TCCD APPENDIX

Shiying Yu

[1] University of Electronic Science and Technology of China, Chengdu, China
[2] Ubiquitous Intelligence and Trusted Services Key Laboratory of Sichuan Province, Chengdu, China

1. EXPERIMENTS AND VISUALIZATIONS OF THE SFID METRIC

CIFAR10, We find that applying only Stage 1: Timesteps-Noise Alignment via Multimodal Contrastive Learning-achieves superior performance under the sFID metric. Upon convergence, the model attains an sFID score of 7.04, representing an improvement of 4.73 over the baseline. A comparative analysis of individual stages (Stage 1 vs. Stage 2), as well as the full TCCD framework versus the baseline DKDM, further demonstrates the effectiveness of our proposed components in enhancing generative quality, is provided in Fig 1.

ImageNet, we observe that our TCCD method achieves earlier convergence in terms of the sFID metric compared to the baseline approaches. Moreover, TCCD consistently outperforms DKDM throughout the training process, demonstrating faster optimization and superior final performance. Specifically, TCCD achieves an sFID score of **4.54**, which represents an improvement of **0.7** over data-based training and **0.39** compared to the DKDM baseline. These results highlight the effectiveness of TCCD in learning more accurate and semantically coherent generative distributions under the data-free distillation setting.Detailed visualizations are provided in Fig 1.

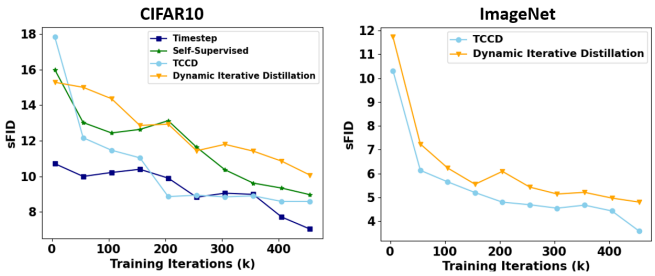

Fig. 1. Comparison of sFID Performance Across Training Iterations: TCCD vs. Dynamic Iterative Distillation on CIFAR10 and ImageNet

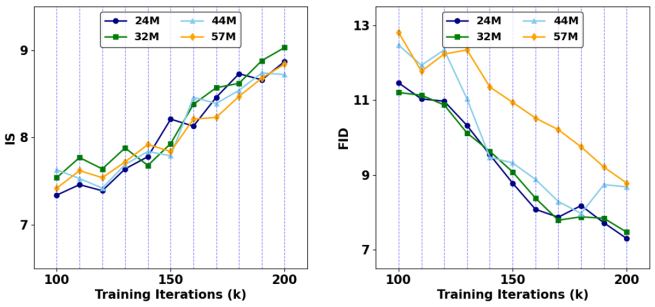

Fig. 2. Performance Evaluation of Different Model Sizes on CIFAR10: IS, FID, Metrics Across Training Iterations

2. VISUALIZATION OF DIFFERENT MODELSIZES

The Fig 2 presents a comparative analysis of four diffusion model variants—24M, 32M, 44M, and 57M parameters—evaluated on the CIFAR10 dataset across training iterations using two widely adopted generative metrics: Inception Score (IS) and Fréchet Inception Distance (FID). In the left panel, the IS curves indicate that both the 24M and 32M models achieve strong and consistent improvements in image quality and diversity, with the 32M model attaining the highest final score of approximately 9.0. The 24M model also demonstrates steady progress, reaching an IS of about 8.6. In contrast, the larger 44M and 57M models exhibit more fluctuating behavior and diminishing returns, suggesting that excessive model capacity may hinder stable learning under the current training regime. In the right panel, the FID results further confirm the superior performance of the smaller architectures. The 24M model achieves the lowest final FID of around 7.0, followed closely by the 32M model at approximately 7.2, both showing significant and consistent reductions in divergence from the real data distribution. Conversely, the 44M and 57M models converge to higher FID values (8.5 and 8.8, respectively), with less stable trajectories. These findings collectively indicate that, within this experimental setup, moderate-sized models (24M–32M) strike an optimal balance between representational capacity and training efficiency, outperforming their larger counterparts in both perceptual quality and distributional fidelity.

This work was supported by the Sichuan Science and Technology Program (granted No. 2024ZDZX0011 and No. 2025ZNSFSC1472).