SUPPLEMENTARY MATERIAL

1.1. Ablation Study

We perform comprehensive ablation experiments to evaluate the impact of each element in our proposed approach, as summarized in Table 3. The studies employed datasets including FFHQ and AFHQ subsets and evaluated performance using FID, where lower values indicate better results.

Initially, the baseline StyleGAN2-ADA achieved FID scores of 8.13, 3.41, 7.40, and 3.55 for FFHQ, AFHQ WILD, AFHQ DOG, and AFHQ CAT, respectively, with an average score of 5.62 across all datasets.

Introducing the ViT-D-StyleGAN2-ADA(+ViT-D) improved the FID scores across multiple datasets: notably reducing the FFHQ score to 7.87, AFHQ DOG to 6.35, and AFHQ CAT to 3.29. However, there was a slight increase in the AFHQ WILD score to 4.12, resulting in an overall improvement to an average FID of 5.41.

When DiffAug is incorporated with + ViT-D, the results were unexpectedly weaker, with higher FID scores of 10.11 (FFHQ), 4.98 (AFHQ WILD), 6.84 (AFHQ DOG), and 4.78 (AFHQ CAT), leading to an increased average FID of 6.68. This indicates that DiffAug alone may introduce excessive augmentation leading to performance degradation.

In contrast, adding +bCR alongside the ViT-D resulted in substantial improvements, yielding FID scores of 7.92 (FFHQ), 3.23 (AFHQ WILD), 5.81 (AFHQ DOG), and 3.12 (AFHQ CAT). The average FID notably decreased to 5.10, underscoring the effectiveness of combining ViT with bCR for maintaining a favorable trade-off between augmentation and regularization.

Overall, the ablation experiments unequivocally show the effectiveness of our ViT-based discriminator when enhanced with bCR, confirming the importance of carefully selected augmentations and transformer-specific regularizations in achieving optimal performance under limited-data conditions.

1.2. Training Observations and Augmentation Impact

We further study augmentation on FFHQ and CIFAR-10. Figure 8(a) contrasts training with vs. without ADA, showing that ADA markedly improves stability and final quality for our ViT-D discriminator. Figure 8(b) examines the augmentation probability p: modest augmentation ($p \approx 0.5$) yields the best trade-off, whereas too little ($p=0.2$) overfits and too much ($p=0.8$) risks leakage-like artifacts consistent with the non-leaking transform principle.

Figure 9 decomposes augmentation categories: geometric and color transforms contribute most to FID reduction, while heavy filtering, noise, and cutout offer diminishing returns. These trends guided our final operator choices for bCR and DiffAug.

1.3. Loss Function and Energy Consumption

TABLE I. ABLATION EXPERIMENTAL RESULT WITH FID ↓ METRICS OF DIFFERENT METHODS ON VARIOUS DATASETS.

Method	FFHQ	AFHQ WILD	AFHQ DOG	AFHQ CAT	Average
Baseline	8.13	3.41	7.40	3.55	5.62
+ViT-D	**7.87**	4.12	6.35	3.29	5.41
+ViT-D + DiffAug	10.11	4.98	6.84	4.78	6.68
+ViT-D + bCR	7.92	**3.23**	**5.81**	**3.12**	**5.10**

Figure 10 tracks discriminator/generator loss signals on FFHQ and shows that token-level R1 and class-token PLR align the CNN generator with the ViT-D decision function, improving stability while preserving global structure.

Finally, Table 5 summarizes computational effort and electricity consumption measured following Green500-style procedures. Experiments were executed on dual RTX 3090

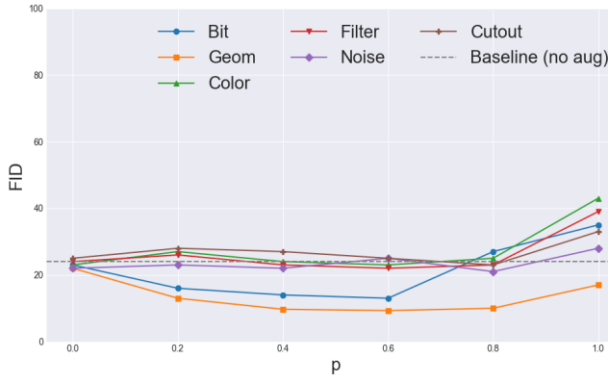

Fig. 1. Impact of parameter p on the performance of different augmentation categories, illustrating how variations in *p* affect the FID score across distinct augmentation techniques.

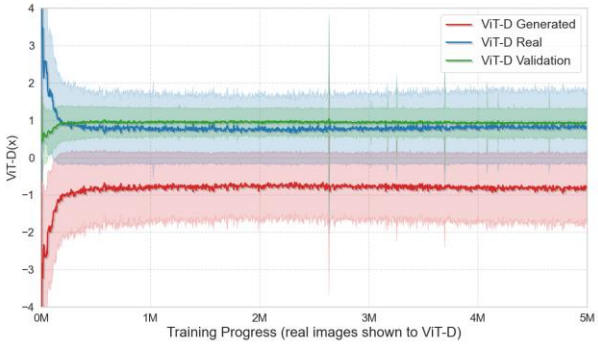

Fig. 2. Visualization of modified loss metrics during training on the FFHQ dataset, depicting the progression of the ViT-D discriminator's performance across generated, real, and validation images over the course of training.

TABLE II. COMPUTATIONAL EFFORT EXPENDITURE AND ELECTRICITY CONSUMPTION DATA FOR THIS RESEARCH.

Dataset	Number of Runs	Memory Usages (GB)	Time per Run (days)	GPU-years (Dule RTX 3090)	Electricity (MWh)
AFHQ	24	8.2 ±0.5	12.5	0.41	
CIFAR-10	16	2.7 ±0.5	2.5	0.05	8.96
FFHQ	100	4.5 ±0.5	4.9	0.67	

GPUs. Notably, a substantial fraction of the total energy was spent during early exploratory runs; the ViT-D swap itself does not impose a large marginal cost relative to a single end-to-end training run.