Classifier-free Diffusion Guidance with a Robust Energy-Based Classifier

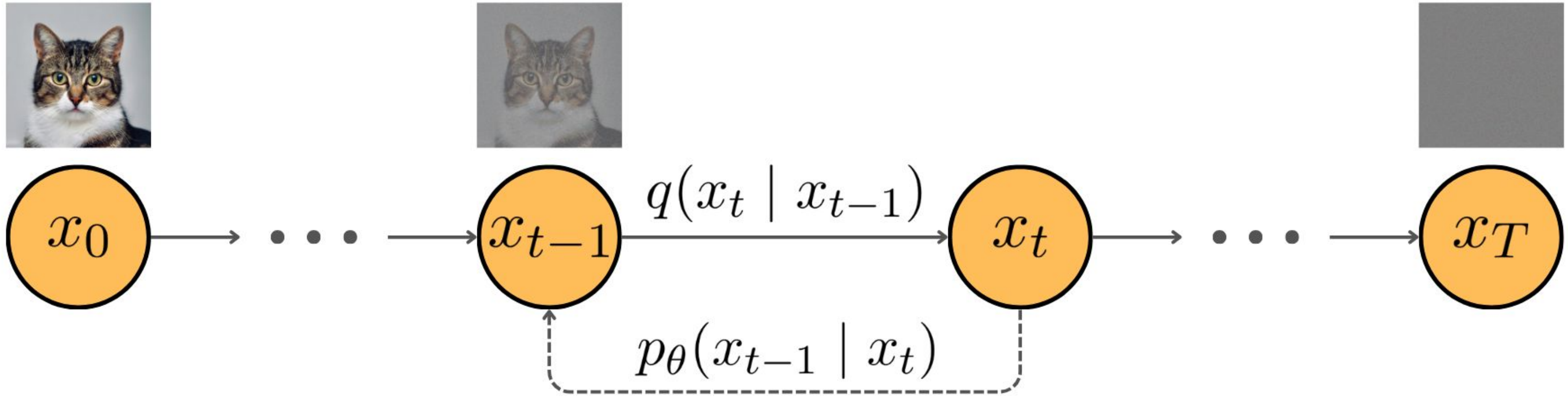
Advisor
Prof. Iacopo Masi

Filippo Wang
Computer Science Department
OmnAI Lab, Sapienza, University of Rome, Italy
https://omnai.di.uniroma1.it

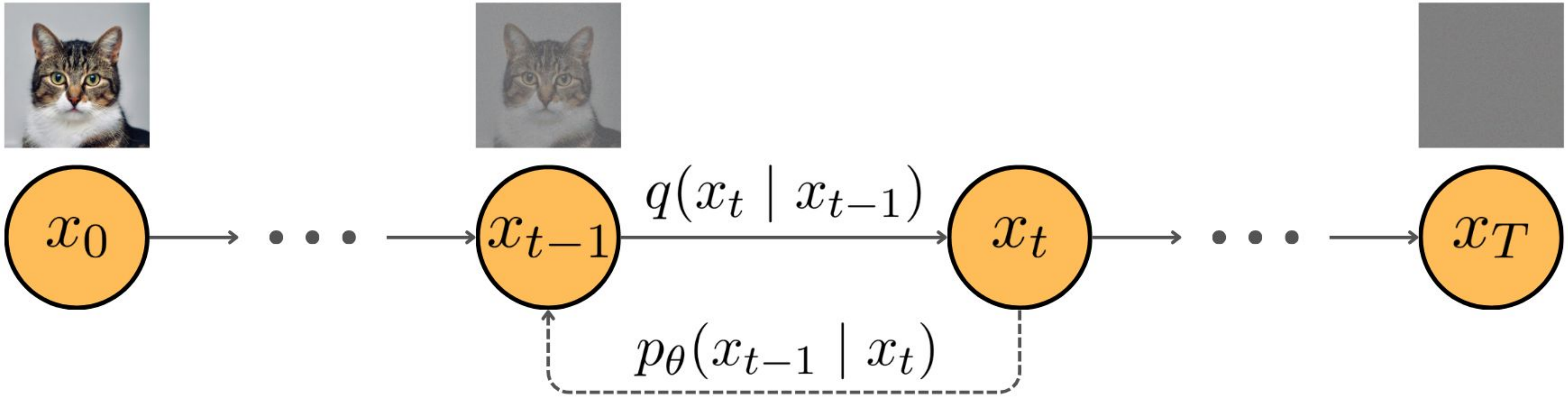
Co-Advisor
Dr. Maria Rosaria Briglia

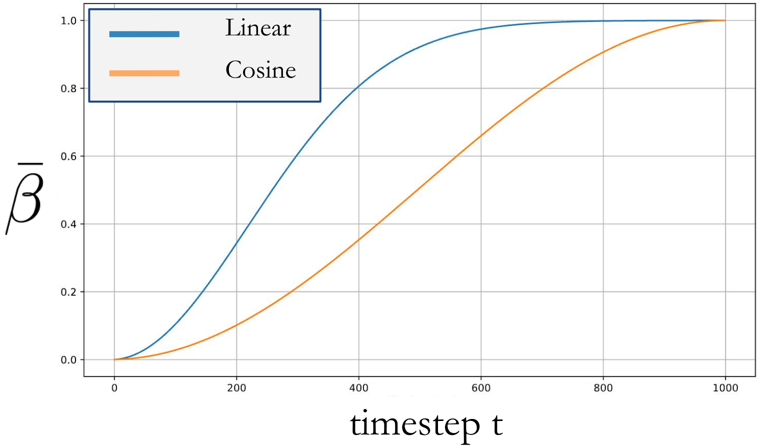
Diffusion Models and Classifier-guidance

Diffusion Models



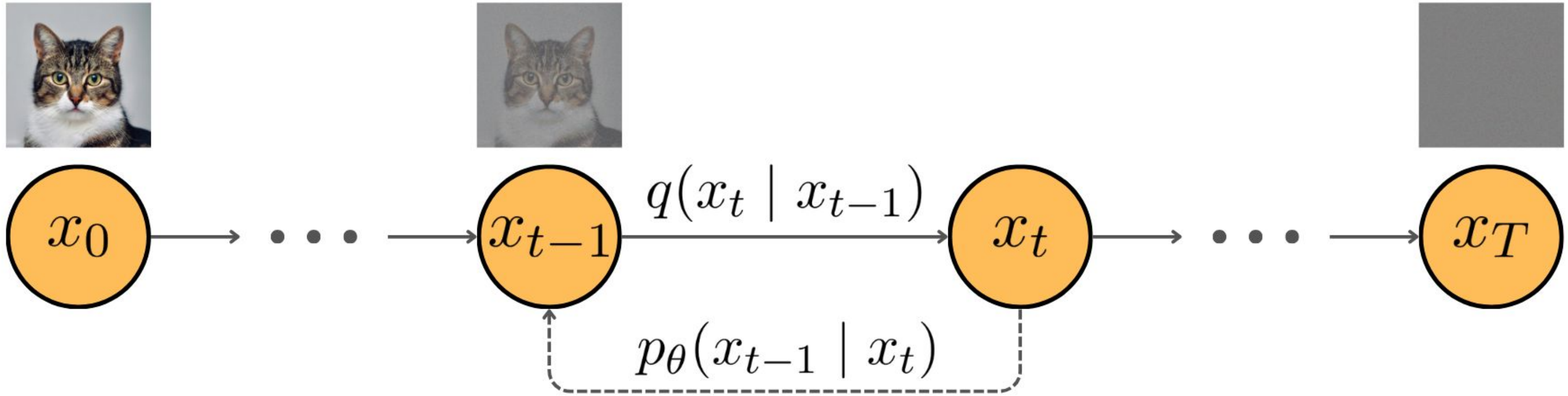
Diffusion Models

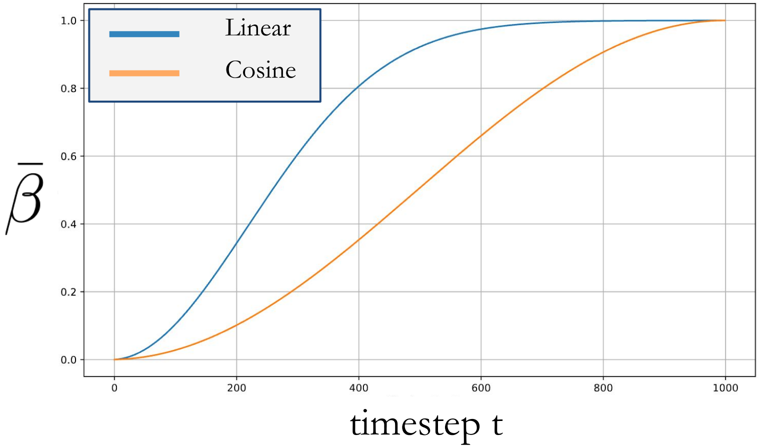




$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1 - \bar{\beta}_t}\; x_0,\ \bar{\beta}_t \mathbf{I}\right)$$

Diffusion Models





$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\bar{\beta}_t}\, x_0,\ \bar{\beta}_t \mathbf{I}\right)$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t),\ \sigma^2\right)$$

Classifier Guidance

for all t *from* T *to* 1 **do**

$\quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;

$\quad \mu \leftarrow \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right)$

$\quad \mathbf{x}_{t-1} \leftarrow \mu + \sigma_t \epsilon$;

return $\mathbf{x}_0$

Classifier Guidance

for all t *from* T *to* 1 **do**

$\quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;

$\quad \mu \leftarrow \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right) + s\sigma_t \nabla_{\mathbf{x}_t} \log p_\phi(y \mid \mathbf{x}_t)$

$\quad \mathbf{x}_{t-1} \leftarrow \mu + \sigma_t \epsilon$;

return $\mathbf{x}_0$

Training is expensive!

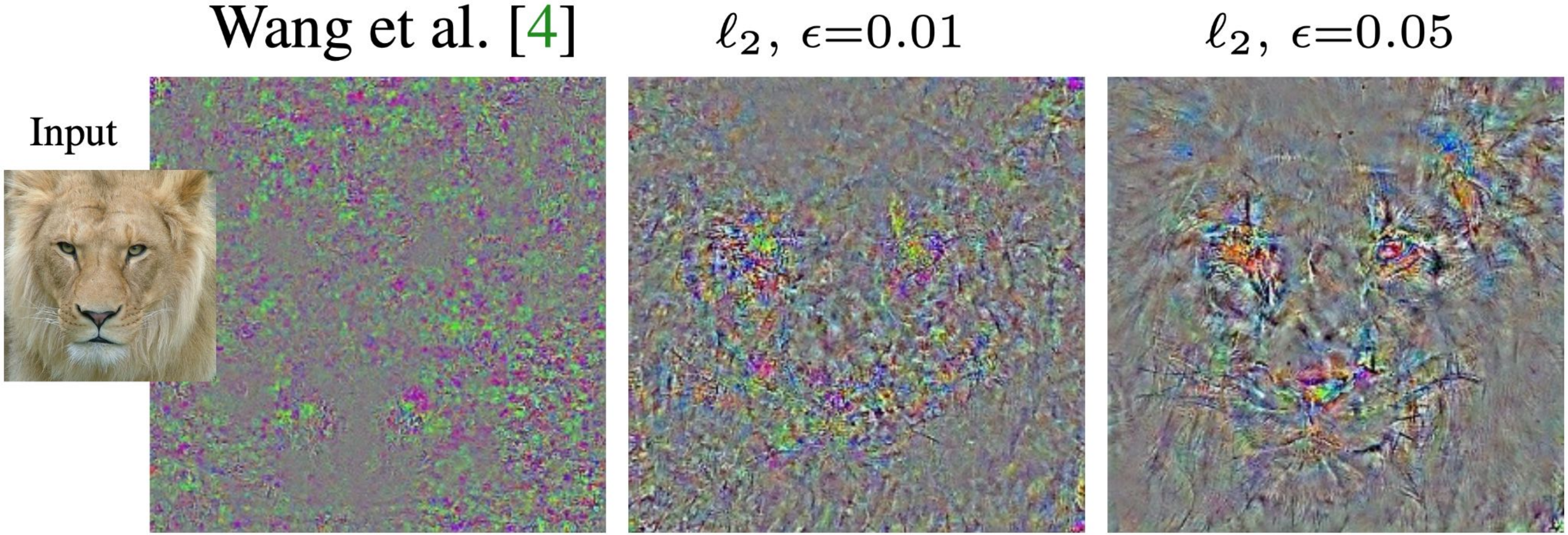
Why Robust Classifiers behave as Generative Models?

Robust Gradients are Perceptually Aligned

$$\nabla_{\mathbf{x}} \mathcal{L}_{\mathrm{CE}(\mathbf{x},\, y;\, \phi)}$$

Non-Robust

Robust



Generating images with a single robust classifier

[Image Synthesis with a Single (Robust) Classifier, NIPS19]

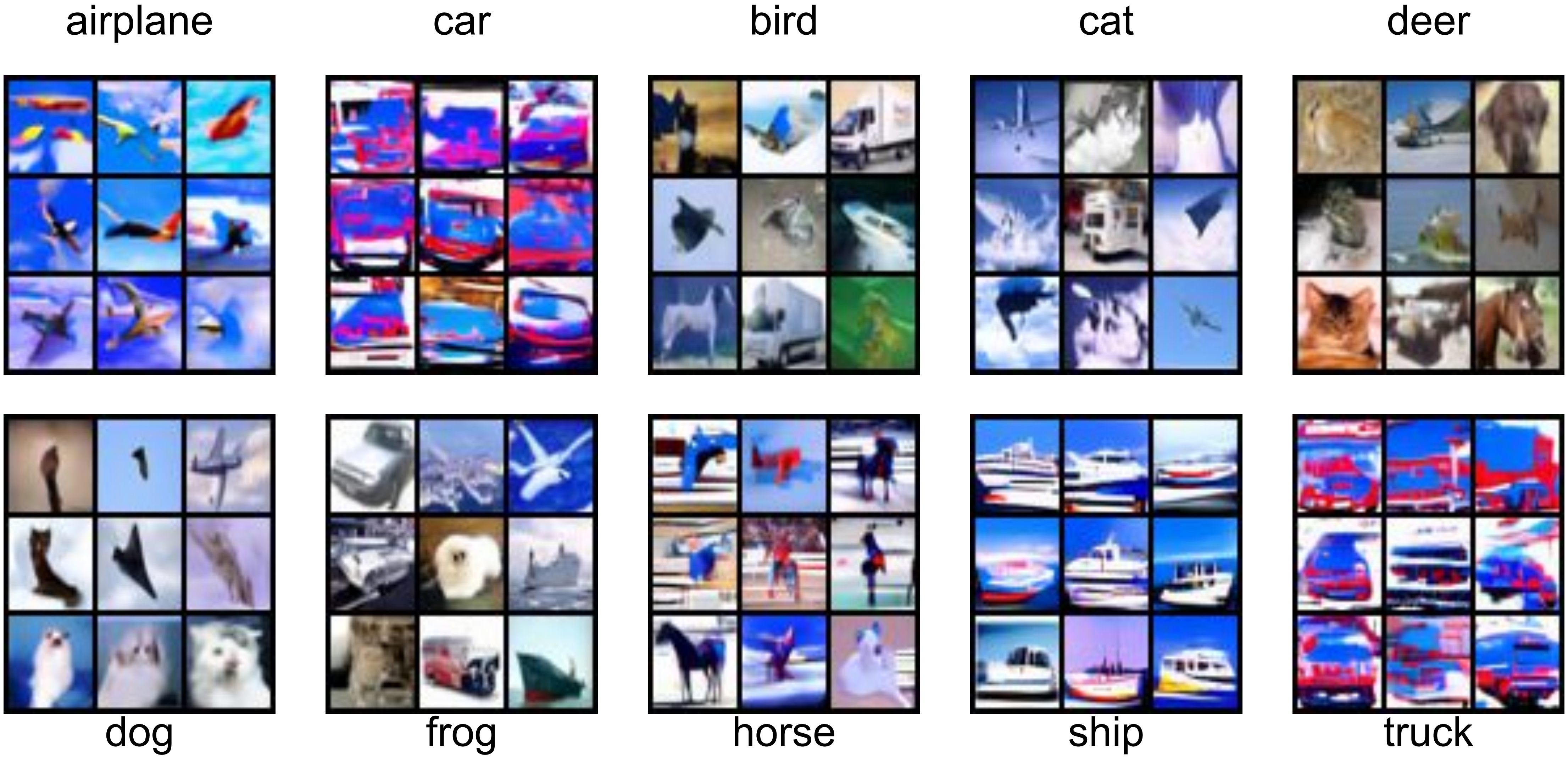
Classifier-free Diffusion Guidance with a Robust Energy-Based Classifier

Robust Classifier Guidance

Naïvely replacing the classifier

airplane

Naïvely replacing the classifier

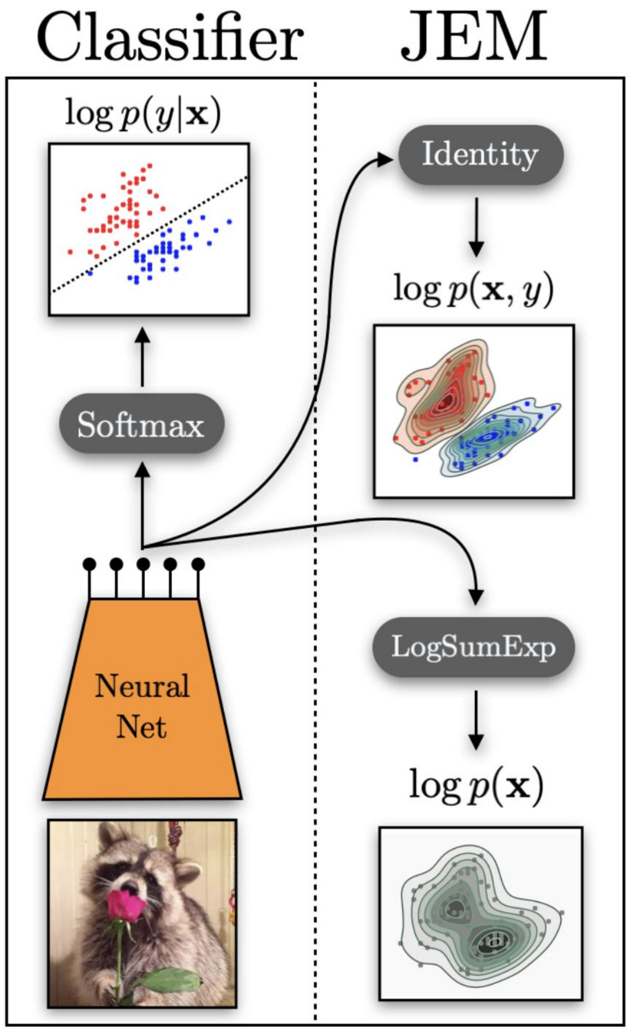


robust model trained on radius ε = 1.0 -- s = 4

Robust Classifier-free guidance

Your classifier is secretly an Energy-Based Model

$$p_\phi(y \mid \mathbf{x}) = \frac{p_\phi(\mathbf{x}, y)}{p_\phi(\mathbf{x})} = \frac{\exp(F_\phi(\mathbf{x})[y])}{\sum_k \exp(F_\phi(\mathbf{x})[y])}$$

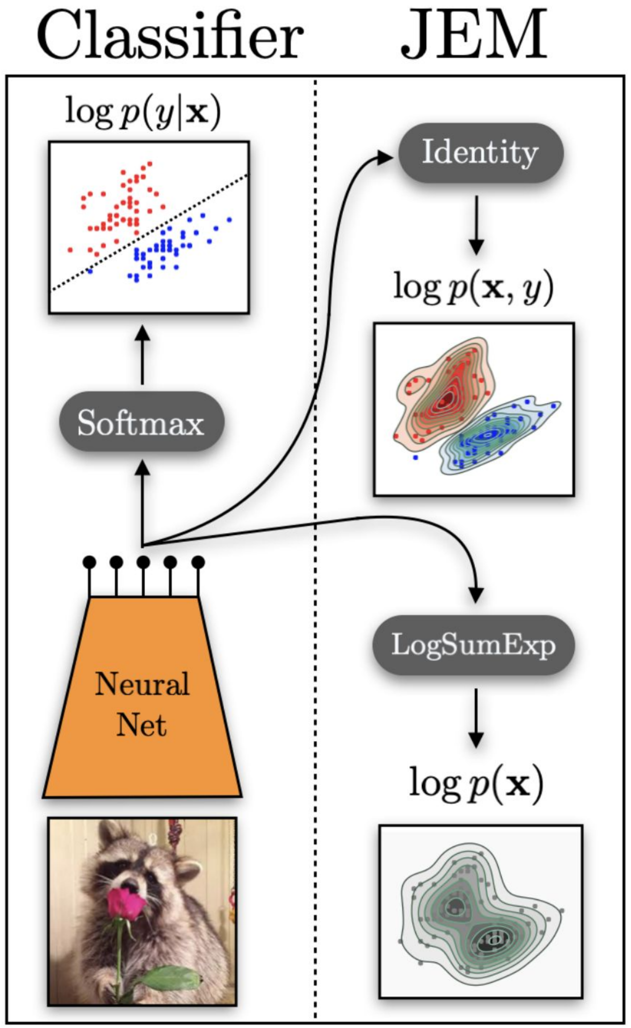


[Your classifier is secretly an energy based model and you should treat it like one, ICLR20]

Your classifier is secretly an Energy-Based Model

$$p_\phi(y \mid \mathbf{x}) = \frac{p_\phi(\mathbf{x}, y)}{p_\phi(\mathbf{x})} = \frac{\exp(F_\phi(\mathbf{x})[y])}{\sum_k \exp(F_\phi(\mathbf{x})[y])}$$

$$\log p_\phi(y \mid \mathbf{x}) = \log p_\phi(\mathbf{x}, y) - \log p_\phi(\mathbf{x})$$



[Your classifier is secretly an energy based model and you should treat it like one, ICLR20]

Your classifier is secretly an Energy-Based Model

$$p_\phi(y \mid \mathbf{x}) = \frac{p_\phi(\mathbf{x}, y)}{p_\phi(\mathbf{x})} = \frac{\exp(F_\phi(\mathbf{x})[y])}{\sum_k \exp(F_\phi(\mathbf{x})[y])}$$

$$\log p_\phi(y \mid \mathbf{x}) = \log p_\phi(\mathbf{x}, y) - \log p_\phi(\mathbf{x})$$

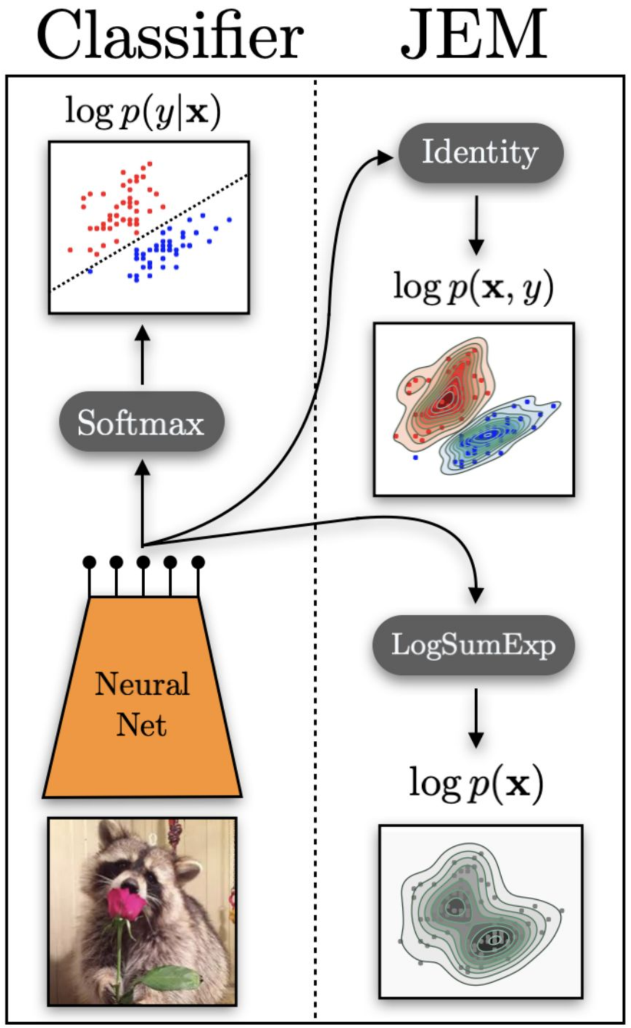
$$= F_\phi(\mathbf{x})[y] - \mathrm{LogSumExp}_y(F_\phi(\mathbf{x})[y])$$

$$= E_\phi(\mathbf{x}) - E_\phi(\mathbf{x}, y)$$

where

$$\mathrm{E}_\phi(\mathbf{x}) = -\log p_\phi(\mathbf{x})$$

$$\mathrm{E}_\phi(\mathbf{x}, y) = -\log p_\phi(\mathbf{x}, y)$$



[Your classifier is secretly an energy based model and you should treat it like one, ICLR20]

Classifier Guidance

for all t *from* T *to* 1 **do**

$\quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;

$\quad \mu \leftarrow \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right) + s\sigma_t \nabla_{\mathbf{x}_t} \log p_\phi(y \mid \mathbf{x}_t)$

$\quad \mathbf{x}_{t-1} \leftarrow \mu + \sigma_t \epsilon$;

return $\mathbf{x}_0$

Classifier Guidance

$$\boxed{-\; s\hat{\sigma}_t \nabla_{\mathbf{x}_t} E_\phi(\mathbf{x}, y)}$$

for all t *from* T *to* 1 **do**

$\quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;

$\quad \mu \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right) \cancel{+ s\sigma_t \nabla_{\mathbf{x}_t} \log p_\phi(y \mid \mathbf{x}_t)}$

$\quad \mathbf{x}_{t-1} \leftarrow \mu + \sigma_t \epsilon$;

return $\mathbf{x}_0$

Classifier Guidance

for all t *from* T *to* 1 **do**

  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;

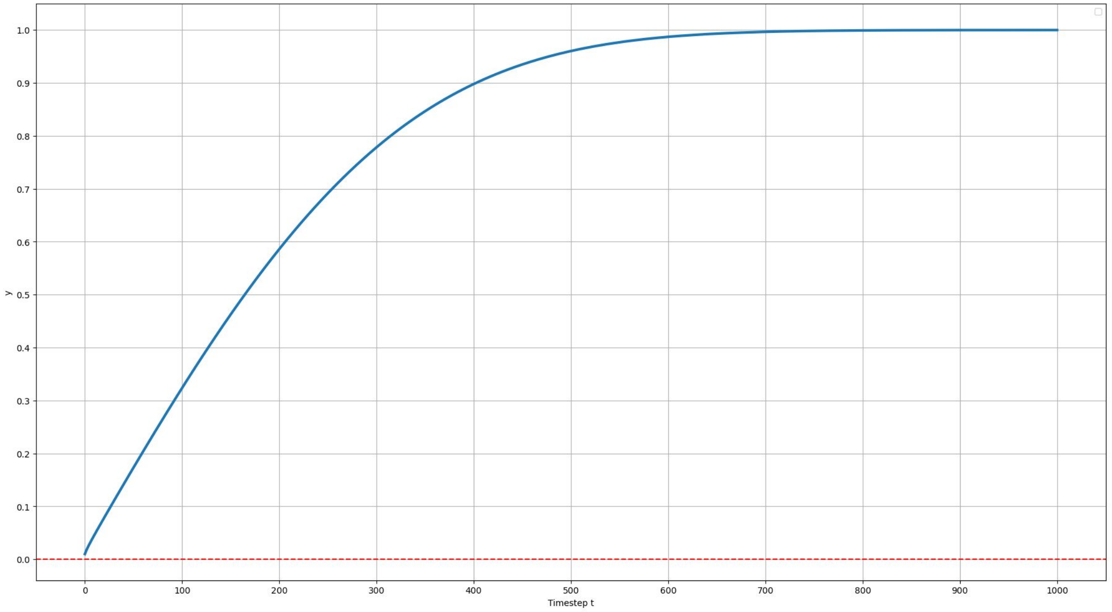
  $\mu \leftarrow \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right) \boxed{- \; s \boxed{\hat{\sigma}_t} \nabla_{\mathbf{x}_t} E_\phi(\mathbf{x}, y)}$

  $\mathbf{x}_{t-1} \leftarrow \mu + \sigma_t \epsilon$;

return $\mathbf{x}_0$

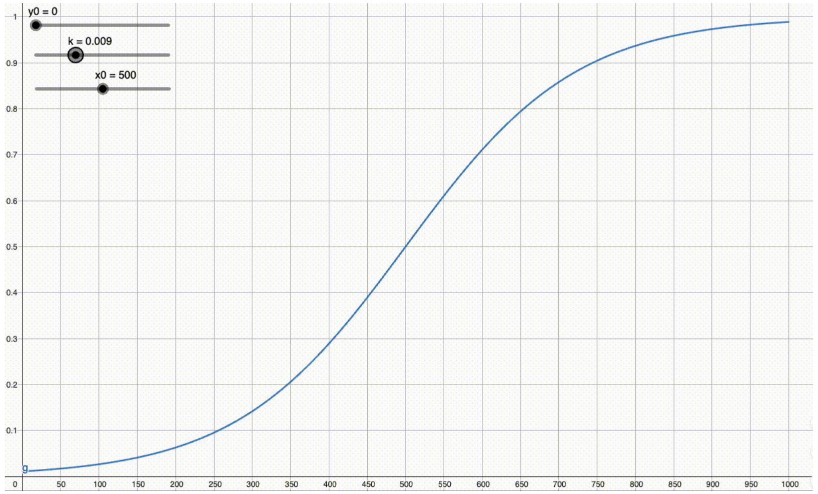
Standard guidance scheduler

$$\hat{\sigma}_t = \sqrt{1 - \bar{\alpha}_t}$$

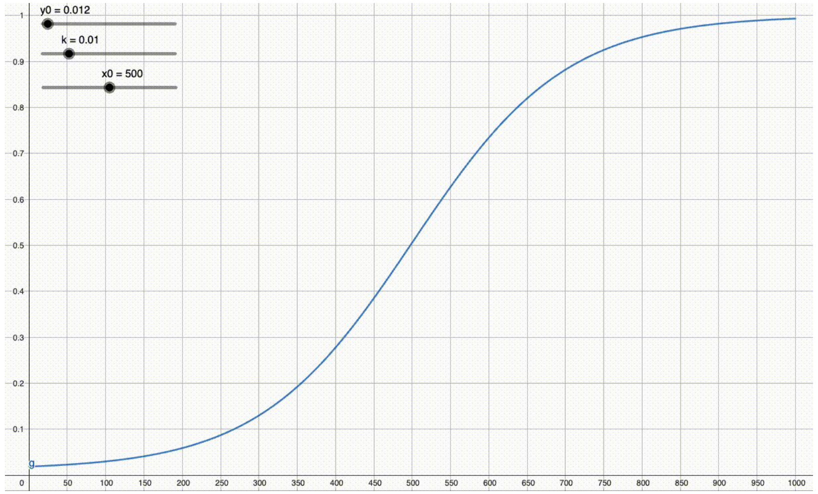


Custom guidance scheduler

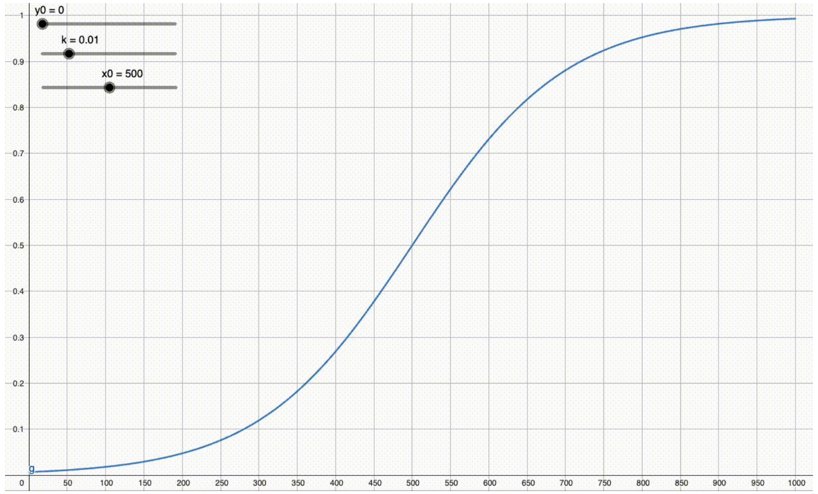
$$\hat{\sigma}_t = (1 - y_0) \cdot \frac{1}{1 + e^{-k(t - t_0)}} + y_0$$



k

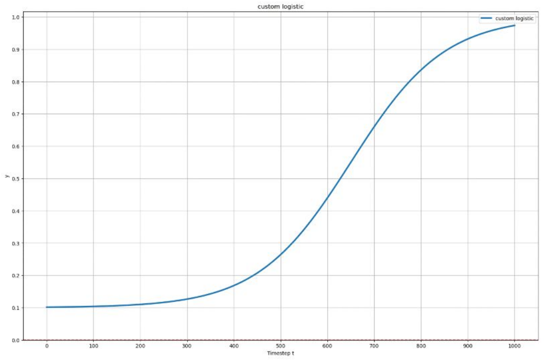


y_0



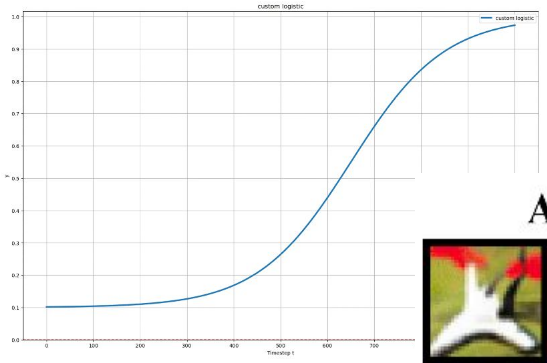
t_0

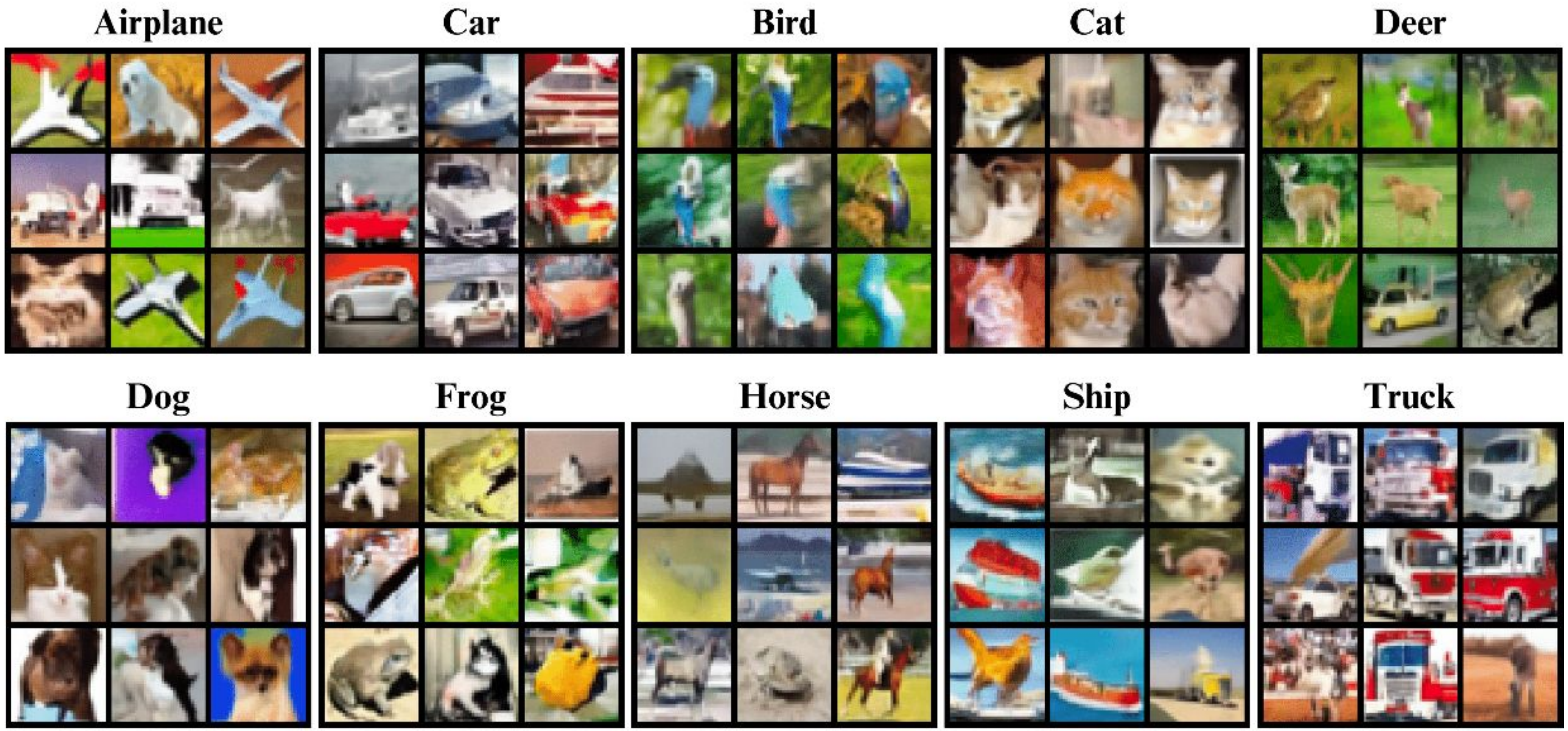
Joint Energy Guidance



Guidance scheduler parameters: t0 = 650, y0 = 0.1, k = 0.01 -- model trained on radius ε = 1.0 -- s = 12

Joint Energy Guidance





Guidance scheduler parameters: t0 = 650, y0 = 0.1, k = 0.01 -- model trained on radius ε = 1.0 -- s = 12

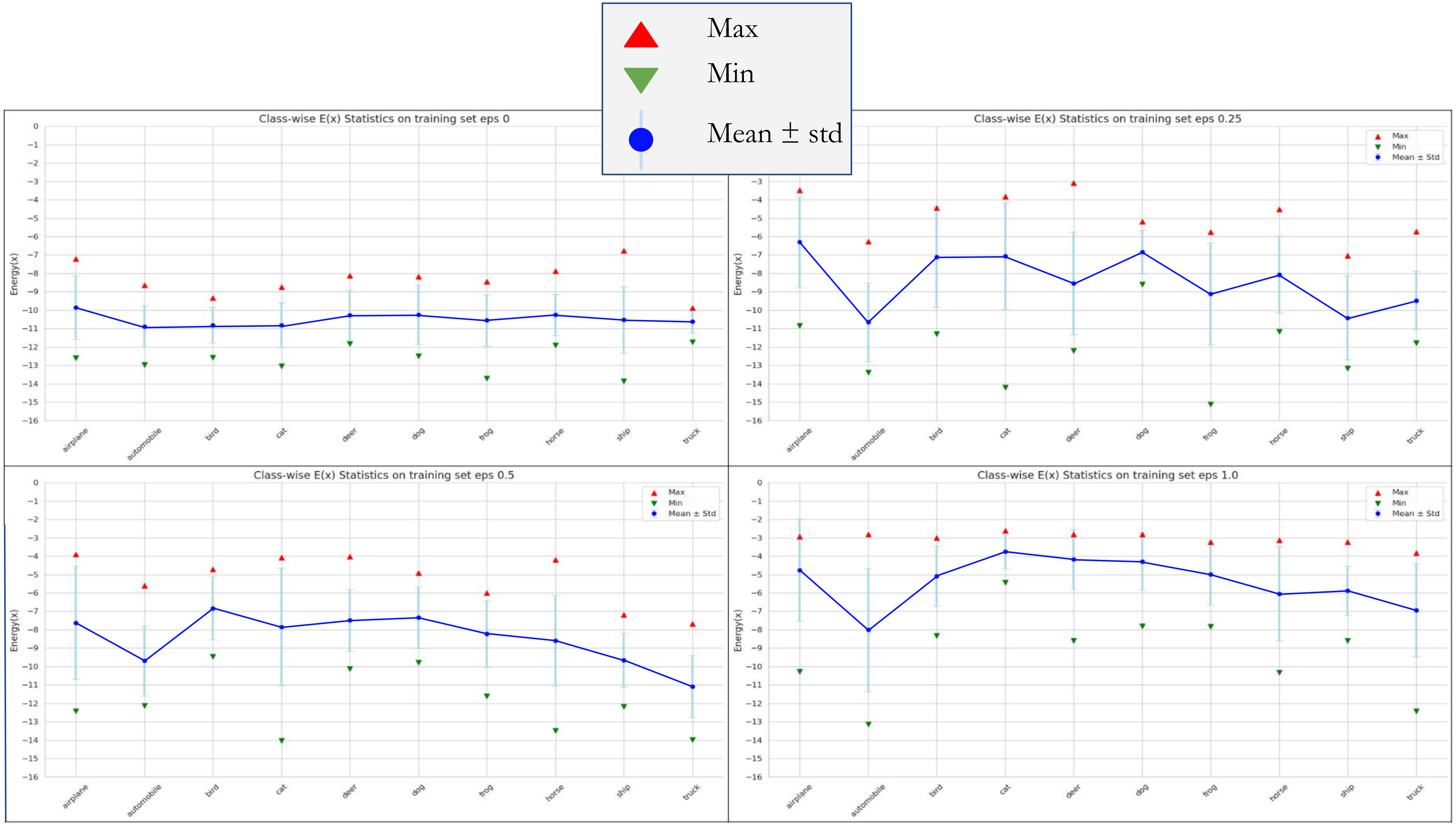
Classifier-free Diffusion Guidance with a Robust Energy-Based Classifier

An analysis on training set Energy

E(x) statistics by Class CIFAR-10 training set

$\varepsilon = 0/255$

$\varepsilon = 0.25/255$



$\varepsilon = 0.5/255$

$\varepsilon = 1.0/255$

Bring back E(x) for guidance

$$\Delta E_\phi(\mathbf{x}) = |E_\phi(\mathbf{x}^*)[y] - E_\phi(\mathbf{x})|$$

Bring back E(x) for guidance

$$\Delta E_\phi(\mathbf{x}) = |E_\phi(\mathbf{x}^*)[y] - E_\phi(\mathbf{x})|$$

$$\mathcal{L} := E_\phi(\mathbf{x}, y) + \lambda \Delta E_\phi(\mathbf{x})$$

Bring back E(x) for guidance

$$\Delta E_\phi(\mathbf{x}) = |E_\phi(\mathbf{x}^*)[y] - E_\phi(\mathbf{x})|$$

$$\mathcal{L} := E_\phi(\mathbf{x}, y) + \lambda \Delta E_\phi(\mathbf{x})$$

$$\boxed{s\hat{\sigma}_t \nabla_{\mathbf{x}_t} \mathcal{L}}$$

for all t *from* T *to* 1 **do**

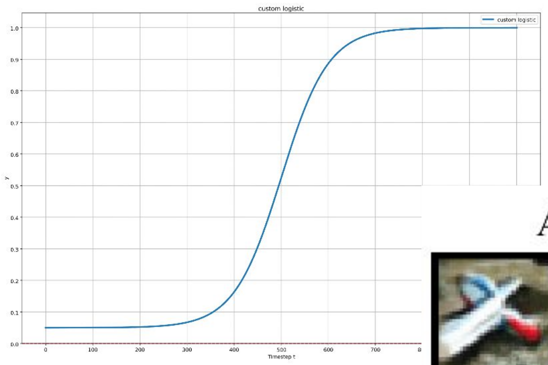
$\quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;

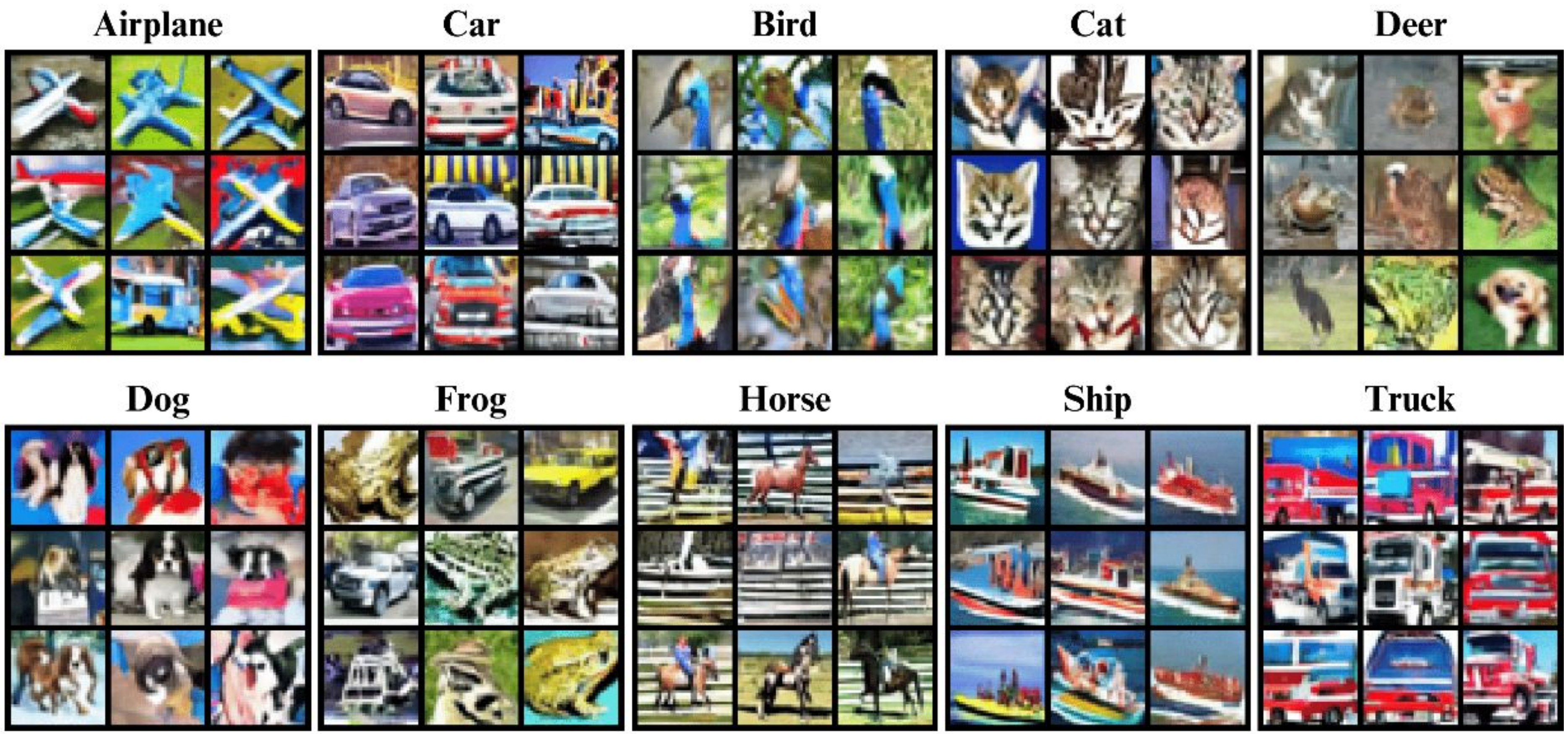
$\quad \mu \leftarrow \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right) - \cancel{s\hat{\sigma}_t \nabla_{\mathbf{x}_t} E_\phi(\mathbf{x}, y)}$

$\quad \mathbf{x}_{t-1} \leftarrow \mu + \sigma_t \epsilon$;

return $\mathbf{x}_0$

Energy is more important than you think

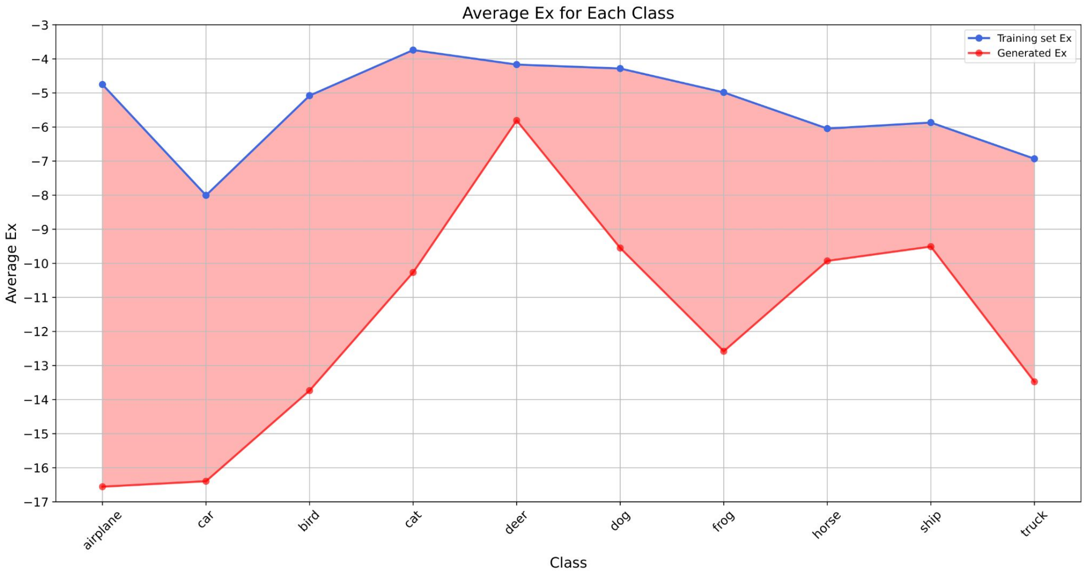




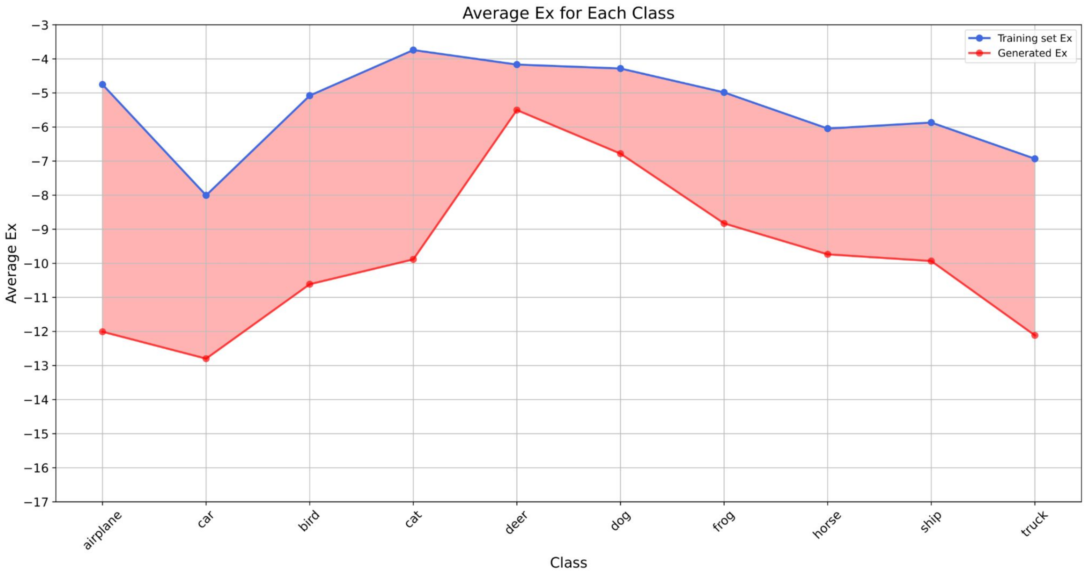
Guidance scheduler parameters: t0 = 500, y0 = 0.05, k = 0.02 -- model trained on radius ε = 1.0 -- s = 4 -- λ = 0.15

Reducing the Energy Divergence

$$\nabla_{\mathbf{x}_t} E_\phi(\mathbf{x}, y)$$



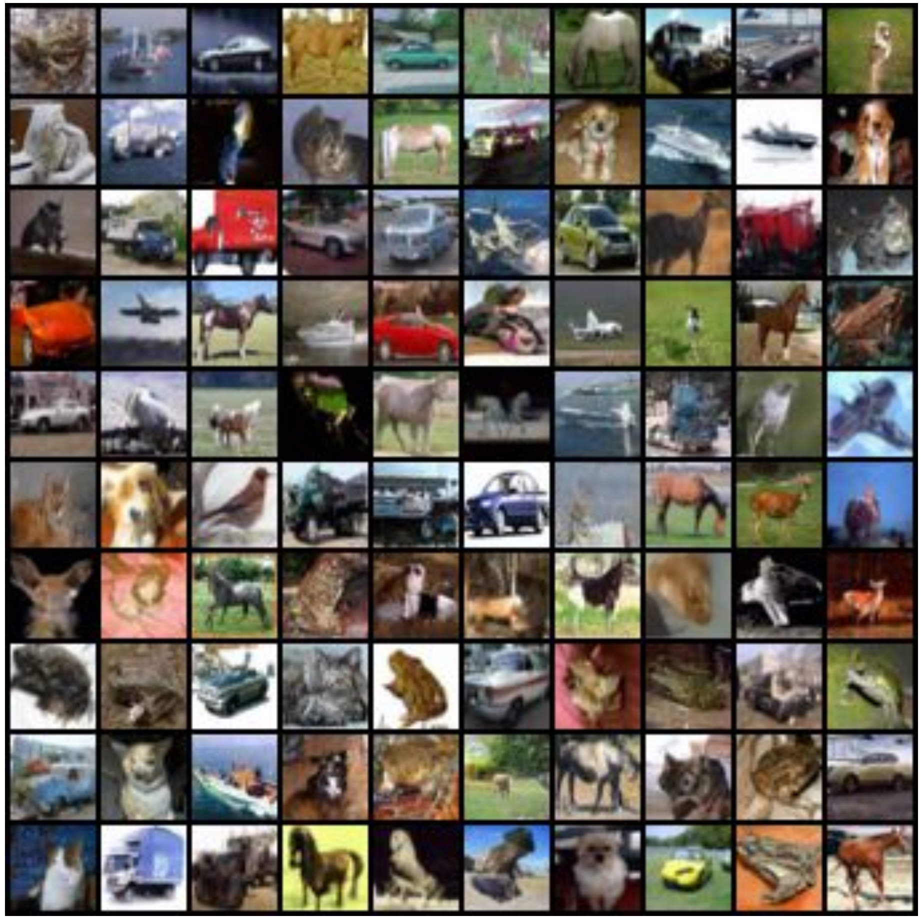
$$\nabla_{\mathbf{x}_t} \mathcal{L}$$



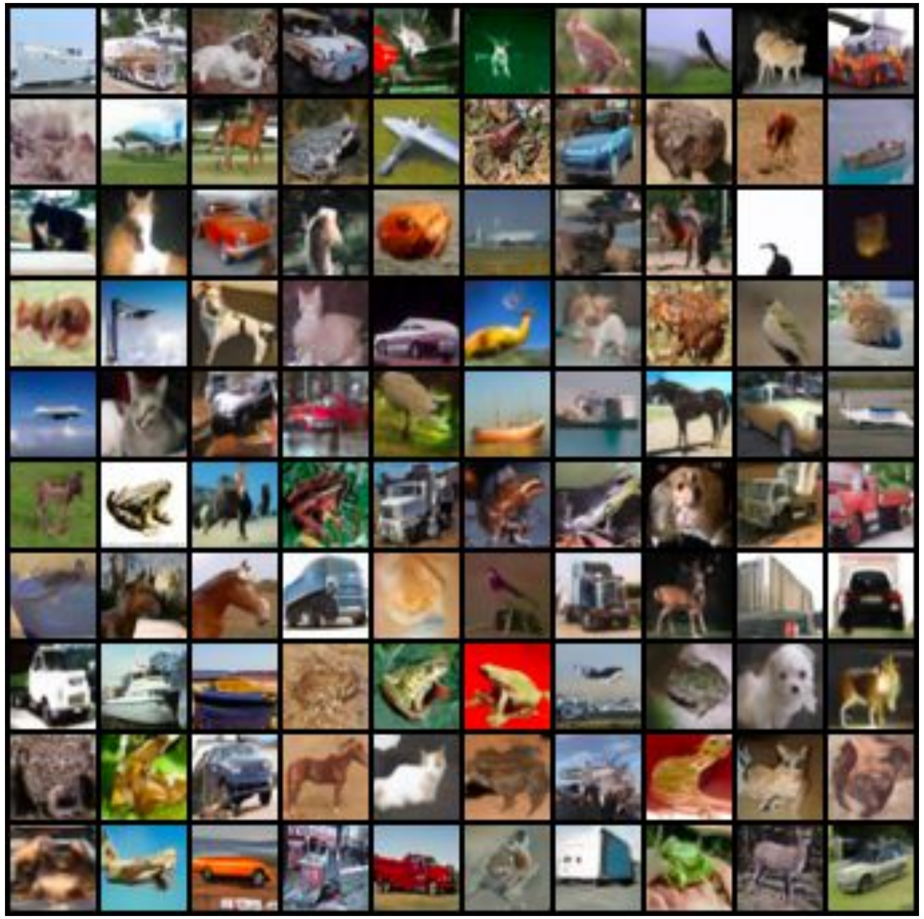
Thanks for listening!

Do you have any questions?

Unconditional Model Quality



DDPM



DDIM

Diffusion Model Evaluation

	IS ↑	FID ↓	Precision ↑	Recall ↑
DDPM	8.68	16.49	0.68	0.59
DDIM	8.40	14.79	0.64	0.59

Table 4.2. Evaluation metrics for unconditional guidance using DDPM and DDIM on CIFAR-10 test set, with 1000 and 100 inference steps respectively

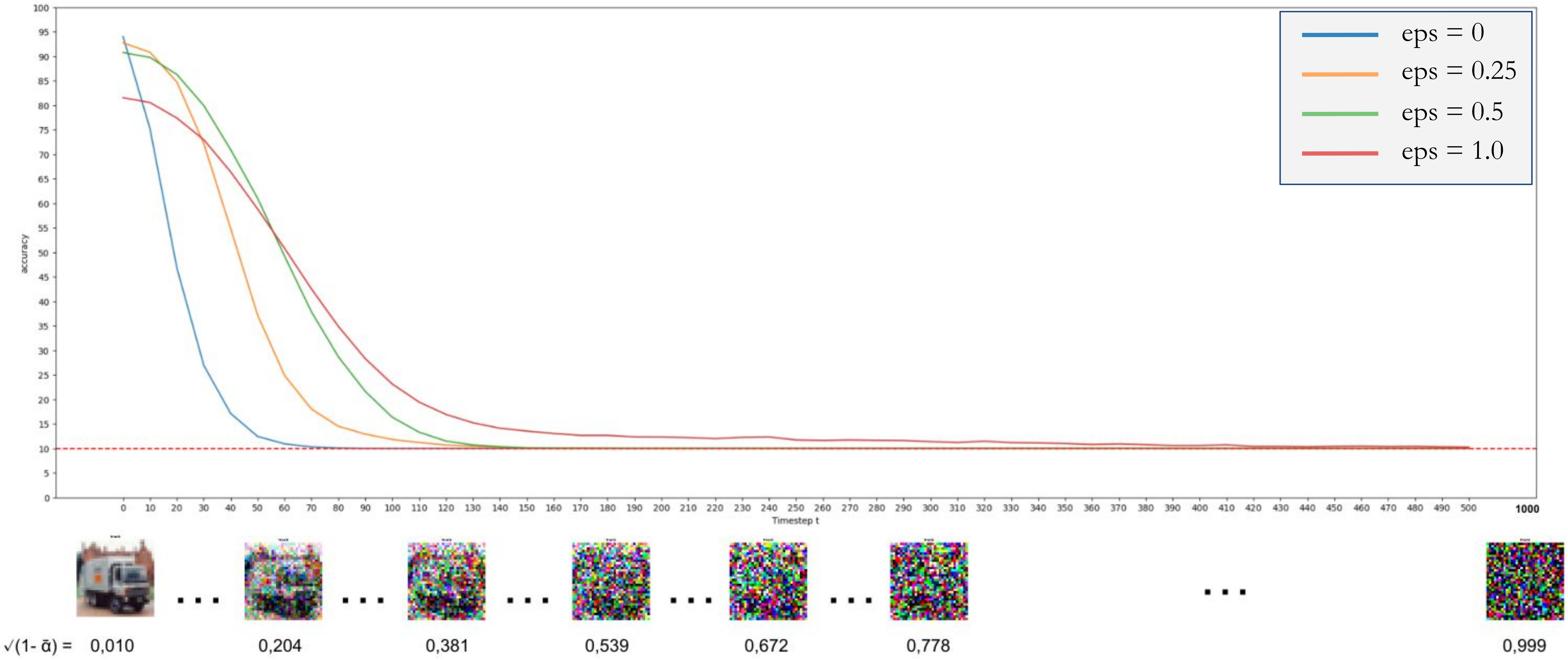
Classifier Evaluation

ε-test \ ε-train	0.0	0.25	0.5	1.0
0.0	**95.25%**	92.77%	90.83%	81.62%
0.25	8.65%	81.21%	**82.40%**	75.53%
0.5	0.28%	62.29%	**70.17%**	68.63%
1.0	0.00%	21.18%	40.48%	**52.72%**
2.0	0.00%	0.53%	5.22%	18.59%

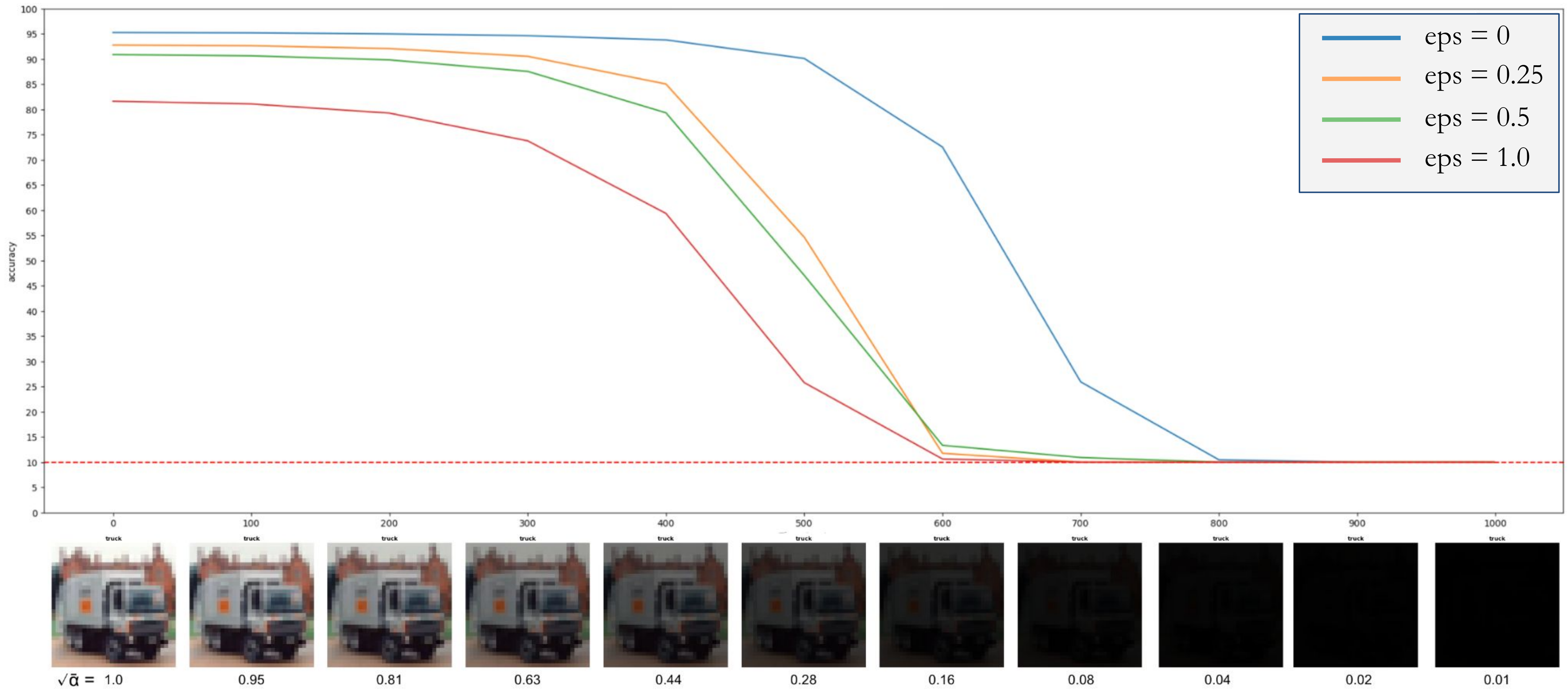
Table 4.1. CIFAR10 L2-norm (ResNet50) 20-steps pgd

test-accuracies on 20-steps PGD-attacks with step size $= 2.5 * \frac{\varepsilon_{test}}{20}$

Classifier Accuracy - Gaussian Noise



Classifier Accuracy - Darkness



Average E(x, y) and E(x)

eps = 0
eps = 0.25
eps = 0.5
eps = 1.0