Supplementary Materials for Submission 11223: TAG

	Deblur		**Super-resolution**		**CIFAR10**		**ImageNet**		**Audio declipping**		**Audio inpainting**	
Method	FID↓	LPIPS↓	FID↓	LPIPS↓	FID↓	Acc.↑	FID↓	Acc.↑	FAD↓	DTW↓	FAD↓	DTW↓
TFG + TAG (ours)	**62.7**	**0.151**	**64.7**	**0.175**	**102.7**	61.5	**219.4**	17.8	**0.74**	120	**0.42**	**51**
TFG + TCS [15]	96.5	0.350	188.7	0.518	160.5	**63.7**	297.9	15.1	20.03	549	6.25	446
TFG + Timestep Guidance [16]	469.8	0.951	483.8	0.973	371.5	11.3	536.5	**25.0**	33.67	**112**	21.96	900
TFG + Self-Guidance [17]	297.7	0.612	426.9	0.711	188.6	42.9	280.5	14.5	35.05	116	19.39	892

	Polarizability α		**Dipole** μ		**Heat capacity** C_v		$\epsilon_{\mathbf{HOMO}}$		$\epsilon_{\mathbf{LUMO}}$		**Gap** ϵ_{Δ}	
Method	MAE↓	Stab.↑	MAE↓	Stab.↑	MAE↓	Stab.↑	MAE↓	Stab.↑	MAE↓	Stab.↑	MAE↓	Stab.↑
TFG + TAG (ours)	**4.46**	43.6	**1.28**	94.3	**2.67**	96.7	**0.43**	**93.9**	**0.89**	92.5	**0.78**	82.8
TFG + TCS [15]	5.40	**99.2**	1.43	**99.2**	3.35	**99.1**	N/A	N/A	1.22	**99.2**	1.31	**99.2**
TFG + Timestep Guidance [16]	10.98	84.4	N/A	N/A	4.09	85.5	0.70	83.5	1.36	73.0	1.30	83.5
TFG + Self-Guidance [17]	7.66	80.2	32.54	80.3	3.80	80.3	N/A	N/A	1.32	80.3	1.27	80.4

Table R1: Comparison of different guidance methods. The best result for each metric is highlighted in **bold**.

Method	FID ↓	Acc. ↑
DPS	217.1	57.5
TAG (ours)	190.4	**63.2**
TCS [15]	213.4	29.4
Timestep Guidance [16]	393.2	9.4
Self-Guidance [17]	205.4	51.6
Epsilon Scaling [10]	**186.0**	53.0
Time Shift Sampler [11]	237.0	60.8
Langevin Dynamics [13]	226.8	58.2

Table R2: Additional baselines when applying DPS on CIFAR-10. TAG improves the performance of DPS while other method struggles.

Method	FID ↓	Acc. ↑
$\eta = 0$	332.0	28.5
$\eta = 0.05$	409.9	23.3
$\eta = 0.10$	376.6	25.4
$\eta = 0.15$	326.7	29.2

Table R3: Effect of Input perturbation on DPS, CIFAR-10. For fair comparison, we train diffusion models with different η from scratch following the official implementation code in [9]. No improvement over original diffusion model ($\eta = 0$) is observed in the presence of off-manifold phenomenon. We report the average value for 512 samples per each conditioning labels.

Method	FID ↓	Acc. ↑
512 samples		
TFG	114.1	55.8
TFG + TAG (ours)	102.7	61.5
50000 samples		
TFG	77.5	54.3
TFG + TAG (ours)	**47.1**	**84.4**

Table R4: Originally, 512 samples were used for rapid, extensive experiments across various tasks. For a more rigorous evaluation, we used 50,000 samples on CIFAR-10 with 100 inference steps. As expected, increasing the number of samples to match standard benchmark protocols led to improved FID scores.

Method	FID ↓	Acc. ↑
Original Update Order		
DPS + TAG	**190.4**	**63.2**
TFG + TAG	**102.7**	**61.5**
Changed Update Order		
DPS + TAG	203.5	60.1
TFG + TAG	116.4	54.1

Table R5: Effect of original vs. changed update order for Algorithm 1 (TAG) on CIFAR-10.

	Training Steps	
	10K	30K
FID ↓	116.0	**102.7**
Acc. ↑	55.3	**61.5**

Table R6: Quantitative evaluation of TFG+TAG across varying training steps on CIFAR-10 confirms the relationship between classifier robustness and TAG performance.

Layers	W1 distance ↓
0 (No TAG)	6.458
1	1.716
2	1.681
3	1.975
4	1.714
5	1.713
6	1.788

Table R7: Robustness of time classifier network on toy experiment. We measure Wasserstein distance (W_1) for 10,000 samples. Consistent improvement compared to original reverse process when applying TAG independent of layer numbers.

(a) Original images of corrupted reverse process (b) Images after applying TAG

Figure R1: Uncurated samples of images in the corrupted reverse process in CIFAR-10 experiment.

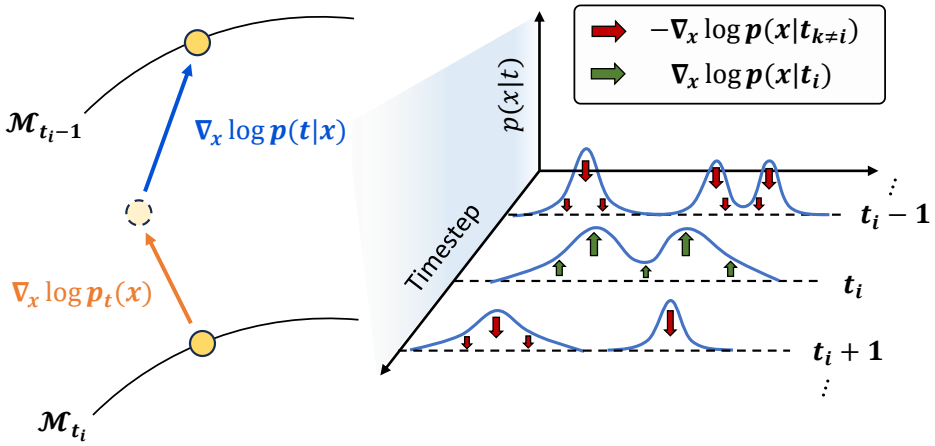

Figure R2: Overview of TAG algorithm. In the left figure, we specifically visualized the decomposition of $\nabla_x \log p(t_i|x)$ according to Eq. 11 in Lemma 3.3.

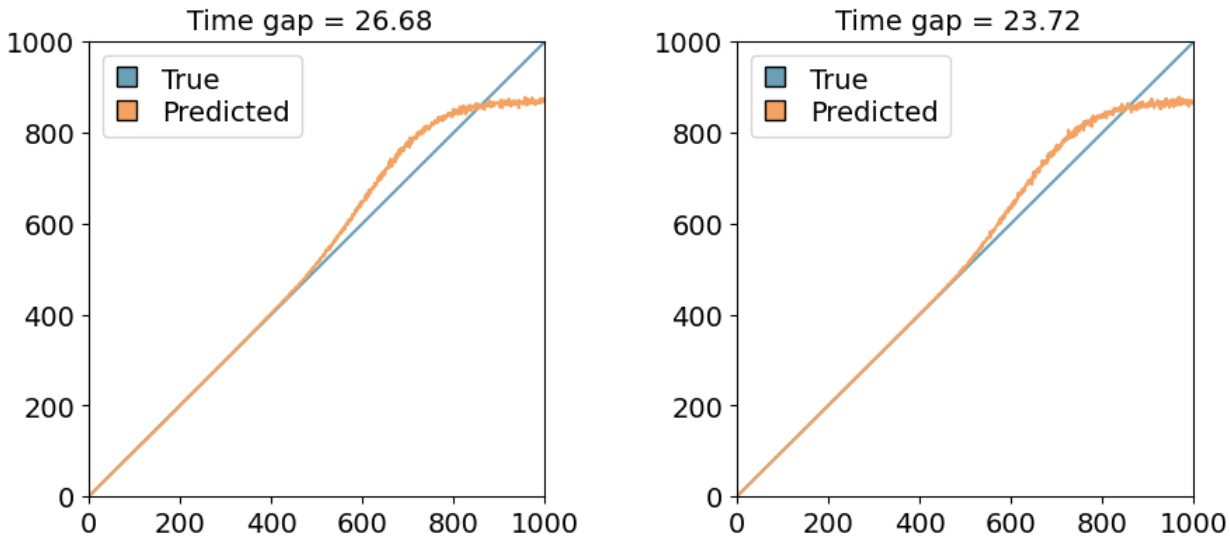

Figure R3: Time Gap (Def. 4.1) in CIFAR-10. The left panel shows 10K training iterations and the right panel 30K iterations. The less-trained classifier exhibits a larger time gap.

References

[1] Han, D., et al. (2024). Understanding training-free diffusion guidance: Mechanisms and limitations. *arXiv:2403.12404*.

[2] Ho, J., & Salimans, T. (2022). Classifier-Free Diffusion Guidance. *arXiv:2207.12598*.

[3] He, Y., et al. (2024). CFG++: Manifold-constrained classifier-free guidance for diffusion models. *arXiv:2406.08070*.

[4] Lin, C. H., et al. (2025). Diffusion models without classifier-free guidance. *arXiv:2502.12154*.

[5] Shi, J., et al. (2023). Language-driven scene synthesis using multi-conditional diffusion model. *arXiv:2310.15948*.

[6] Schneuing, A., et al. (2024). Inverse molecular design with multi-conditional diffusion guidance. *arXiv:2401.13858*.

[7] Geon, P., et al. (2024). Inference-time diffusion model distillation. *arXiv:2412.08871*.

[8] Li, J., et al. (2023). On error propagation of diffusion models. *arXiv:2308.05021*.

[9] Ning, Z., Li, W., He, D., & Zhang, L. (2023). Input perturbation reduces exposure bias in diffusion models. In *Proceedings of the International Conference on Machine Learning (ICML)* (arXiv:2301.11706).

[10] Ning, Z., Li, W., He, D., & Zhang, L. (2024). Elucidating the exposure bias in diffusion models. In *Proceedings of the International Conference on Learning Representations (ICLR)* (arXiv:2308.15321).

[11] Li, Z., Liu, J., & Zhang, L. (2024). Alleviating exposure bias in diffusion models through sampling with shifted time steps. In *Advances in Neural Information Processing Systems*.

[12] Le, N. A. K., Nguyen, T., & Tran, A. T. (2024). Classification diffusion models: Revitalizing density ratio estimation. In *Proceedings of the Neural Information Processing Systems (NeurIPS)* (arXiv:2402.10095).

[13] Song, Y., & Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. In *Proceedings of the Neural Information Processing Systems (NeurIPS)* (arXiv:1907.05600).

[14] Ye, H., et al. (2024). TFG: Unified training-free guidance for diffusion models. *arXiv:2409.15761*.

[15] Jung, H., Park, Y., Schmid, L., Jo, J., Lee, D., Kim, B., Yun, S.-Y., & Shin, J. (2024). Conditional synthesis of 3D molecules with time correction sampler. In *Advances in Neural Information Processing Systems*, 37.

[16] Sadat, S., Kansy, M., Hilliges, O., & Weber, R. M. (2024). No training, no problem: Rethinking classifier-free guidance for diffusion models. *arXiv:2407.02687*.

[17] Li, T., Luo, W., Chen, Z., Ma, L., & Qi, G. J. (2024). Self-guidance: Boosting flow and diffusion generation on their own. *arXiv:2412.05827*.

[18] Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-based generative modeling through stochastic differential equations. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[19] Zhang, L., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[20] Chung, H., et al. (2022). Diffusion posterior sampling for general noisy inverse problems. *arXiv:2209.14687*.

[21] Bhat, S. F., Mitra, N., & Wonka, P. (2024). Loosecontrol: Lifting ControlNet for generalized depth conditioning. In *ACM SIGGRAPH 2024 Conference Papers*, 1–11.

[22] Lin, H., Cho, J., Zala, A., & Bansal, M. (2024). Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. *arXiv:2404.09967*.

[23] Yang, J., Zhao, J., Wang, P., Wang, Z., & Liang, Y. (2025). Meta ControlNet: Enhancing task adaptation via meta learning. In *Proceedings of The Second Conference on Parsimony and Learning (Proceedings Track)*. Available at `https://openreview.net/forum?id=ju63pUpq0N`.

[24] Rout, L., et al. (2025). RB-modulation: Training-free personalization of diffusion models using stochastic optimal control. In *Proceedings of the International Conference on Learning Representations (ICLR)* (arXiv:2405.17401).

[25] Bar-Tal, O., et al. (2023). MultiDiffusion: Fusing diffusion paths for controlled image generation. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[26] Yu, J., Wang, Y., Zhao, C., Ghanem, B., & Zhang, J. (2023). FreeDoM: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[27] He, Y., Murata, N., Lai, C. H., Takida, Y., Uesaka, T., Kim, D., Liao, W. H., Mitsufuji, Y., Kolter, J. Z., Salakhutdinov, R., & Ermon, S. (2024). Manifold preserving guided diffusion (MPGD). In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[28] Du, Y., Mao, J., & Tenenbaum, J. B. (2024). Learning iterative reasoning through energy diffusion. *arXiv preprint arXiv:2406.11179*.

[29] Bansal, A., Chu, H. M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., & Goldstein, T. (2023). Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 843–852).

[30] Song, J., Zhang, Q., Yin, H., Mardani, M., Liu, M. Y., Kautz, J., & Vahdat, A. (2023, July). Loss-guided diffusion models for plug-and-play controllable generation. In *International Conference on Machine Learning* (pp. 32483–32498). PMLR.

[31] Yu, J., et al. (2022). Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*.

[32] Saleh, B., & Elgammal, A. (2015). Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*.

[33] Deng, J., et al. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

[34] Schuhmann, C., et al. (2022). Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 25278–25294).

[35] Wu, X., et al. (2023). Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*.

[36] Kirstain, Y., et al. (2023). Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems* (Vol. 36, pp. 36652–36663).

[37] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (pp. 8748–8763). PMLR.

[38] Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., & Choi, Y. (2021). CLIPScore: A Reference-Free Evaluation Metric for Image Captioning. *arXiv preprint arXiv:2104.08718*.

[39] Chung, H., Kim, J., Park, G. Y., Nam, H., & Ye, J. C. (2024). Cfg++: Manifold-constrained classifier free guidance for diffusion models. *arXiv preprint arXiv:2406.08070*.