of ADM [12] and our SADM on unconditional FFHQ.

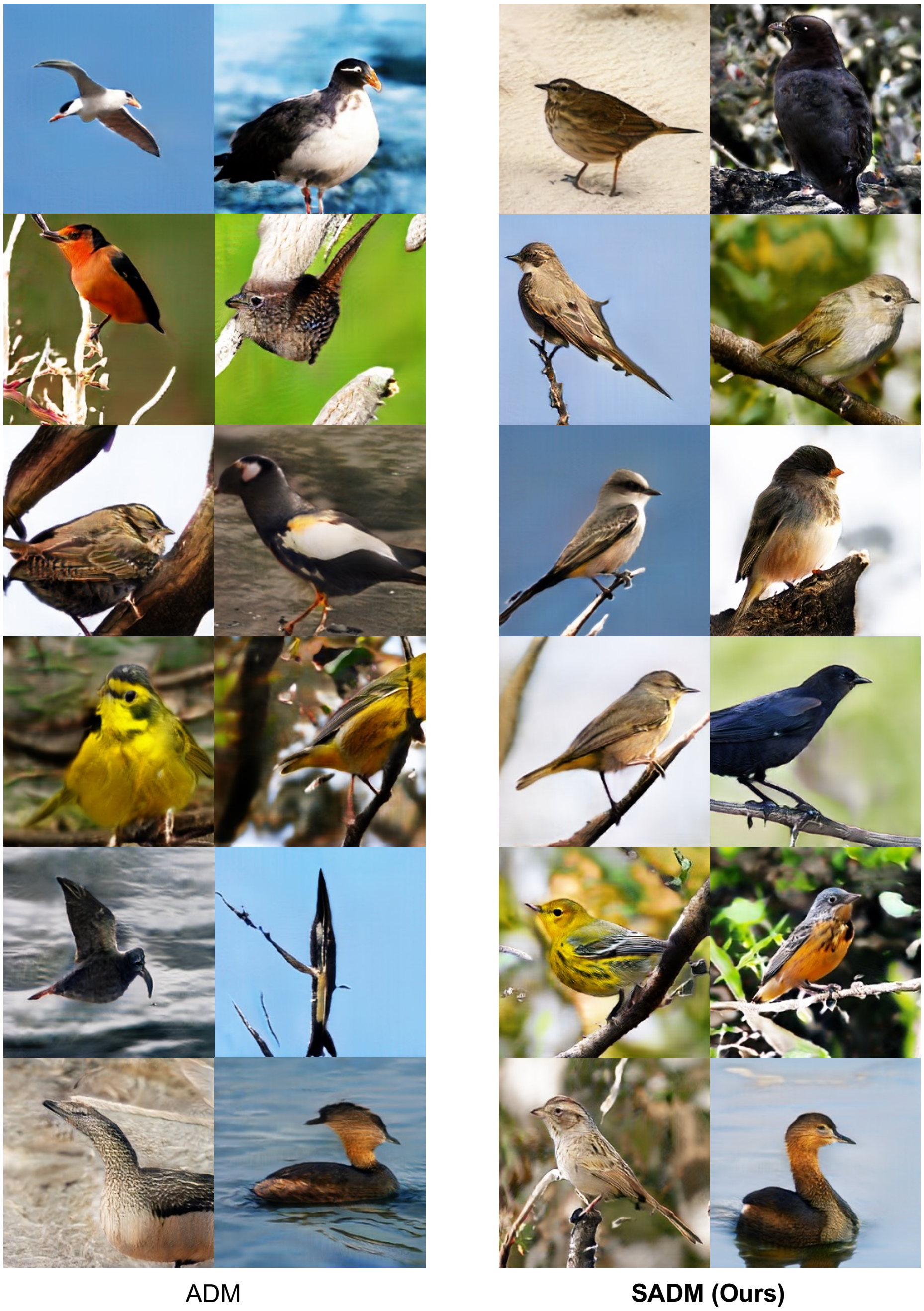


Figure 11. Random generated samples of the diffusion model fine-tuned by ADM [12] and our SADM on unconditional CUB-200.

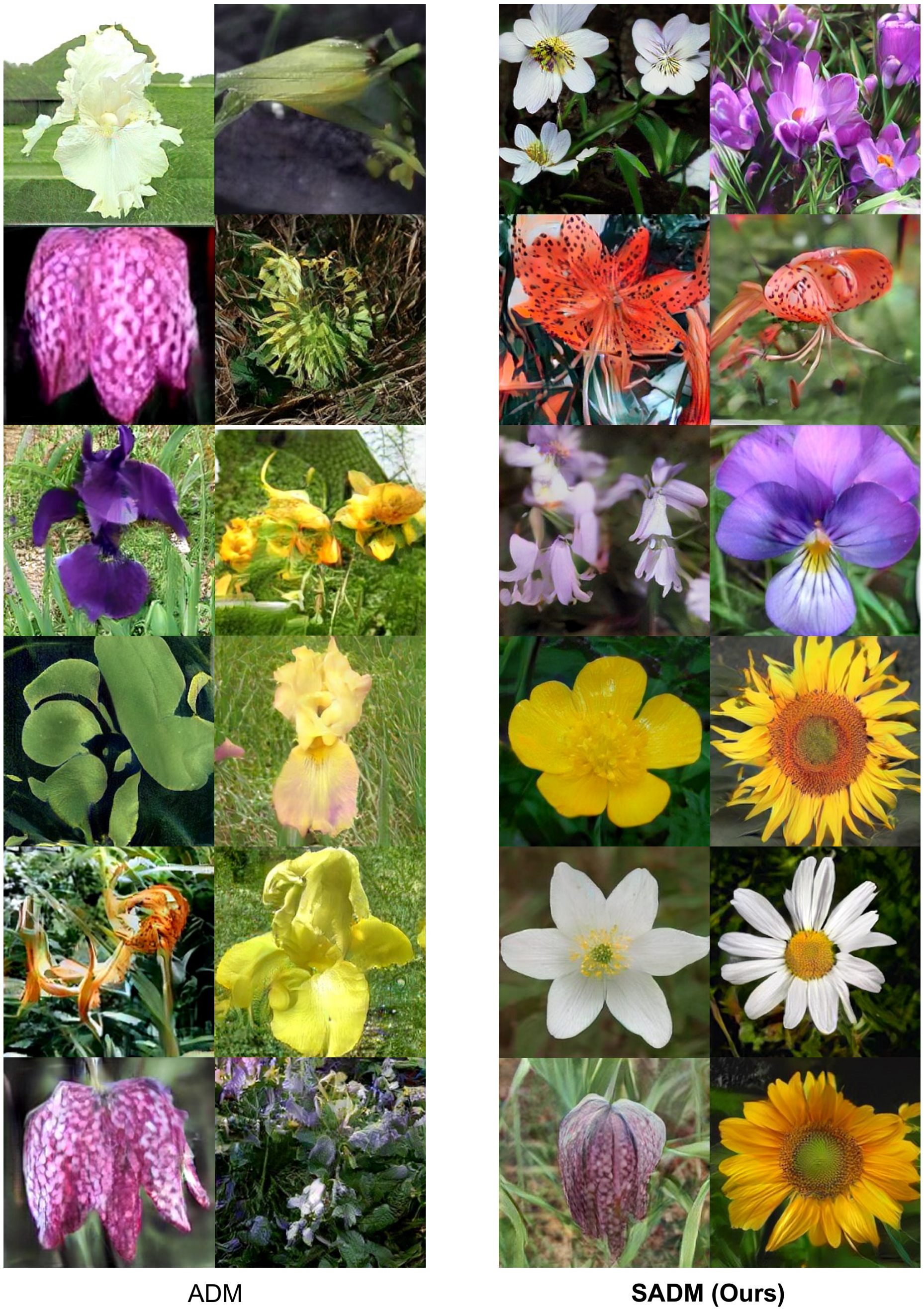


Figure 12. Random generated samples of the diffusion model fine-tuned by ADM [12] and our SADM on unconditional Oxford-Flowers.