Intriguing Properties of Data Attribution on Diffusion Models

SMU SINGAPORE MANAGEMENT UNIVERSITY

sea AI Lab

Xiaosen Zheng [1], Tianyu Pang [2], Chao Du [2], Jing Jiang [1], Min Lin [2]

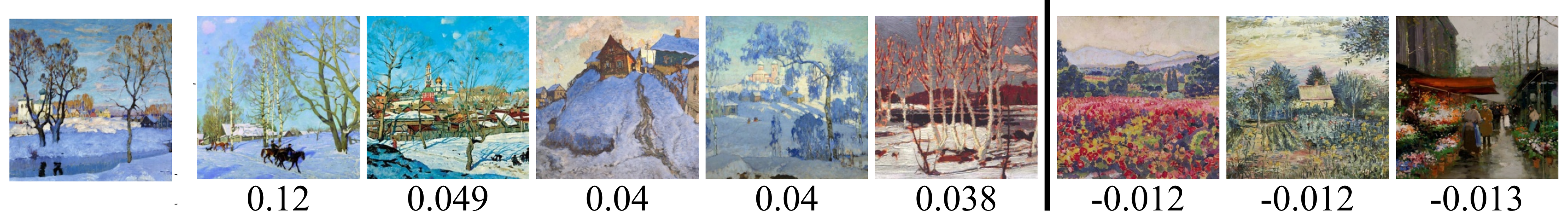
[1]*Singapore Management University* [2]*Sea AI Lab*

ICLR | 2024

SCAN ME

• Data Attribution

Data attribution method τ assigns an importance score to each training sample given a test sample



τ can predict model output $g_\tau(\boldsymbol{x}, \mathcal{D}'; \mathcal{D}) \triangleq \sum_{\boldsymbol{x}^n \in \mathcal{D}'} \tau(\boldsymbol{x}, \mathcal{D})_n$

evaluated by $\mathrm{LDS}(\tau, \boldsymbol{x}) \triangleq \rho(\{\boldsymbol{\mathcal{F}}(\boldsymbol{x}; \theta^*(\mathcal{D}^m)) : m \in [M]\}, \{g_\tau(\boldsymbol{x}, \mathcal{D}^m; \mathcal{D}) : m \in [M]\})$

An implementation of **τ**: TRAK (Park et al. ICML 2023)

$$\tau_{\mathrm{TRAK}}(\boldsymbol{x}, \mathcal{D}) = \left[\frac{1}{S}\sum_{s=1}^{S} \phi^s(\boldsymbol{x})^\top \left({\Phi^s_{\mathrm{TRAK}}}^\top \Phi^s_{\mathrm{TRAK}}\right)^{-1} {\Phi^s_{\mathrm{TRAK}}}^\top\right]\left[\frac{1}{S}\sum_{s=1}^{S} \mathcal{Q}^s_{\mathrm{TRAK}}\right]$$

$$\Phi^s_{\mathrm{TRAK}} = \left[\phi^s(\boldsymbol{x}^1); \cdots; \phi^s(\boldsymbol{x}^N)\right]^\top, \text{ where } \phi^s(\boldsymbol{x}) = \mathcal{P}_s^\top \nabla_\theta \boldsymbol{\mathcal{F}}(\boldsymbol{x}; \theta^*_s);$$

$$\mathcal{Q}^s_{\mathrm{TRAK}} = \mathrm{diag}\left(Q^s(\boldsymbol{x}^1), \cdots, Q^s(\boldsymbol{x}^N)\right), \text{ where } Q^s(\boldsymbol{x}) = \frac{\partial \boldsymbol{\mathcal{L}}}{\partial \boldsymbol{\mathcal{F}}}(\boldsymbol{x}; \theta^*_s).$$

• Diffusion Model

The commonly used loss function is

$$\mathcal{L}_{\mathrm{Simple}}(\boldsymbol{x}; \theta) = \mathbb{E}_{\boldsymbol{\epsilon}, t}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\overline{\alpha}_t}\boldsymbol{x} + \sqrt{1 - \overline{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|_2^2\right]$$

We set $\boldsymbol{\mathcal{F}}(\boldsymbol{x}; \theta) = \boldsymbol{\mathcal{L}}(\boldsymbol{x}; \theta) = \mathcal{L}_{\mathrm{Simple}}(\boldsymbol{x}, \theta)$

Thus, TRAK's gradient is $\phi^s(\boldsymbol{x}) = \mathcal{P}_s^\top \nabla_\theta \mathcal{L}_{\mathrm{Simple}}(\boldsymbol{x}, \theta^*_s)$

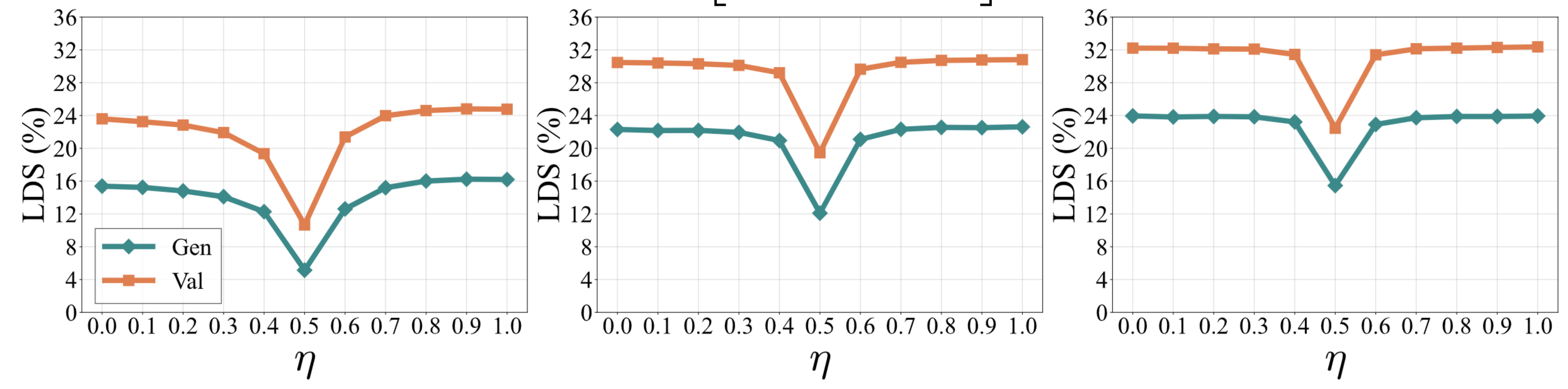
$$\nabla_\theta \mathcal{L}_{\mathrm{Simple}} = \mathbb{E}_{t, \boldsymbol{\epsilon}}\left[2 \cdot \underbrace{(\boldsymbol{\epsilon}_\theta - \boldsymbol{\epsilon})^\top}_{} \nabla_\theta \boldsymbol{\epsilon}_\theta\right]$$

Would this term lead to gradient saturation?

Hence, we consider the following interpolation:

$$\phi^s(\boldsymbol{x}) = \mathcal{P}_s^\top \nabla_\theta \left[\eta \mathcal{L}_{\mathrm{Square}} + (1-\eta)\left(\mathcal{L}_{\mathrm{Simple}} - \mathcal{L}_{\mathrm{Square}}\right)\right](\boldsymbol{x}, \theta^*_s)$$
$$= \mathcal{P}_s^\top \mathbb{E}_{t, \boldsymbol{\epsilon}}\left[2 \cdot \left(\eta \boldsymbol{\epsilon}_\theta - (1-\eta)\boldsymbol{\epsilon}\right)^\top \nabla_\theta \boldsymbol{\epsilon}_\theta\right]$$

Note that $\mathcal{L}_{\mathrm{Square}}(\boldsymbol{x}, \theta) = \mathbb{E}_{t, \boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\|_2^2\right]$



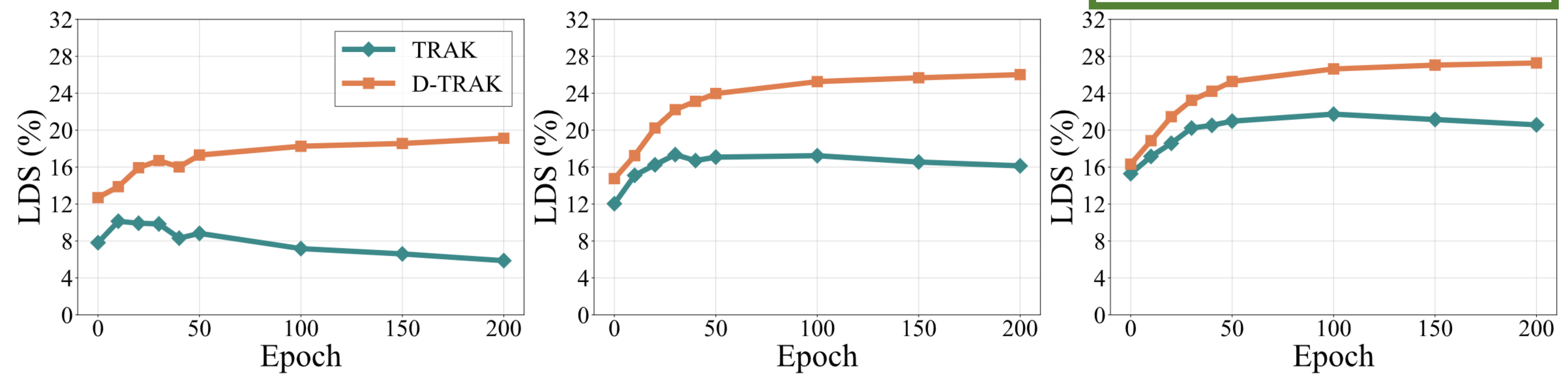
(a) # of timesteps is 10 (b) # of timesteps is 100 (c) # of timesteps is 1000

Method	Construction of $\phi^s(\boldsymbol{x})$	Validation 10	Validation 100	Validation 1000	Generation 10	Generation 100	Generation 1000
TRAK	$\mathcal{P}_s^\top \nabla_\theta \mathcal{L}_{\mathrm{Simple}}(\boldsymbol{x}, \theta^*_s)$	10.66	19.50	22.42	5.14	12.05	15.46
D-TRAK (**Ours**)	$\mathcal{P}_s^\top \nabla_\theta \mathcal{L}_{\mathrm{ELBO}}(\boldsymbol{x}, \theta^*_s)$	8.46	9.07	13.19	3.49	3.83	5.80
	$\mathcal{P}_s^\top \nabla_\theta \mathcal{L}_{\mathrm{Square}}(\boldsymbol{x}, \theta^*_s)$	24.78	30.81	32.37	16.20	22.62	23.94
	$\mathcal{P}_s^\top \nabla_\theta \mathcal{L}_{\mathrm{Avg}}(\boldsymbol{x}, \theta^*_s)$	24.91	29.15	30.39	16.76	20.82	21.48
	$\mathcal{P}_s^\top \nabla_\theta \mathcal{L}_{1\text{-norm}}(\boldsymbol{x}, \theta^*_s)$	23.44	30.36	32.29	15.10	21.99	23.78
	$\mathcal{P}_s^\top \nabla_\theta \mathcal{L}_{2\text{-norm}}(\boldsymbol{x}, \theta^*_s)$	24.72	30.91	32.35	15.75	22.44	23.82
	$\mathcal{P}_s^\top \nabla_\theta \mathcal{L}_{\infty\text{-norm}}(\boldsymbol{x}, \theta^*_s)$	5.22	11.54	22.25	3.99	8.11	15.94

$\nabla_\theta \mathcal{L}_{\mathrm{Simple}}(\boldsymbol{x}, \theta^*_s)$ **(TRAK)**

⬇

$\nabla_\theta \mathcal{L}_{\mathrm{Square}}(\boldsymbol{x}, \theta^*_s)$ **(D-TRAK)**

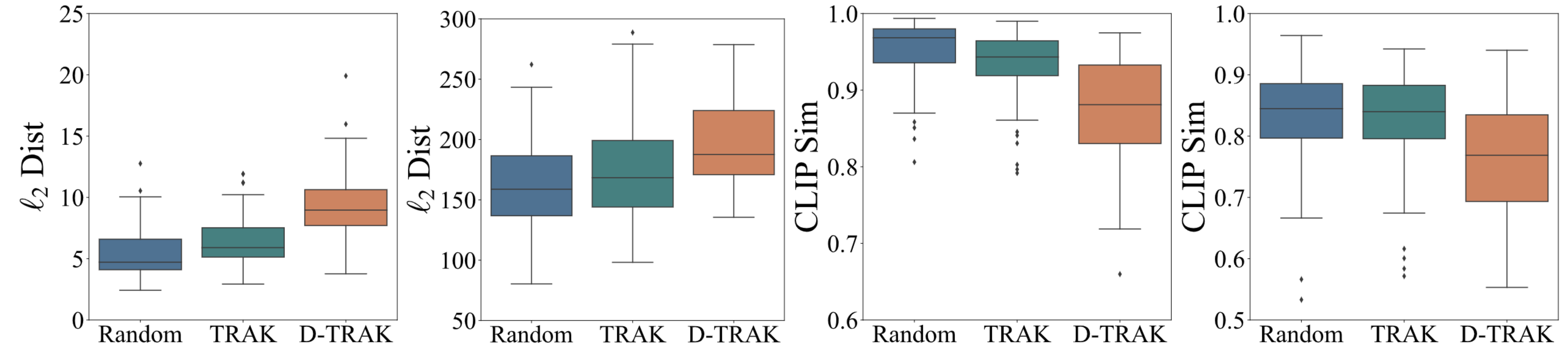


(a) # of timesteps is 10 (b) # of timesteps is 100 (c) # of timesteps is 1000

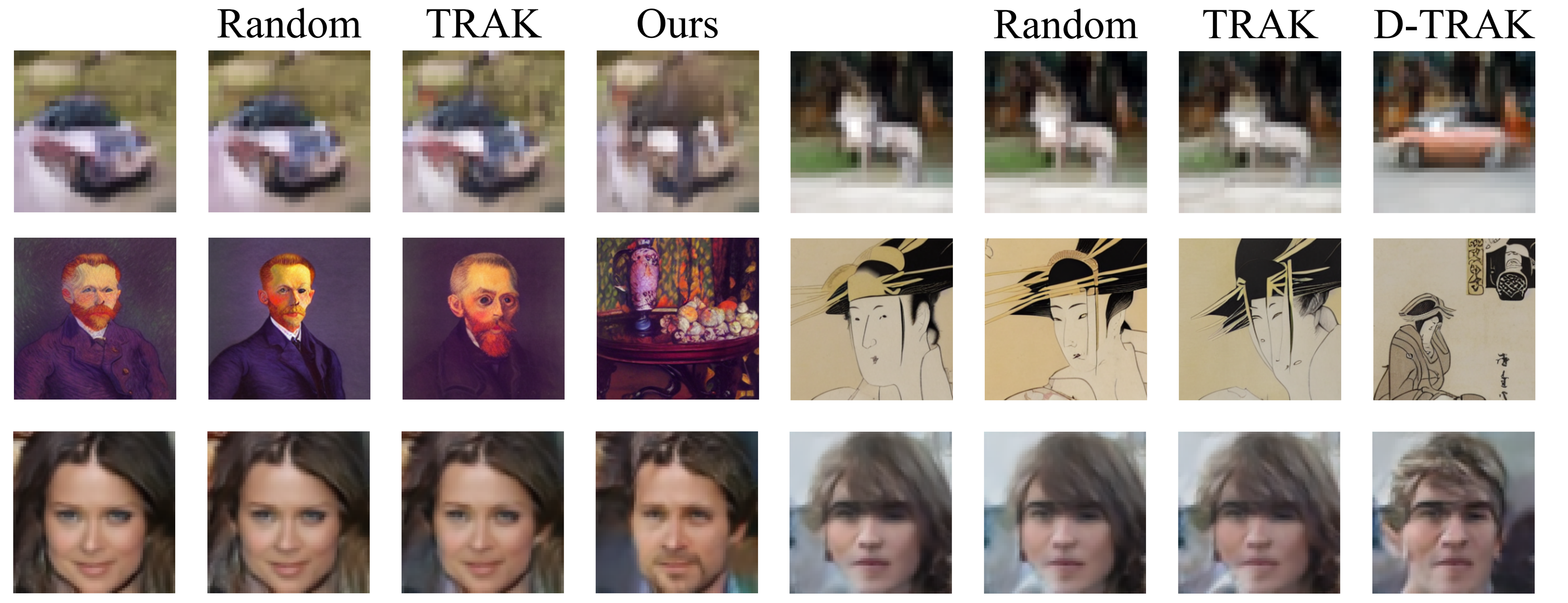
• LDS Evaluation

Method	Results on CIFAR-2: Validation 10	Validation 100	Generation 10	Generation 100
Raw pixel (dot prod.)	7.77 ± 0.57		4.89 ± 0.58	
Raw pixel (cosine)	7.87 ± 0.57		5.44 ± 0.57	
CLIP similarity (dot prod.)	6.51 ± 1.06		3.00 ± 0.95	
CLIP similarity (cosine)	8.54 ± 1.01		4.01 ± 0.85	
Gradient (dot prod.) (Charpiat et al., 2019)	5.14 ± 0.60	5.07 ± 0.55	2.80 ± 0.55	4.03 ± 0.51
Gradient (cosine) (Charpiat et al., 2019)	5.08 ± 0.59	4.89 ± 0.50	2.78 ± 0.54	3.92 ± 0.49
TracInCP (Pruthi et al., 2020)	6.26 ± 0.84	5.47 ± 0.87	3.76 ± 0.61	3.70 ± 0.66
GAS (Hammoudeh & Lowd, 2022a)	5.78 ± 0.82	5.15 ± 0.87	3.34 ± 0.56	3.30 ± 0.68
Journey TRAK (Georgiev et al., 2023)	/	/	7.73 ± 0.65	12.21 ± 0.46
Relative IF† (Barshan et al., 2020)	11.20 ± 0.51	23.43 ± 0.46	5.86 ± 0.48	15.91 ± 0.39
Renorm. IF† (Hammoudeh & Lowd, 2022a)	10.89 ± 0.46	21.46 ± 0.42	5.69 ± 0.45	14.65 ± 0.37
TRAK (Park et al., 2023)	11.42 ± 0.49	23.59 ± 0.46	5.78 ± 0.48	15.87 ± 0.39
D-TRAK (**Ours**)	**26.79 ± 0.33**	**33.74 ± 0.37**	**18.82 ± 0.43**	**25.67 ± 0.40**

• Counterfactual Evaluation



(a) CIFAR-2 (b) ArtBench-2 (c) CIFAR-2 (d) ArtBench-2



Find more interesting conclusions in our paper!