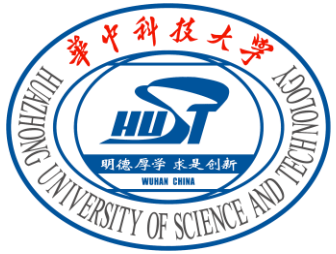


華中科技大學
HUAZHONG UNIVERSITY SCIENCE AND TECHNOLOGY

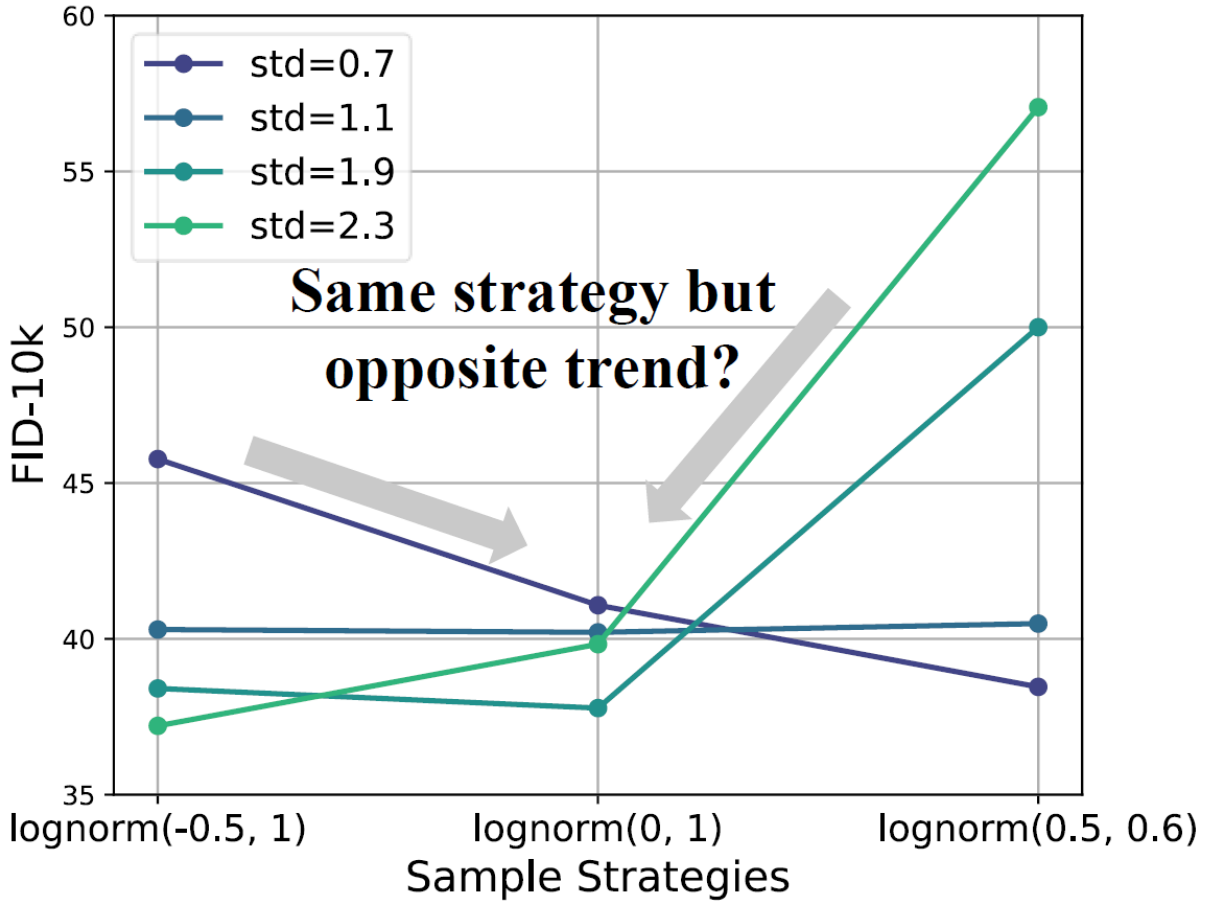
FasterDiT: Towards Faster Diffusion Transformers Training without Architecture Modification

Jingfeng Yao, Cheng Wang, Wenyu Liu, Xinggang Wang*

Highlights

- Interpreting the performance robustness with the perspective of the Probability Density Function (PDF) of SNR during training.
- Conducting extensive experiments and report about one hundred experiment results in our paper to empirically analyze the association between training performance and robustness with PDF.
- Introducing a new supervision method for velocity prediction-based approaches.

Problem Source



Probability Density Function of SNR during Training

- A tool to analyse the SNR

$f_s(t)$ denotes the timestep sampling function. $f_l(t)$ denotes the loss weight function. $f_t(t)$ is to unify them into a single distribution of t.

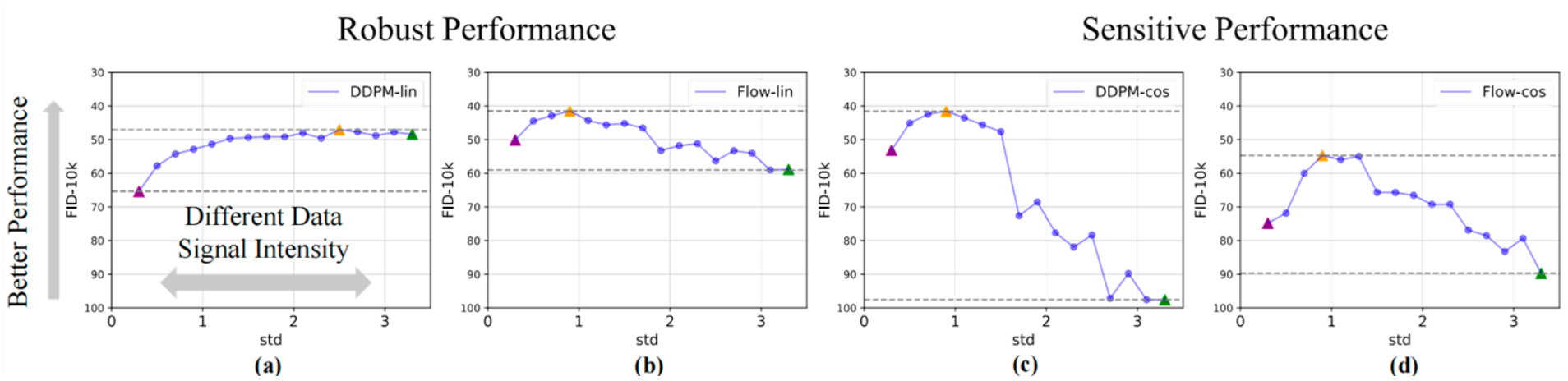
$$\mathrm{SNR}(t) = \frac{K(I)}{\mathrm{std}^2}\frac{\mathrm{std}^2\alpha_t^2}{\sigma_t^2} \approx C(I)\frac{\mathrm{std}^2\alpha_t^2}{\sigma_t^2}$$

$$f_t(t) = \frac{f_l(t)f_s(t)}{\int_0^1 f_l(t)f_s(t)\,dt}$$

$$f_Y(y) = f_t(g(y))\left|\frac{d}{dy}g(y)\right|$$

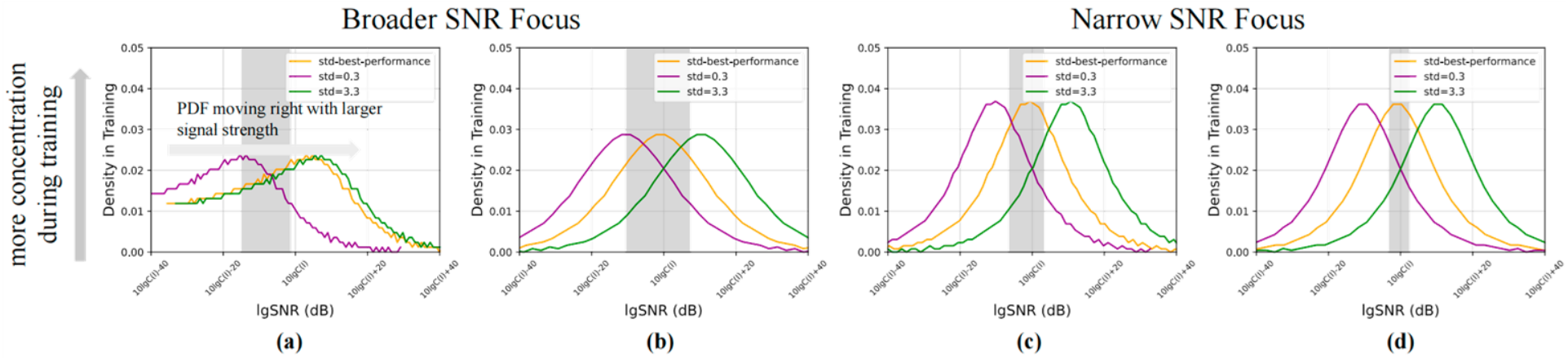
What we can learn from SNR PDF

- Different data signal intensities lead to different training effects.



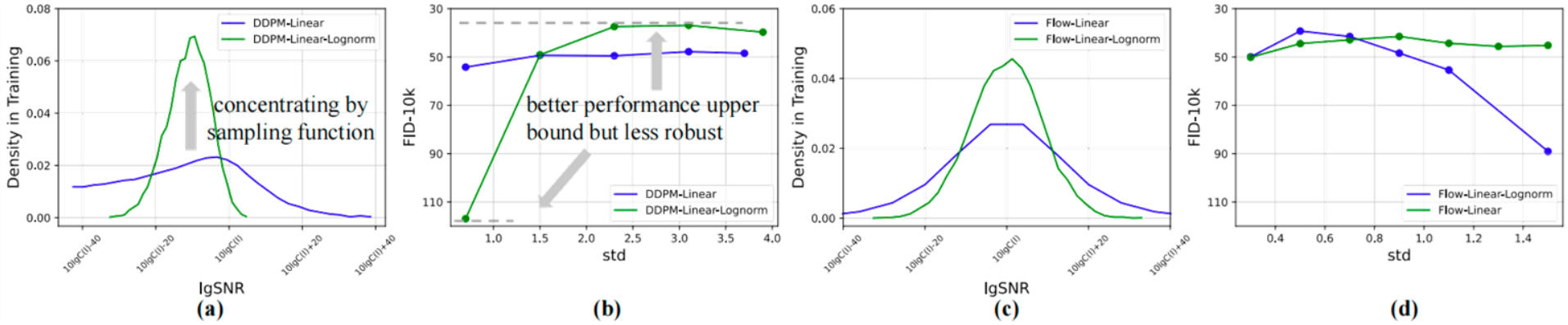
What we can learn from SNR PDF

- Different schedules exhibit significant differences in data robustness.
- Schedule with a broad SNR focus could have a robust performance.
- The optimal SNR ranges for different schedules seem to be similar



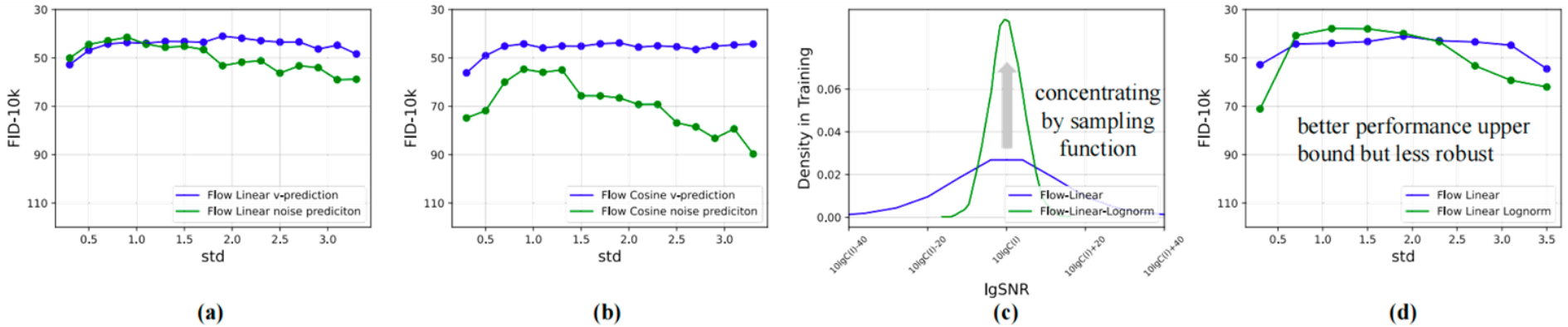
What we can learn from SNR PDF

- There is a trade-off between performance and robustness.



What we can learn from SNR PDF

- Flow matching with v-prediction gets more robust performance.
- The trade-off still exists in flow matching with v-prediction.



Improving DiT Training

- Improving Single Step Supervision
- we not only use the Mean Squared Error (MSE) to supervise velocity predictions but also apply cosine similarity to further supervise the directionality of velocity.

$$L_d = 1 - \frac{1}{HW} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} \frac{v_{gt}^{(h,w)} \cdot v_{pred}^{(h,w)}}{|v_{gt}^{(h,w)}||v_{pred}^{(h,w)}|}$$

Results

Method	Model	Training Iters	FID↓	sFID↓	IS↑	Prec.↑	Rec.↑
BigGAN [4]	BigGAN-deep	-	6.95	7.36	171.4	**0.87**	0.28
MaskGIT [6]	MaskGIT	1387k×256	6.18	-	182.1	-	-
ADM-*G* [14]	ADM	1980k×256	4.59	5.25	186.70	0.82	0.52
CDM [23]	CDM	-	4.88	-	158.71	-	-
RIN [26]	RIN	-	3.42	-	-	-	-
Simple Diffusion [25]	U-Net	2000k×512	3.76	-	-	-	-
Simple Diffusion	U-ViT-L	500k×2048	2.77	-	-	-	-
LDM-4-*G* [40]	LDM	178k×1200	3.60	-	247.67	0.87	0.48
U-ViT-*G* [2]	U-ViT	300k×1024	3.40	-	-	-	-
StyleGAN [42]	StyleGAN-XL	-	2.30	**4.02**	265.12	0.78	0.53
MDT-*G* [18]	MDT	2500k×256	2.15	4.52	249.27	0.82	0.58
DiT [37]	DiT-XL/2	7000k×256	9.62	6.85	121.50	0.67	0.67
SiT [32]	DiT-XL/2	7000k×256	8.61	6.32	131.65	0.68	0.67
FasterDiT	DiT-XL/2	**1000k**×256	8.72	5.23	121.17	0.68	0.67
	DiT-XL/2	**2000k**×256	7.91	5.46	131.27	0.67	**0.69**
DiT *(cfg=1.5)* [37]	DiT-XL/2	7000k×256	2.27	4.60	**278.24**	0.83	0.57
SiT *(cfg=1.5)* [32]	DiT-XL/2	7000k×256	2.06	4.50	270.27	0.82	0.59
FasterDiT *(cfg=1.5)*	DiT-XL/2	**1000k**×256	2.30	4.80	249.34	0.82	0.58
	DiT-XL/2	**2000k**×256	**2.03**	4.63	263.95	0.81	0.60

Results


index:333 hamster
index:388 panda
index:99 goose
index:207 golden retriever