Diffusion Models and Thermodynamic Singularity

Particle Physics, Quantum Information Science, and Machine Learning 2026

July 13 (Mon.), 2026

Yung-Kyun Noh

Hanyang University & Korea Institute for Advanced Study

Contents

Motivation (Free Energy)

$$P(X = \mathrm{H}) = 0.9\,, \quad P(X = \mathrm{T}) = 0.1$$

H : 

T : 

$$P(\underbrace{\mathrm{HHHHHHHHHH}}_{10}) = 0.9^{10}$$

$$P(\underbrace{\mathrm{HHHHHHTHHH}}_{10}) = 0.9^{9} \cdot 0.1$$

Among the total N of samples,

Let $P(X = a) > P(X = b)$

$X_i \in \{a, b\}$

Number of a samples:

$$N_a = P(a)N$$

Number of b samples:

$$N_b = N - N_a = P(b)N$$

$$\prod_{i=1}^{N} P(X_i = a) > \prod_{i=1}^{N_a} P(X_i = a) \cdot \prod_{j=N_a+1}^{N} P(X_j = b)$$

Why do we always see N_1 number of as and N_2 number of bs?

- Consider $N = 2 \qquad X_1, X_2 \in \{a, b\}$

- Four assignments cover all possibilities:

$$X_1 = a, X_2 = a \qquad P(X_1, X_2) = P(a)^2 \qquad \Longleftarrow \quad P(a)^2$$

$$X_1 = a, X_2 = b \qquad P(X_1, X_2) = P(a)P(b)$$

$$X_1 = b, X_2 = a \qquad P(X_1, X_2) = P(a)P(b)$$

$$\Longleftarrow \quad 2P(a)P(b)$$

$$X_1 = b, X_2 = b \qquad P(X_1, X_2) = P(b)^2 \qquad \Longleftarrow \quad P(b)^2$$

Why do we always see N_1 number of as and N_2 number of bs?

- With N trials, $X_1, X_2, \ldots, X_N \in \{a, b\}$

- Probability of having N_a number of as and N_b number of bs is

$$P(N_a, N_b) = \left(\begin{array}{c} N \\ N_a, N_b \end{array} \right) P(a)^{N_a} P(b)^{N_b}$$

$$\left(\begin{array}{c} N \\ N_a, N_b \end{array} \right) = \left(\begin{array}{c} N \\ N_a \end{array} \right) = \frac{N!}{N_a! N_b!}$$

Why do we always see N_1 number of as and N_2 number of bs?

- With N trials, $X_1, X_2, \ldots, X_N \in \{a, b\}$

- Probability of having N_a number of as and N_b number of bs is

$$P(N_a, N_b) = \binom{N}{N_a, N_b} P(a)^{N_a} P(b)^{N_b}$$

$$\binom{N}{N_1, N_2} \gg \binom{N}{N, 0}$$

$$P(a)^{N_1} P(b)^{N_2} < P(a)^N$$

Now let's derive the equation like this!

$$P(N_a, N_b) = \binom{N}{N_a, N_b} P(a)^{N_a} P(b)^{N_b}$$

Stirling's approximation for large n,

$$\log n! \approx n \log n - n$$

$$\Rightarrow \quad \log \binom{N}{N_a, N_b} = \log \frac{N!}{N_a! N_b!}$$

$$\approx N \log N - \cancel{N} - \{N_a \log N_a - \cancel{N_a}\} - \{N_b \log N_b - \cancel{N_b}\}$$

$$= N \cdot \left\{ -\frac{N_a}{N} \log \frac{N_a}{N} - \frac{N_b}{N} \log \frac{N_b}{N} \right\}$$

Now let's derive the equation like this!

$$\log \binom{N}{N_a, N_b} \approx N \cdot \left\{ -\frac{N_a}{N} \log \frac{N_a}{N} - \frac{N_b}{N} \log \frac{N_b}{N} \right\}$$

Let $Q_i = \dfrac{N_i}{N}$

$$\log \binom{N}{N_a, N_b} \approx N \sum_{i \in \{a,b\}} -Q_i \log Q_i$$

$$= N \cdot \mathrm{Ent}[Q_i]$$

Entropy of empirical distribution

Free Energy

- Let

$$P(N_a, N_b) = \exp\Bigg(- \Bigg\{ - N_a \log P(x=a|\lambda) - N_b \log P(x=b|\lambda) \\ -N\left(-\frac{N_a}{N}\log\frac{N_a}{N} - \frac{N_b}{N}\log\frac{N_b}{N}\right)\Bigg\}\Bigg)$$

$$\equiv \exp\left(-\frac{E-TS}{T}\right)$$

Free Energy

$$P = \exp\left(-\frac{E - TS}{T}\right)$$

Data fitting

$$E = -\frac{N_a}{N}\log P(x = a|\lambda) - \frac{N_b}{N}\log P(x = b|\lambda)$$

$$S = N \cdot \left(-\frac{N_a}{N}\log\frac{N_a}{N} - \frac{N_b}{N}\log\frac{N_b}{N}\right)$$

$$T = \frac{1}{N}$$

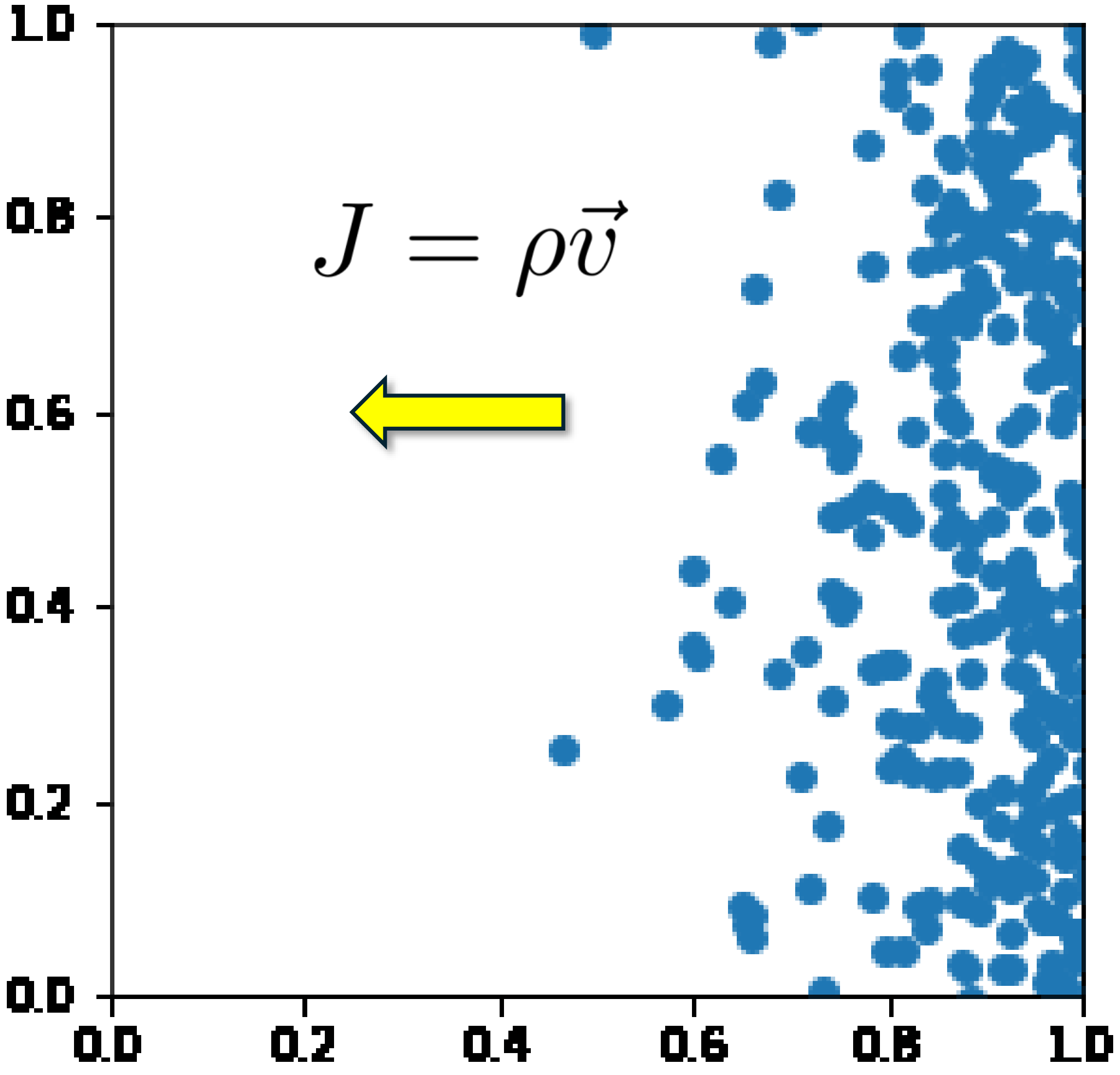
Temperature

$$\frac{1}{T} = \frac{dS}{dE}$$

$$T = \frac{1}{N} \frac{\log P(a) - \log P(b)}{\log Q_a - \log Q_b}$$

At $T = \dfrac{1}{N}$, the equilibrium occurs at

$$Q_a = P(a), Q_b = P(b)$$



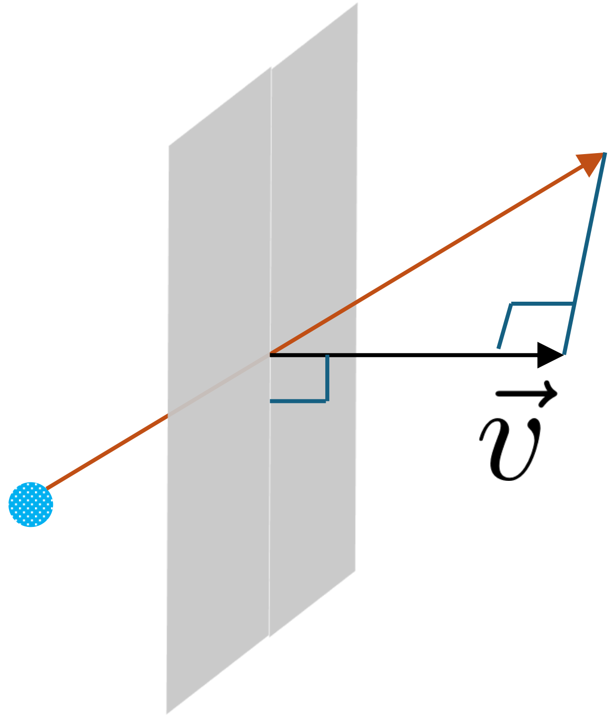
$$J = \rho\vec{v}$$

Flux J: Net number of particles passing through the membrane per unit time

$$J \equiv \rho \vec{v}$$

ρ : mass density

$\vec{v}$: net orthogonal velocity



A velocity field $\vec{v}$ associated with J (flux) is of our interest!

Helmholtz Free Energy (Dynamics with Fixed T)

$$F[\rho_t] = \langle U \rangle_{\rho_t} - TS = \int \rho_t(\mathbf{x}) U(\mathbf{x}) d\mathbf{x} + T \int \rho_t(\mathbf{x}) \log \rho_t(\mathbf{x}) d\mathbf{x}$$

$$\frac{dF}{dt} = \int \left. \frac{\partial F}{\partial \rho_t} \right|_{\rho_t} \frac{\partial \rho_t}{\partial t} d\mathbf{x} = - \int \left. \frac{\partial F}{\partial \rho_t} \right|_{\rho_t} \nabla \cdot (\rho_t \vec{v}) d\mathbf{x}$$

$$= \int \rho_t \vec{v}(\mathbf{x})^\top \nabla \left. \frac{\partial F}{\partial \rho_t} \right|_{\rho_t} d\mathbf{x} = \left\langle \vec{v}, \nabla \left. \frac{\partial F}{\partial \rho_t} \right|_{\rho_t} \right\rangle_{\rho_t}$$

$\longleftarrow$ Maximally decrease

$\vec{v} \propto -\nabla \left. \frac{\partial F}{\partial \rho_t} \right|_{\rho_t}$ with fixed $K = \frac{1}{2} \int \rho_t ||\vec{v}||^2 d\mathbf{x}$

Flow for Diffusion Models

- Flow that maximally decreases free energy

$$\vec{v}(\mathbf{x}) = -\nabla \frac{\partial F}{\partial \rho_t} \;=\; -\nabla U - T\nabla \log \rho_t$$

$$J(\mathbf{x}) = \rho_t(\mathbf{x})\vec{v}(\mathbf{x}) = -\rho_t(\mathbf{x})\nabla U - T\nabla \rho_t$$

- Forward process

Fixed potential energy and temperature $U(\mathbf{x}) = \frac{\beta}{4}\mathbf{x}^2 \;, \quad T = \frac{\beta}{2}$

$$J_F = -\rho_{\mathrm{eq}}\nabla U - T\nabla \rho_{\mathrm{eq}} = 0 \;\longrightarrow\; \rho_{\mathrm{eq}}(\mathbf{x}) \propto \exp(-U(\mathbf{x})/T) = \exp\left(-\mathbf{x}^2/2\right)$$

Forward flux

Flow for Diffusion Models

- Flow that maximally decreases free energy

$$\vec{v}(\mathbf{x}) = -\nabla \frac{\partial F}{\partial \rho_t} \; = \; -\nabla U - T\nabla \log \rho_t$$

$$J(\mathbf{x}) = \rho_t(\mathbf{x})\vec{v}(\mathbf{x}) = -\rho_t(\mathbf{x})\nabla U - T\nabla \rho_t$$

- Reverse process

With learned (time-varying) potential energy,

$$J_R = -J_F$$

Properties of Helmholtz Free Energy Dynamics

[1] Helmholtz free energy does not increase

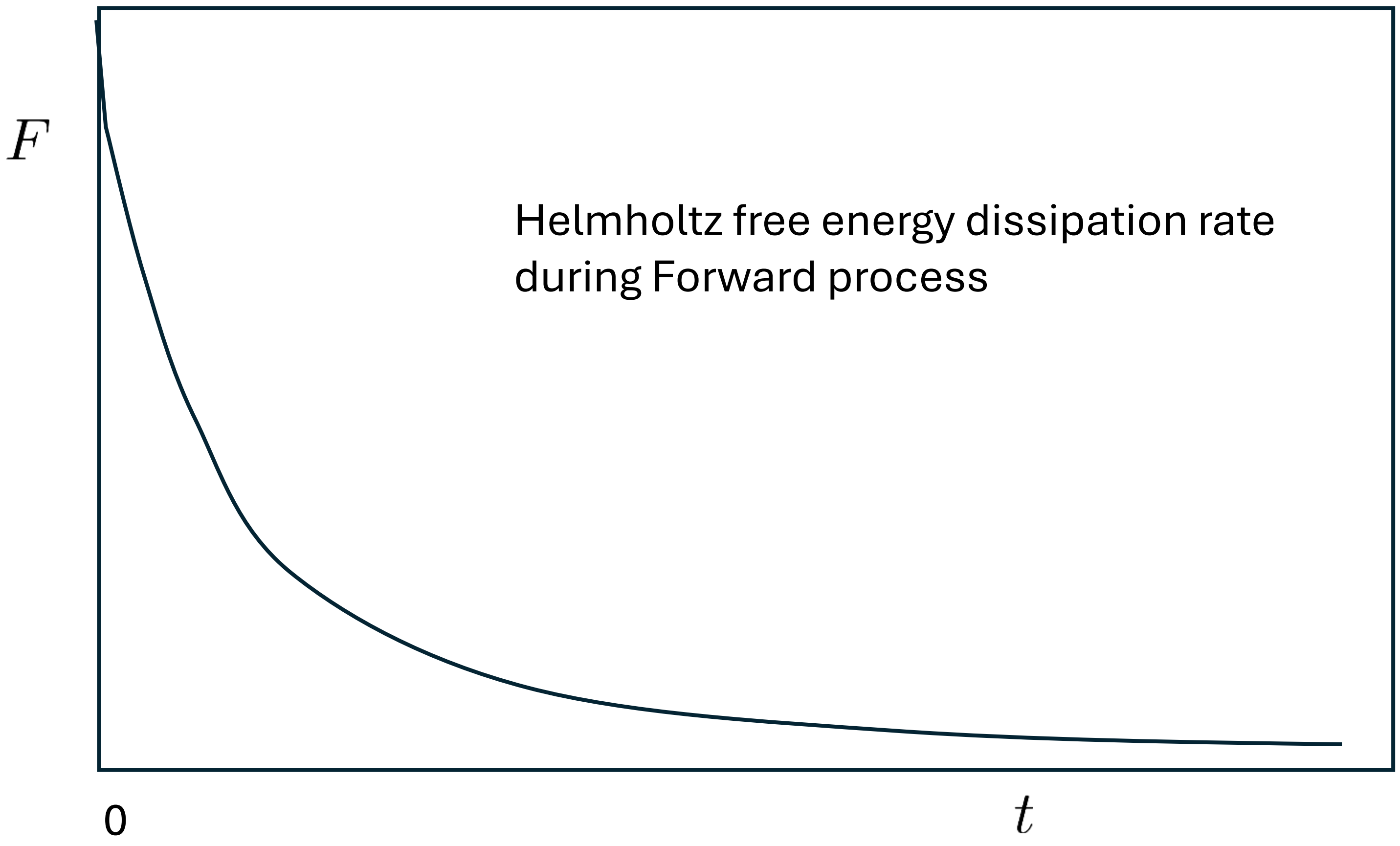
$$\vec{v}(\mathbf{x}) = -\nabla \frac{\partial F}{\partial \rho_t} \;=\; -\nabla U - T\nabla \log \rho_t$$
$$\frac{dF}{dt} = -\langle \vec{v}, \vec{v} \rangle_{\rho_t} \leq 0$$

[2] The rate of change decreases over time during the forward process

- The second derivative $\frac{\partial^2 F}{dt^2}$ is positive, as implied by results on the dissipation of Fisher information provided that the potential is "convex".

Flow to Maximize Free Energy Decrease

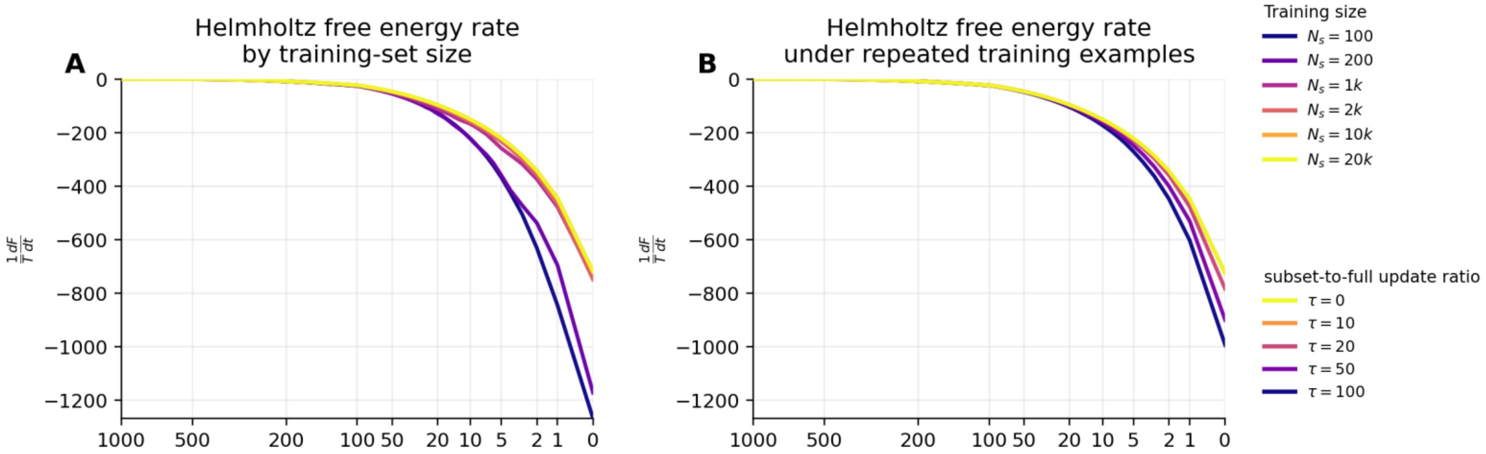
$$W \equiv \frac{1}{T}\left.\widehat{\frac{dF}{dt}}\right|_t = -\frac{\beta}{2}\frac{1}{N}\sum_{i=1}^{N}||\widehat{\mathbf{v}}_t(\mathbf{x}_{t,i})||^2$$



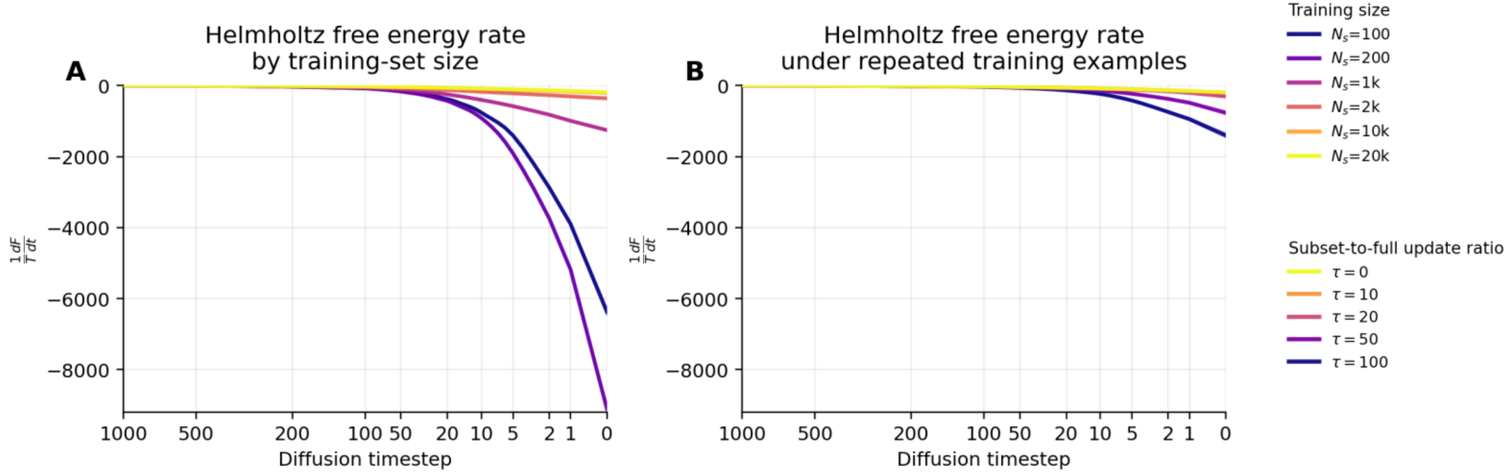
Flow to Maximize Free Energy Decrease

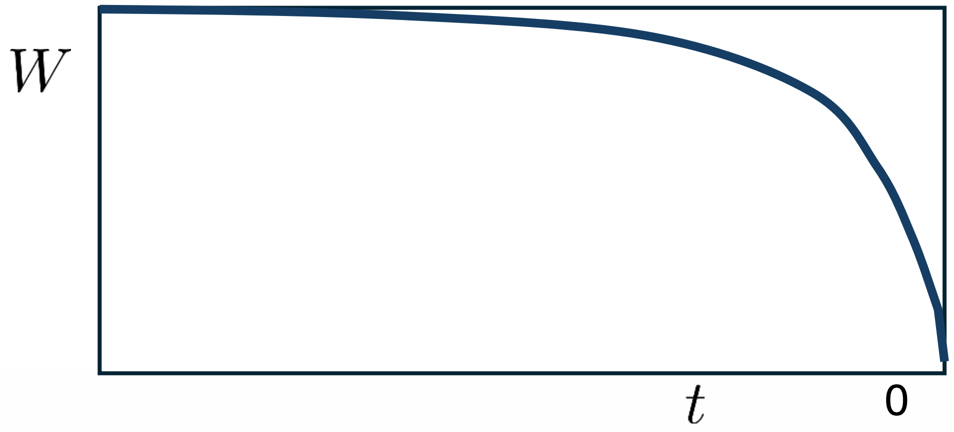
$$W \equiv \frac{1}{T}\widehat{\frac{dF}{dt}}\bigg|_t = -\frac{\beta}{2}\frac{1}{N}\sum_{i=1}^{N}||\widehat{\mathbf{v}}_t(\mathbf{x}_{t,i})||^2$$

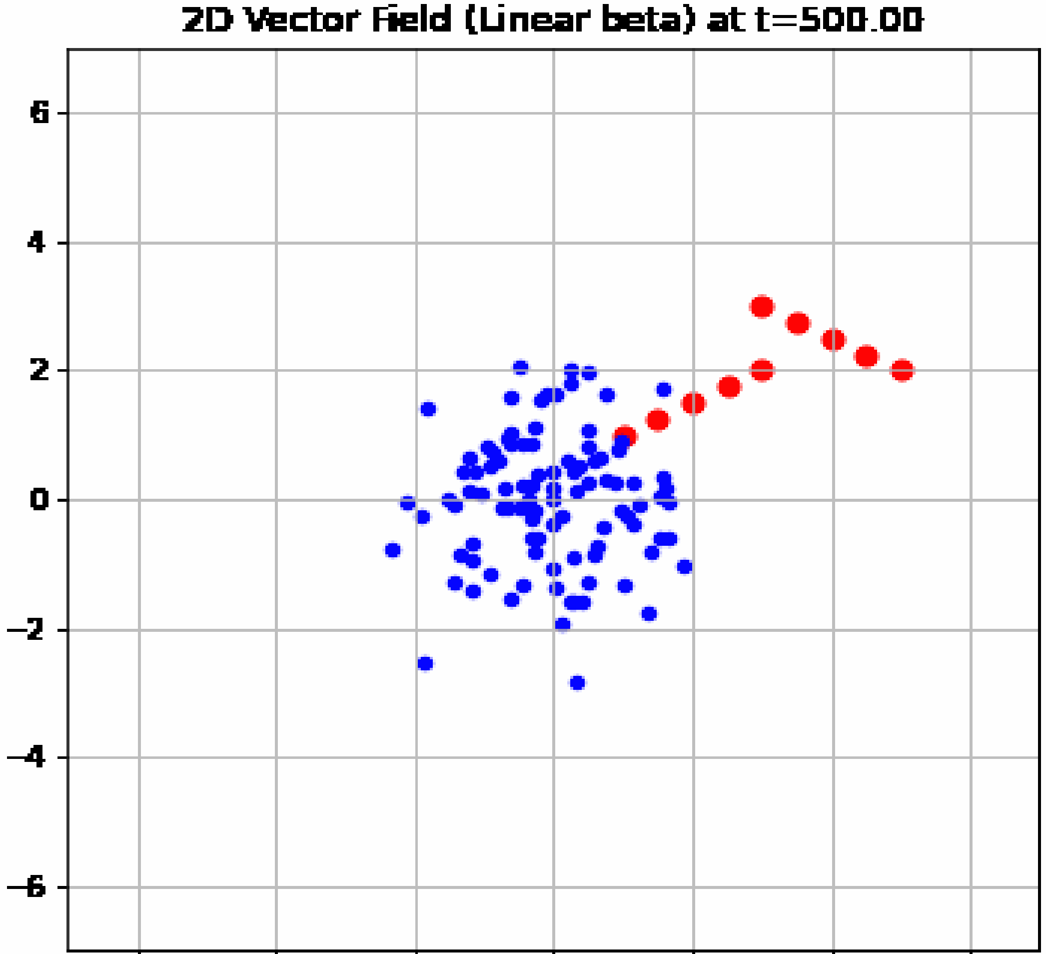
- CIFAR-10

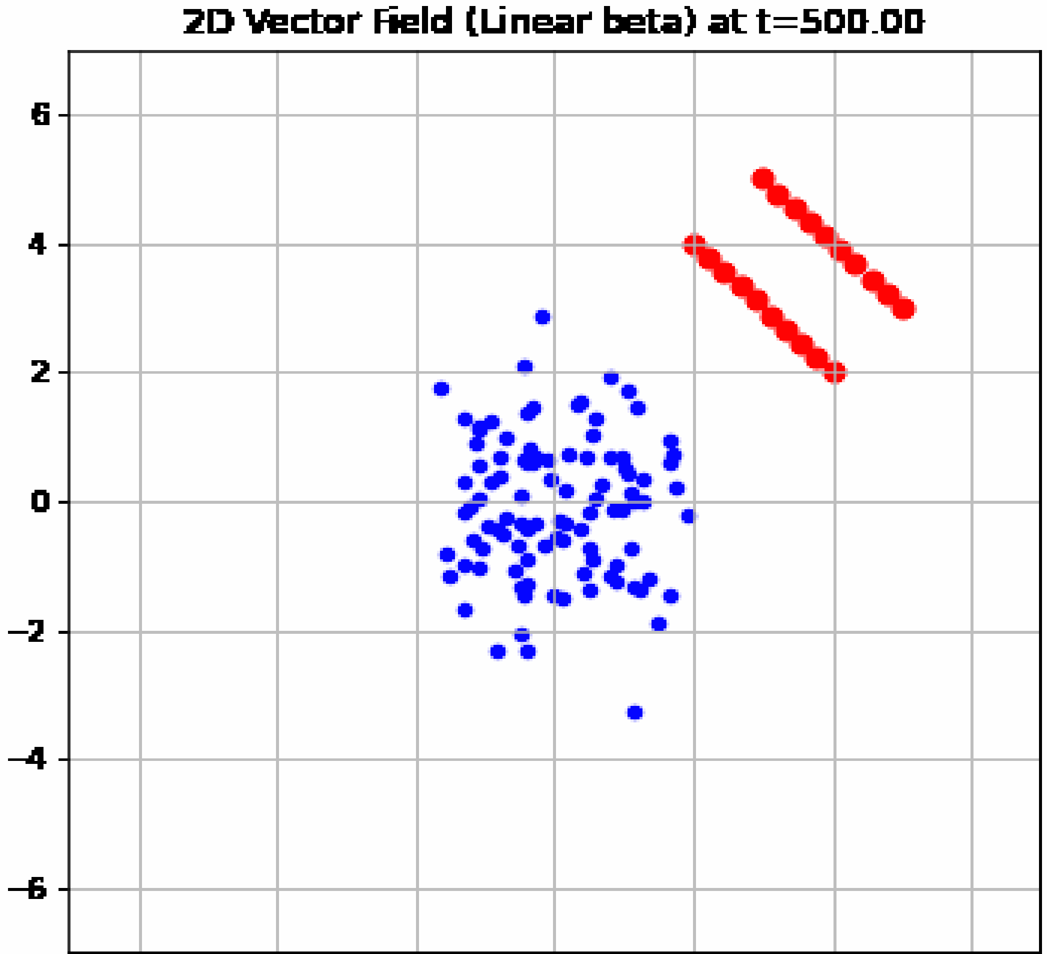


- CelebA-HQ




W
t
0


2D Vector Field (Linear beta) at t=500.00
6
4
2
0
−2
−4
−6


2D Vector Field (Linear beta) at t=500.00
6
4
2
0
−2
−4
−6

Continuous-time Diffusion of a δ-function

- Diffusion of delta (or mixture of delta) function

$$\rho(\mathbf{x};t) = k(\mathbf{x};t) = \mathcal{N}(0, \beta t I)$$

$$\mathbf{x}_{t+\Delta t} = \mathbf{x}_t + \sqrt{\beta \Delta t}\ \epsilon_t \qquad \epsilon_t \sim \mathcal{N}(0, I)$$

Continuous-time Diffusion of a δ-function

- Diffusion of delta function

$$\nabla \cdot (\vec{v}\ k(\mathbf{x};t)) = -\frac{dk(\mathbf{x};t)}{dt} \quad \& \quad k(\mathbf{x};t+\Delta t) = \int k(\mathbf{x}';t)k(\mathbf{x}'-\mathbf{x};\Delta t)d\mathbf{x}'$$

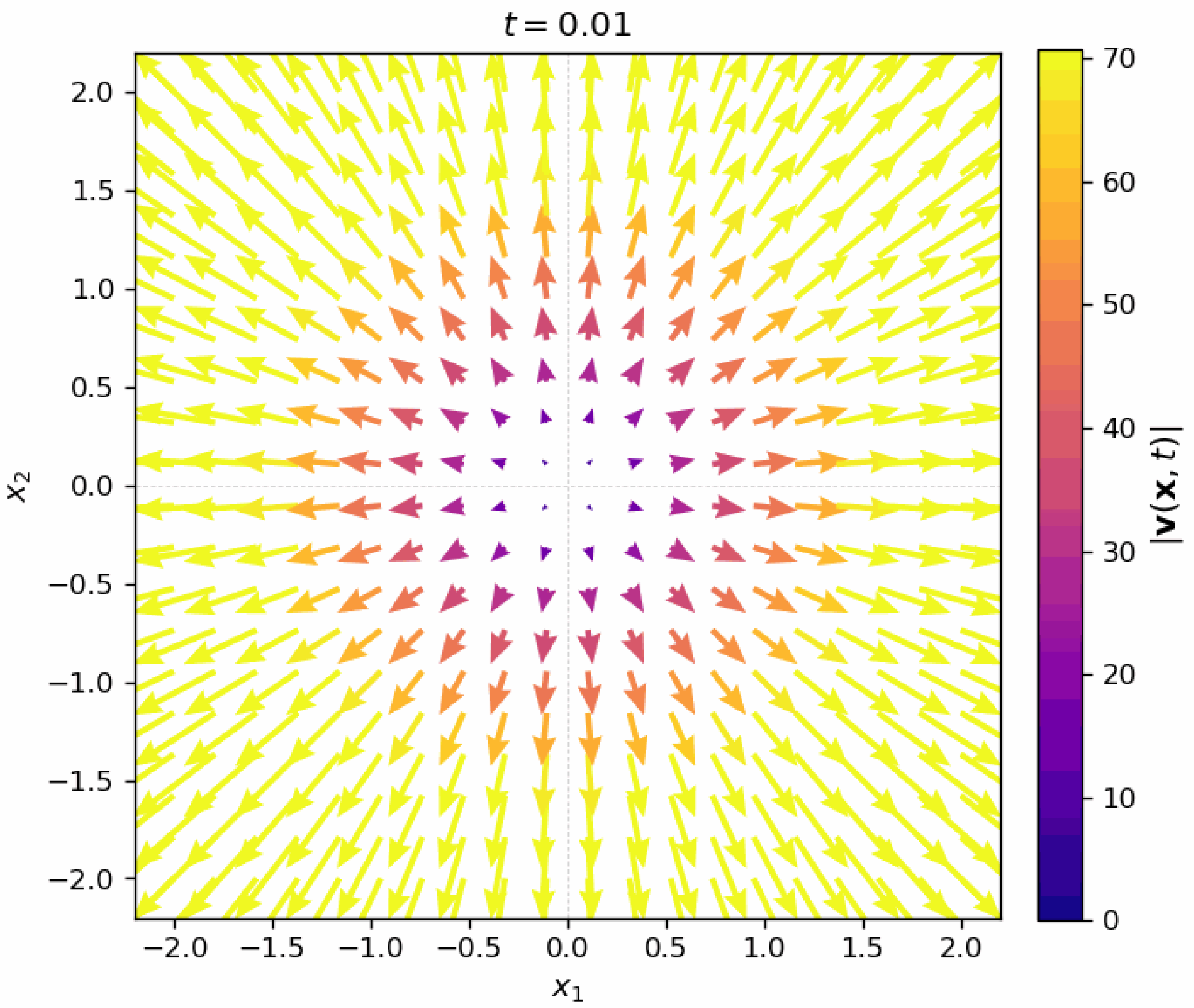
$$k(\mathbf{x};t)\nabla \cdot \vec{v} + \vec{v} \cdot \nabla k(\mathbf{x};t) = -\frac{dk(\mathbf{x};t)}{dt}$$

$$k(\mathbf{x};t)\left[\nabla \cdot \vec{v} - \frac{\mathbf{x}}{\beta t} \cdot \vec{v}\right] = -\frac{1}{2t}\left(\frac{\mathbf{x}^2}{\beta t} - d\right) k(\mathbf{x};t)$$

$$\nabla \cdot \vec{v} - \frac{\mathbf{x}}{\beta t} \cdot \vec{v} = -\frac{1}{2}\left(\frac{\mathbf{x}^2}{\beta t} - d\right)$$

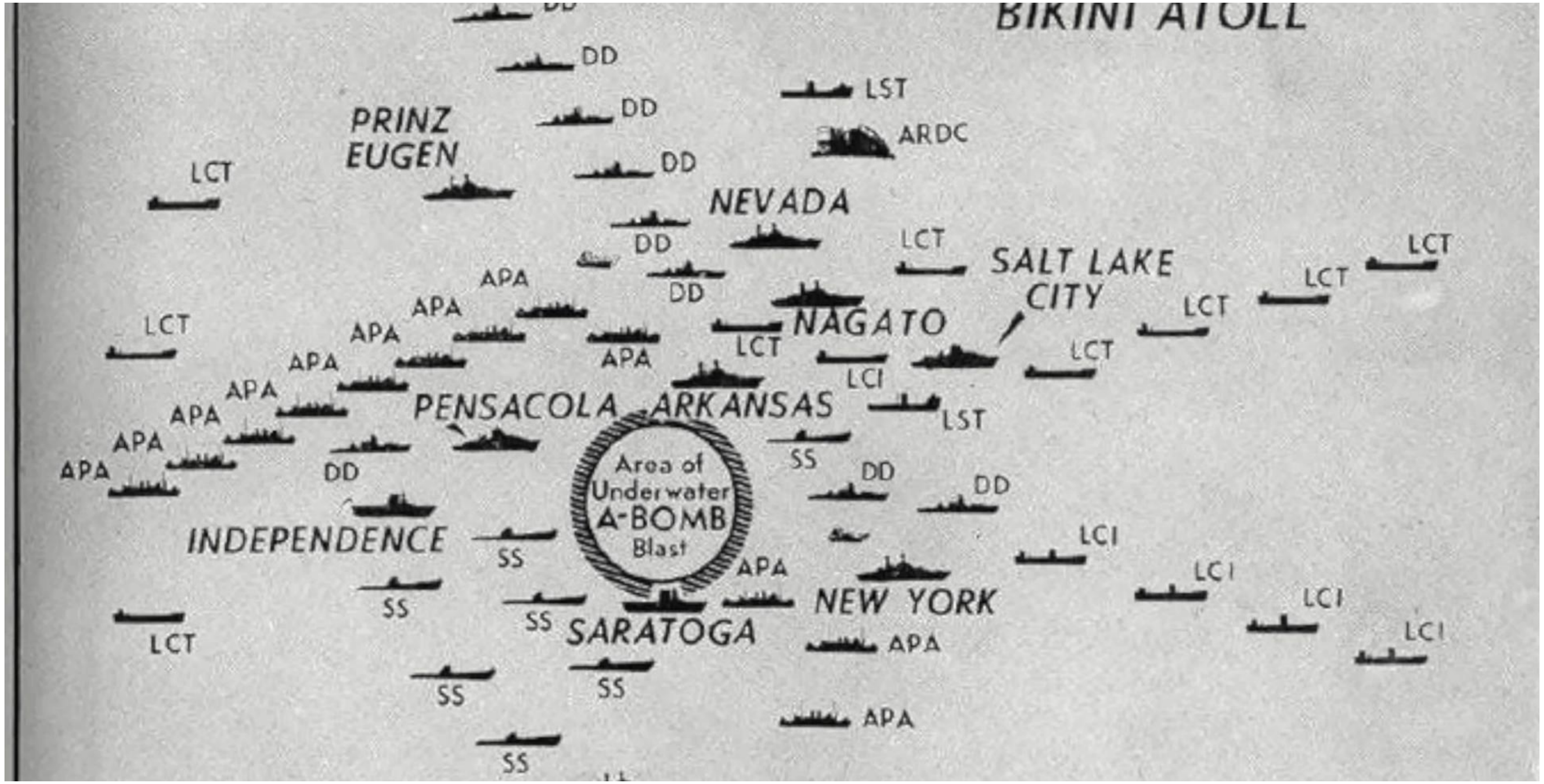
$$\Rightarrow \quad \vec{v}_t = \frac{1}{2t}\mathbf{x}$$

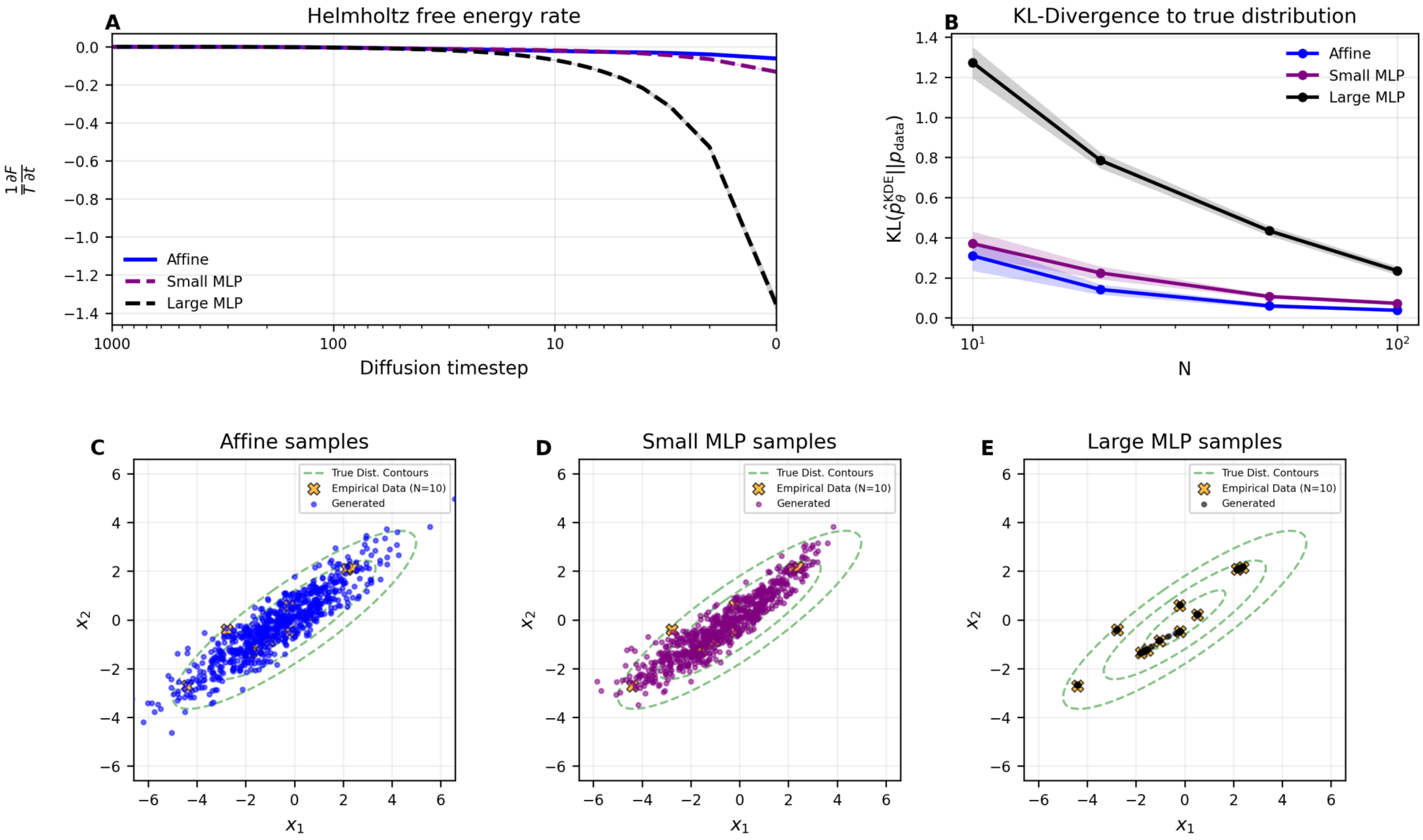
Velocity is β -independent



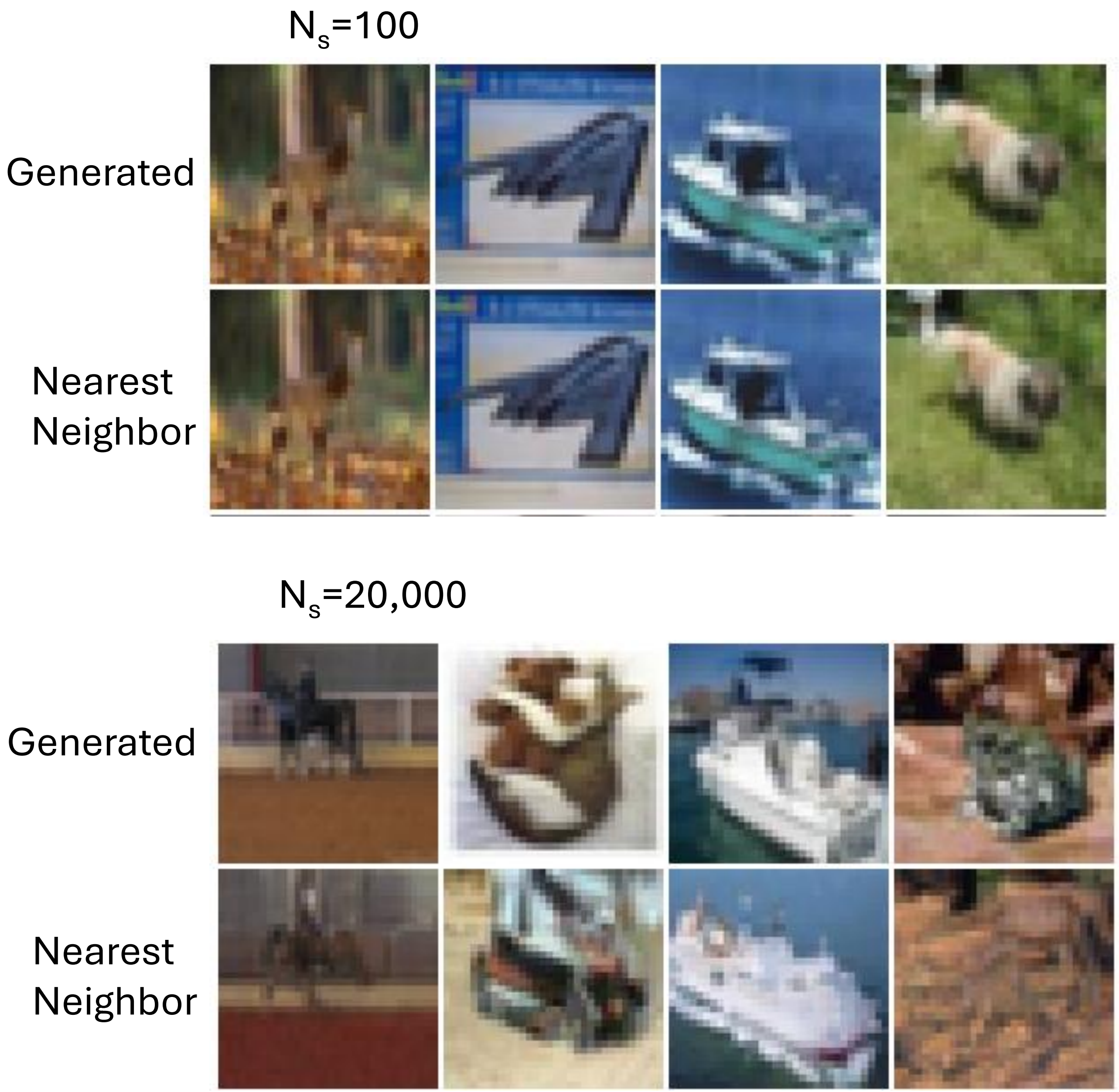
$$\vec{v}_t = \frac{1}{2t}\mathbf{x}$$

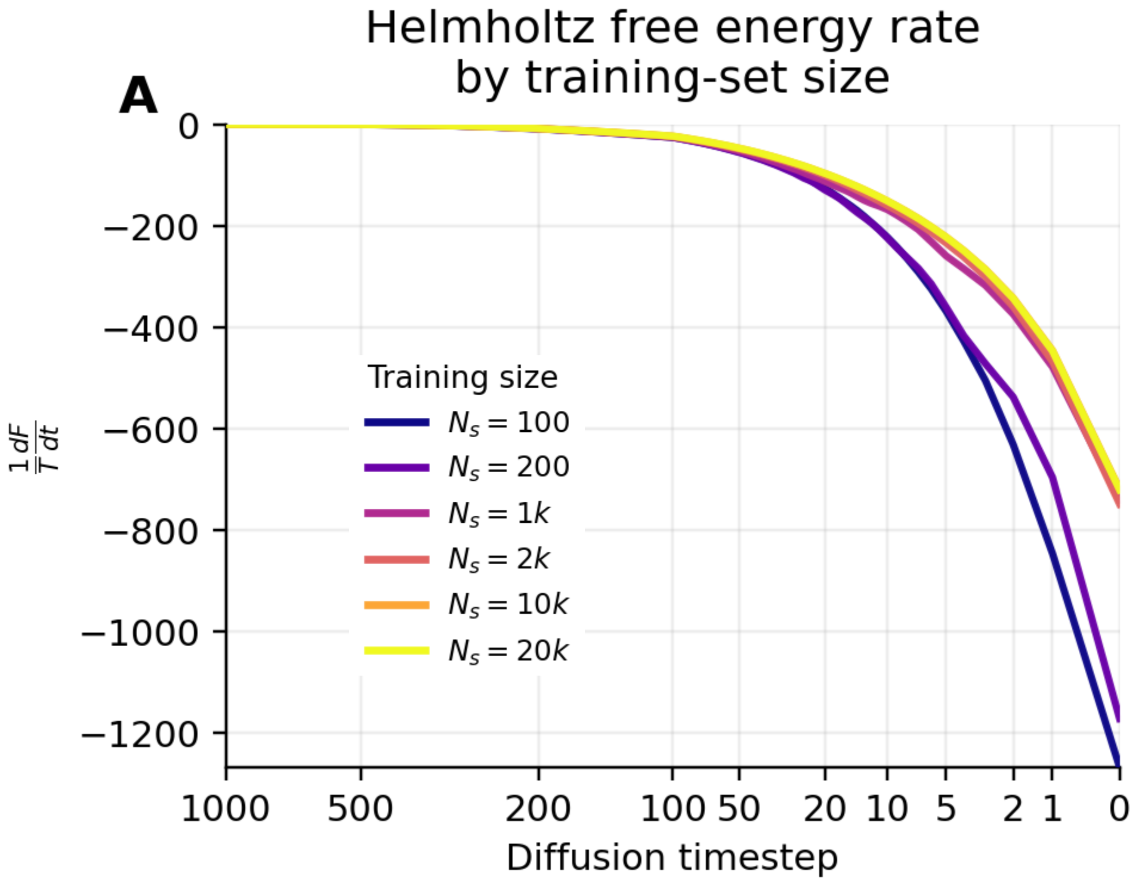
https://www.youtube.com/watch?v=AL-WRlEr54M

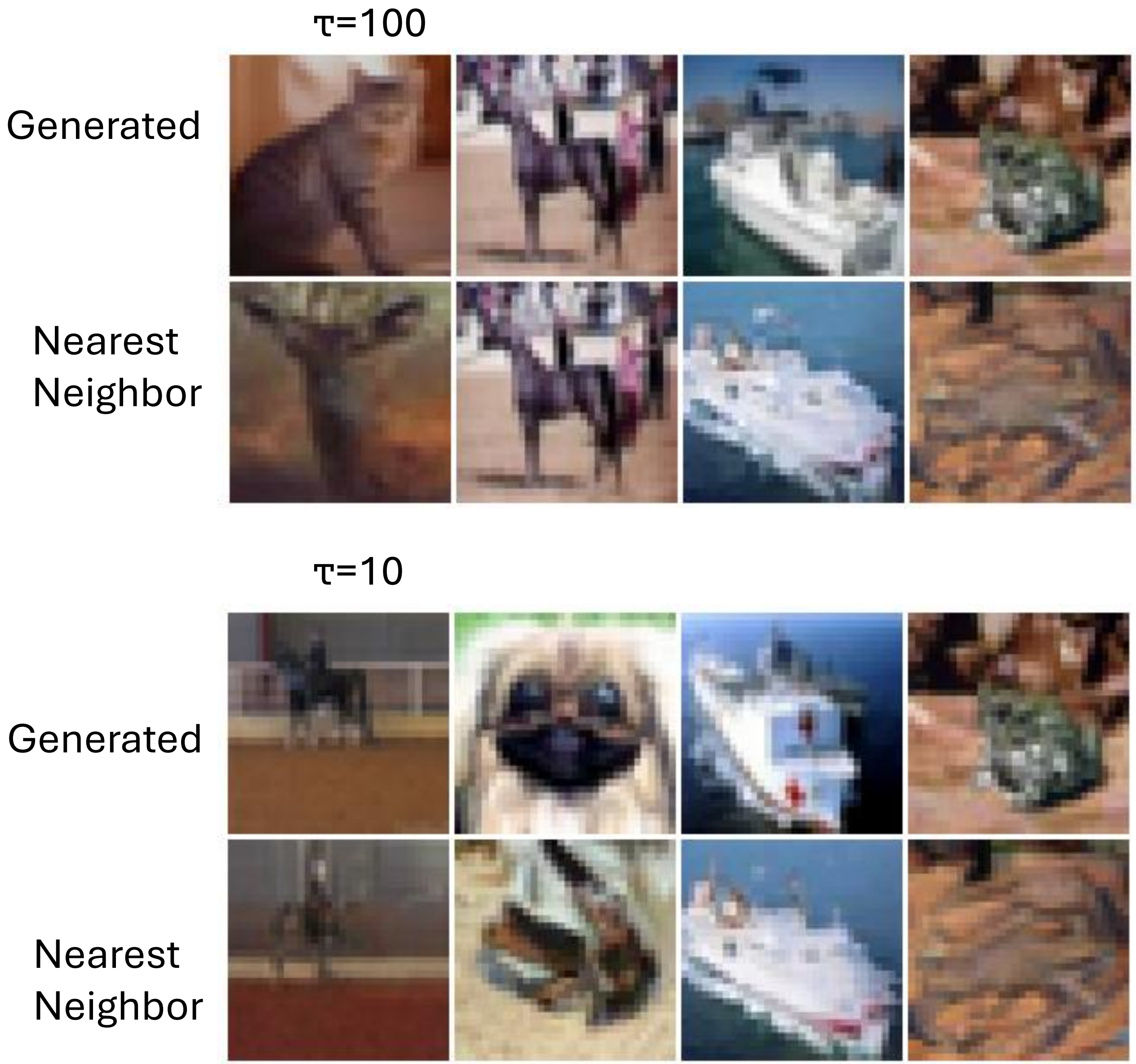


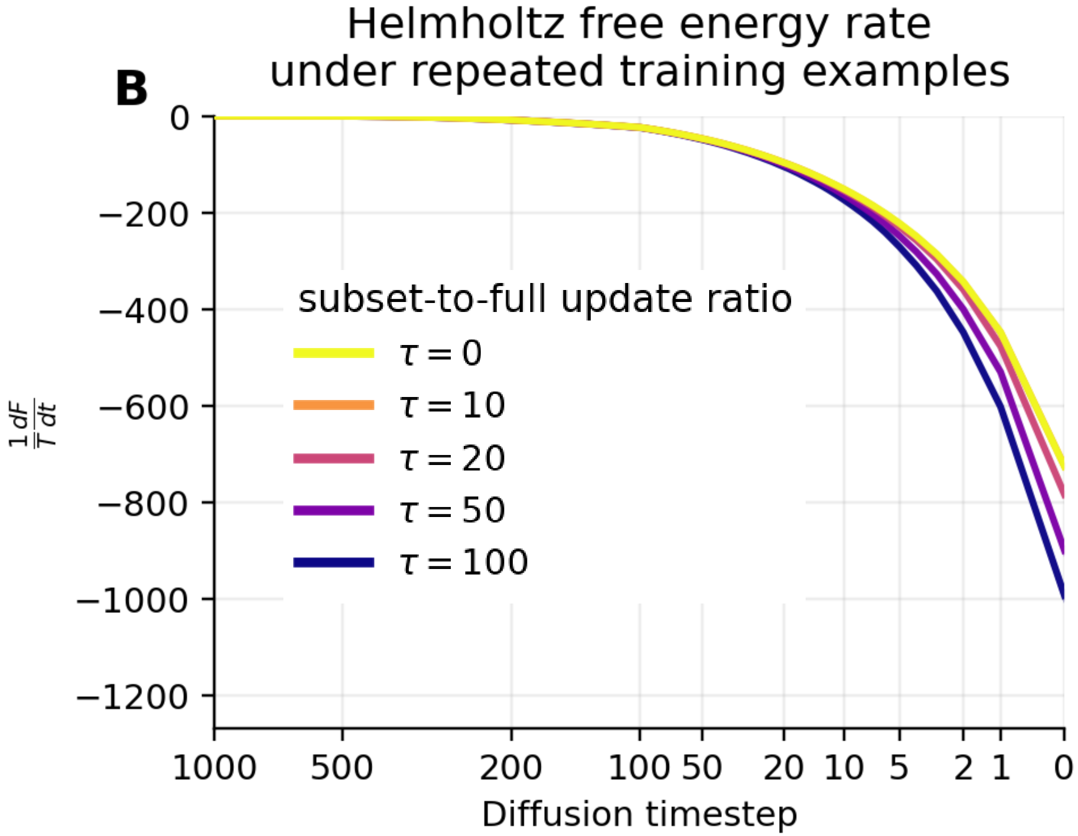


A
Helmholtz free energy rate
$\frac{1}{T}\frac{\partial F}{\partial t}$
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
−1.2
−1.4
1000
100
10
0
Diffusion timestep
Affine
Small MLP
Large MLP
B
KL-Divergence to true distribution
$\mathrm{KL}(\hat{p}_\theta^{\mathrm{KDE}}||p_{\mathrm{data}})$
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
10^1
10^2
N
Affine
Small MLP
Large MLP
C
Affine samples
D
Small MLP samples
E
Large MLP samples
True Dist. Contours
Empirical Data (N=10)
Generated
x_1
x_2


N_s=100
Generated
Nearest Neighbor
N_s=20,000
Generated
Nearest Neighbor


Helmholtz free energy rate
by training-set size
A
$\frac{1}{T}\frac{dF}{dt}$
0
−200
−400
−600
−800
−1000
−1200
Training size
$N_s = 100$
$N_s = 200$
$N_s = 1k$
$N_s = 2k$
$N_s = 10k$
$N_s = 20k$
1000 500 200 100 50 20 10 5 2 1 0
Diffusion timestep


τ=100
Generated
Nearest Neighbor
τ=10
Generated
Nearest Neighbor


Helmholtz free energy rate
under repeated training examples
B
$\frac{1}{T}\frac{dF}{dt}$
0
−200
−400
−600
−800
−1000
−1200
subset-to-full update ratio
$\tau = 0$
$\tau = 10$
$\tau = 20$
$\tau = 50$
$\tau = 100$
1000
500
200
100
50
20
10
5
2
1
0
Diffusion timestep

N_s=100

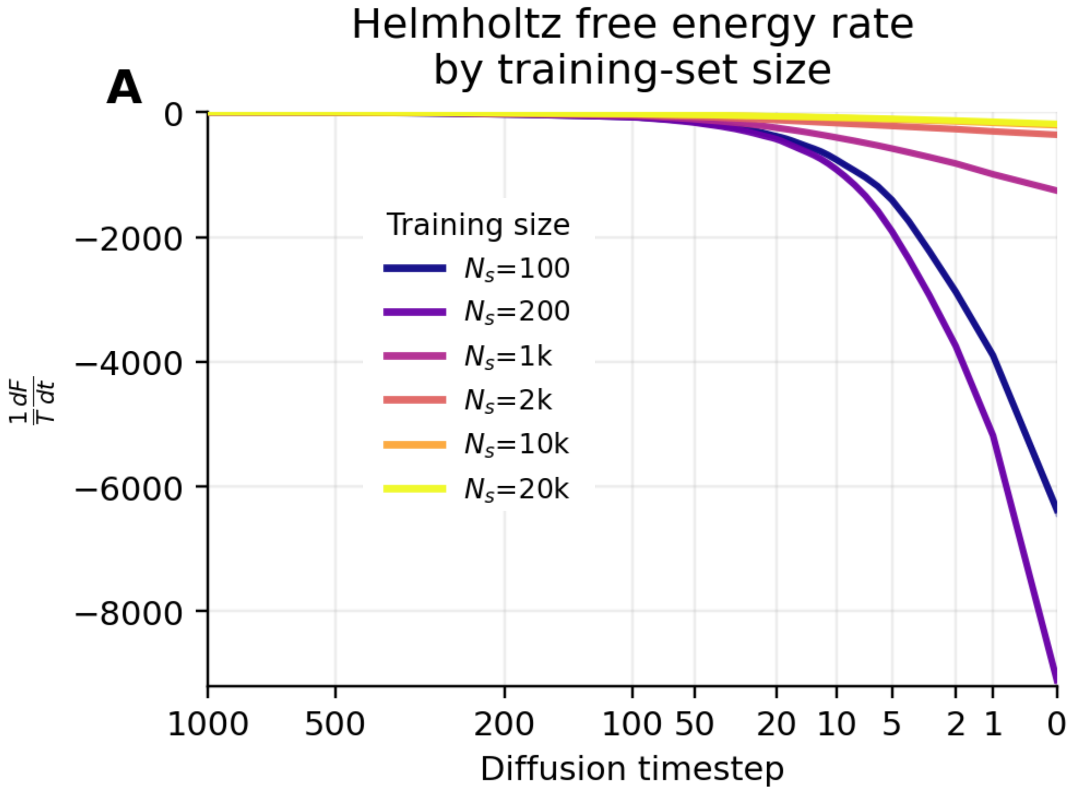
Generated

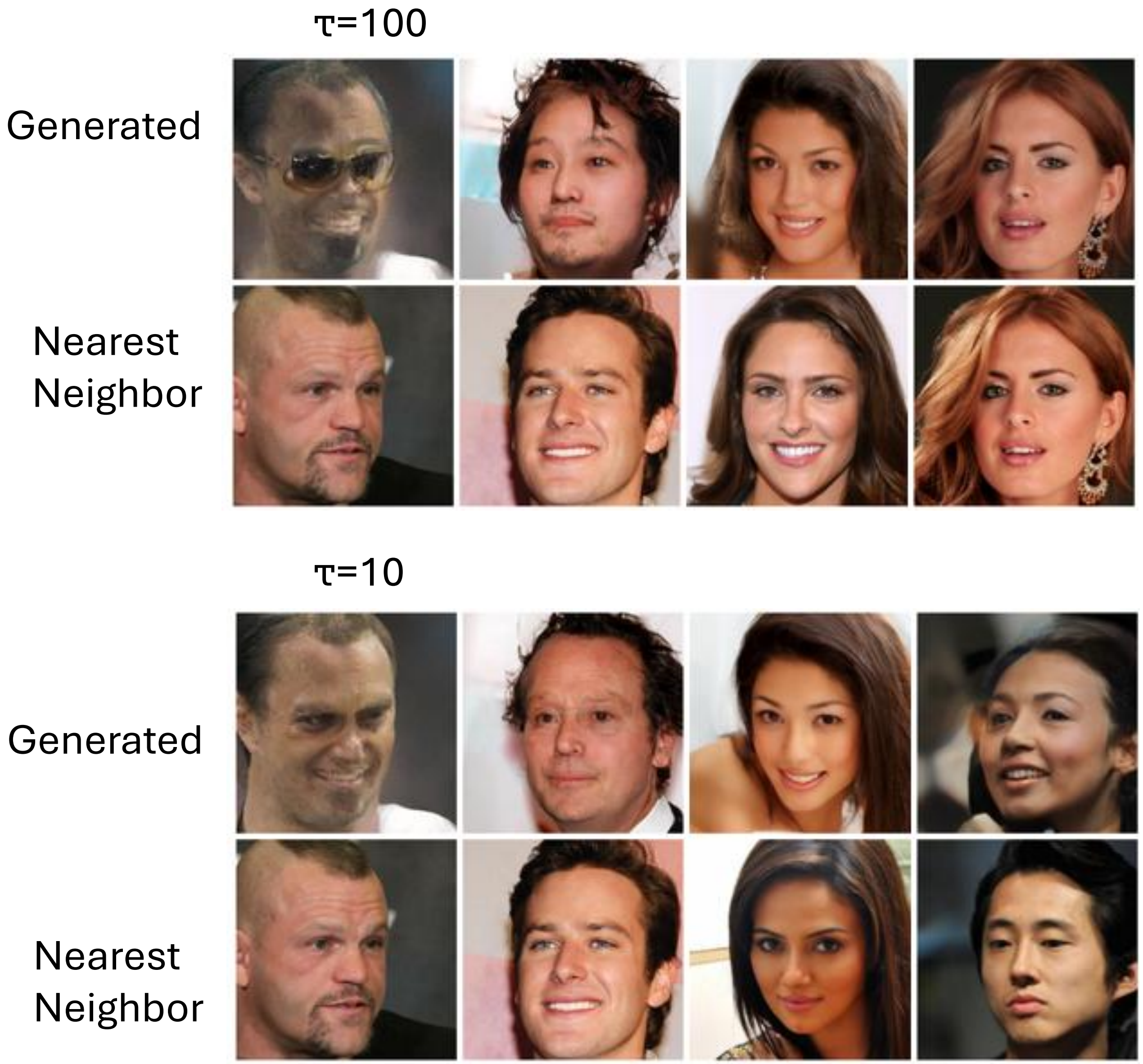
Nearest Neighbor

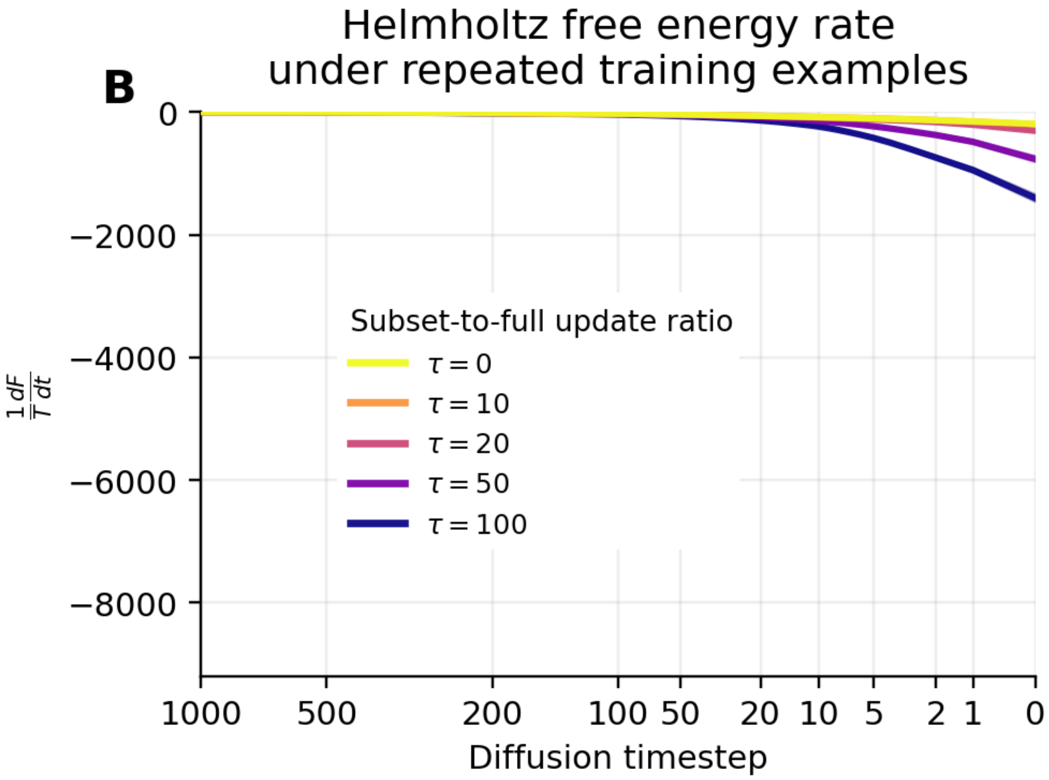
N_s=20,000

Generated

Nearest Neighbor




τ=100
Generated
Nearest
Neighbor
τ=10
Generated
Nearest
Neighbor


B
Helmholtz free energy rate
under repeated training examples
0
−2000
−4000
−6000
−8000
$\frac{1}{T}\frac{dF}{dt}$
Subset-to-full update ratio
$\tau = 0$
$\tau = 10$
$\tau = 20$
$\tau = 50$
$\tau = 100$
1000 500 200 100 50 20 10 5 2 1 0
Diffusion timestep

Sample-level Memorization Detection

- Stable diffusion v1.4

Method	Time step	AUC	TPR@1%FPR
$\left\|\mathbf{s}_\theta^\Delta(\mathbf{x}_t)\right\|$	1,000	0.937 ±0.012	0.771 ±0.027
$= \left\|\mathbf{s}_{\theta,\mathrm{cond}} - \mathbf{s}_{\theta,\mathrm{uncond}}\right\|$ [19]	1	0.972 ±0.005	**0.909** ±0.020
$\left\|H_\theta^\Delta(\mathbf{x}_T)\mathbf{s}_\theta^\Delta(\mathbf{x}_T)\right\|^2$ [7]	1,000	0.962 ±0.008	0.856 ±0.024
	1	0.970 ±0.005	**0.874** ±0.021
$\|\widehat{\mathbf{v}}_{\mathrm{cfg}}\|^2 - \|\widehat{\mathbf{v}}_{\mathrm{uncond}}\|^2$ (Ours)	1	**0.982** ±0.003	**0.905** ±0.019

CelebA-HQ Fine Tunning with Velocity Suppression

Condition	LPIPS ≤ 0.1	LPIPS ≤ 0.3	FID
Pretrained model	0.00%	0.08%	14.93
No reference scale	28.53%	33.11%	24.33

	Full velocity $\widehat{\mathbf{v}}_t = x_t + \mathbf{s}_\theta(x_t, t) \mid \lambda = 10^{-3}$			**Score-only** $\widehat{\mathbf{v}}_t = \mathbf{s}_\theta(x_t, t) \mid \lambda = 10^{-5}$		
Condition	LPIPS ≤ 0.1	LPIPS ≤ 0.3	FID	LPIPS ≤ 0.1	LPIPS ≤ 0.3	FID
$\gamma = 1.0$	31.85%	36.79%	26.19	23.14%	31.59%	22.71
$\gamma = 0.8$	11.39%	28.14%	17.99	2.47%	18.01%	53.80
$\gamma = 0.7$	3.50%	20.65%	15.88	1.46%	15.75%	25.74
$\gamma = 0.5$	0.05%	15.31%	**14.38**	0.29%	2.31%	156.08
$\gamma = 0.3$	**0.00%**	**3.80%**	15.87	0.02%	0.23%	295.58

Summary

- Data generation and the shift of distributions are naturally associated with Helmholtz free energy dissipation under diffusion dynamics (using the equation of continuity).

- Thermodynamic singularities in the diffusion of delta functions are characteristic of memorization, and its absence indicates non-memorization.

Thank you!

Yung-Kyun Noh

nohyung@hanyang.ac.kr