Diagnosing and Improving Diffusion Models by Estimating the Optimal Loss Value

Yixian Xu[1*], Shengjie Luo[1*], Liwei Wang[1], Di He[1†], Chang Liu[2†]

[1] *State Key Laboratory of General Artificial Intelligence, Peking University, Beijing, China*

[2] *Zhongguancun Academy, Beijing, China*

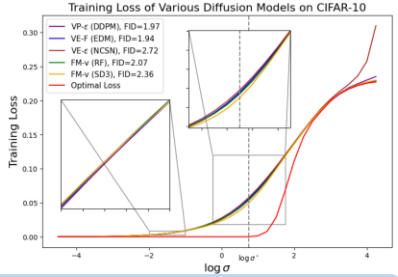
Diffusion-Model Training Loss

$$\min_{\theta} \mathbb{E}_{p(t)} w_t \left\{ J_t(\boldsymbol{\theta}) := \mathbb{E}_{p_0(\mathbf{x}_0) p(\mathbf{x}_t|\mathbf{x}_0)} \left\| \mathbf{x}_{0\boldsymbol{\theta},t}(\mathbf{x}_t) - \mathbf{x}_0 \right\|^2 \right\}$$

➔ Optimal loss $J_t^\star = \operatorname{tr} \operatorname{Cov}_{p(\mathbf{x}_0|\mathbf{x}_t)}[\mathbf{x}_0] > 0$ and unknown!

Loss value does not reflect **absolute fit to data**:

- Cannot diagnose **training status** from training loss.
- Cannot compare **data fitness** across time steps.



Optimal Loss Estimators

$$J_t^\star = \underbrace{\mathbb{E}_{p_0(\mathbf{x}_0)} \|\mathbf{x}_0\|^2}_{=:A} - \underbrace{\mathbb{E}_{p_t(\mathbf{x}_t)} \left\| \mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_t)}[\mathbf{x}_0] \right\|^2}_{=:B_t}$$

(1) For the first term: $\hat{A} = \frac{1}{N} \sum_{n\in[N]} \left\| \mathbf{x}_0^{(n)} \right\|^2$.

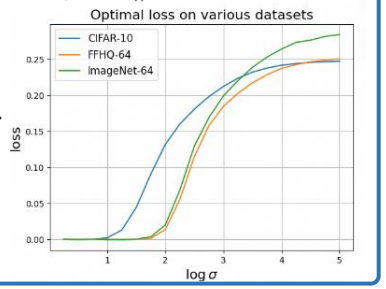
(2) For the second term:

$$\mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_t)}[\mathbf{x}_0] = \mathbb{E}_{p(\mathbf{x}_0)} \left[\frac{p(\mathbf{x}_0|\mathbf{x}_t)}{p(\mathbf{x}_0)} \mathbf{x}_0 \right] \propto \mathbb{E}_{p(\mathbf{x}_0)} [p(\mathbf{x}_t|\mathbf{x}_0)\, \mathbf{x}_0]$$

➔ $$\hat{B}_t = \frac{1}{M} \sum_{m\in[M]} \left\| \frac{\sum_{n\in[N]} \mathbf{x}_0^{(n)} K_t(\mathbf{x}_t^{(m)}, \mathbf{x}_0^{(n)})}{\sum_{n'\in[N]} K_t(\mathbf{x}_t^{(m)}, \mathbf{x}_0^{(n')})} \right\|^2$$

where

$$K_t(\mathbf{x}_t, \mathbf{x}_0) := \exp\left\{ -\frac{\|\mathbf{x}_t - \alpha_t \mathbf{x}_0\|^2}{2\sigma_t^2} \right\}$$



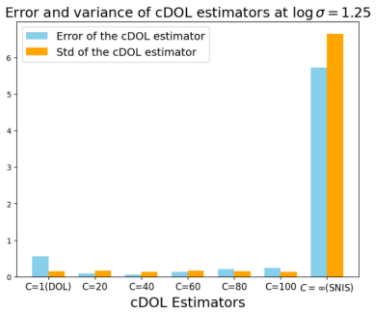
Scalable Estimators by Dataset Subsampling

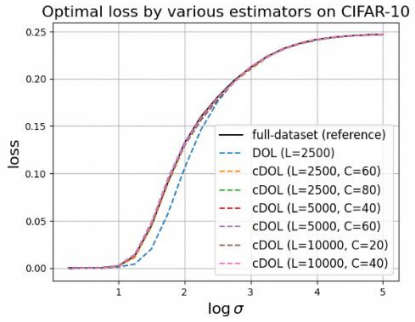
$$\hat{B}_t^{\mathrm{SNIS}} := \frac{1}{M} \sum_{m\in[M]} \left\| \frac{\sum_{l\in[L]} \mathbf{x}_0^{(l)} K_t(\mathbf{x}_t^{(m)}, \mathbf{x}_0^{(l)})}{\sum_{l'\in[L]} K_t(\mathbf{x}_t^{(m)}, \mathbf{x}_0^{(l')})} \right\|^2$$ ✗ large variance

$$\hat{B}_t^{\mathrm{DOL}} := \frac{1}{M} \sum_{\tilde{m}\in[M]} \left\| \frac{\sum_{l\in[L]} \mathbf{x}_0^{(l)} K_t(\mathbf{x}_t^{(\tilde{m})}, \mathbf{x}_0^{(l)})}{\sum_{l'\in[L]} K_t(\mathbf{x}_t^{(\tilde{m})}, \mathbf{x}_0^{(l')})} \right\|^2$$ ✗ large bias

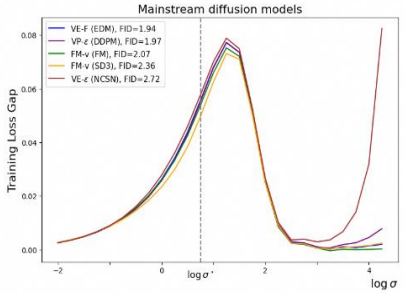
✓ **Corrected: Properly balancing variance and bias!**

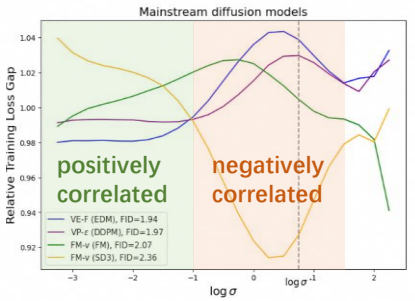
$$\hat{B}_t^{\mathrm{cDOL}} := \frac{1}{M} \sum_{\tilde{m}\in[M]} \left\| \frac{\sum_{l\in[L], l\neq l_{\tilde{m}}} \mathbf{x}_0^{(l)} K_t(\mathbf{x}_t^{(\tilde{m})}, \mathbf{x}_0^{(l)}) + \frac{1}{C} \mathbf{x}_0^{(l_{\tilde{m}})} K_t(\mathbf{x}_t^{(\tilde{m})}, \mathbf{x}_0^{(l_{\tilde{m}})})}{\sum_{l'\in[L], l'\neq l_{\tilde{m}}} K_t(\mathbf{x}_t^{(\tilde{m})}, \mathbf{x}_0^{(l')}) + \frac{1}{C} K_t(\mathbf{x}_t^{(\tilde{m})}, \mathbf{x}_0^{(l_{\tilde{m}})})} \right\|^2$$





Stepwise loss gap vs. **generation performance**:





Principled Training Schedule Design

$$w_\sigma = \underbrace{a \min\left\{1/J_\sigma^\star, w^\star\right\}}_{\text{normalize the scale}} \underbrace{+ f(\sigma)\, \mathbb{I}_{\sigma<\sigma^\star}}_{\text{upweight the pos. correl. region}}$$

$$p(\sigma) \propto w_\sigma (J_\sigma(\boldsymbol{\theta}) - J_\sigma^\star)$$

Table 3: Comparison between existing training schedules and ours on ImageNet-256 dataset.

Method	Generation w/o CFG				Generation w/ CFG			
	FID(↓)	IS(↑)	Pre.(↑)	Rec.(↑)	FID(↓)	IS(↑)	Pre.(↑)	Rec.(↑)
Pixel-space Diffusion Models								
ADM (Dhariwal and Nichol, 2021)	10.94	–	0.69	0.63	3.94	215.9	0.83	0.53
RIN (Jabri et al., 2022)	3.42	182.0	–	–	–	–	–	
Simple Diffusion (Hoogeboom et al., 2023)	2.77	211.8	–	–	2.12	256.3	–	–
VDM++ (Kingma and Gao, 2023)	2.40	**225.3**	–	–	2.12	267.7	–	–
SiD2 (Hoogeboom et al., 2024)	–	–	–	–	1.38	–	–	–
Latent Diffusion Models								
MaskDiT (Zheng et al., 2023)	5.69	177.9	0.74	0.60	2.28	276.6	0.80	0.61
DiT (Peebles and Xie, 2023)	9.62	121.5	0.67	0.67	2.27	278.2	**0.83**	0.57
SiT (Ma et al., 2024)	8.61	131.7	0.68	0.67	2.06	270.3	0.82	0.59
FasterDiT (Yao et al., 2024)	7.91	131.3	0.67	**0.69**	2.03	264.0	0.81	0.60
MDT (Gao et al., 2023a)	6.23	143.0	0.71	0.65	1.79	283.0	0.81	0.61
MDTv2 (Gao et al., 2023b)	–	–	–	–	1.58	**314.7**	0.79	0.65
REPA (Yu et al., 2024)	5.90	–	–	–	1.42	305.7	0.80	0.65
LightningDiT (Yao et al., 2025)	2.17	205.6	**0.77**	0.65	1.35	295.3	0.79	0.65
+ reproduction	2.29	206.2	0.76	0.66	1.42	292.9	0.79	0.65
+ our schedule	**2.08**	220.8	**0.77**	0.66	**1.30**	301.3	0.79	**0.66**

Principled Scaling Law Study

$$J(F) = \beta F^\alpha \quad \Rightarrow \quad J(F) - J^\star = \beta F^\alpha$$