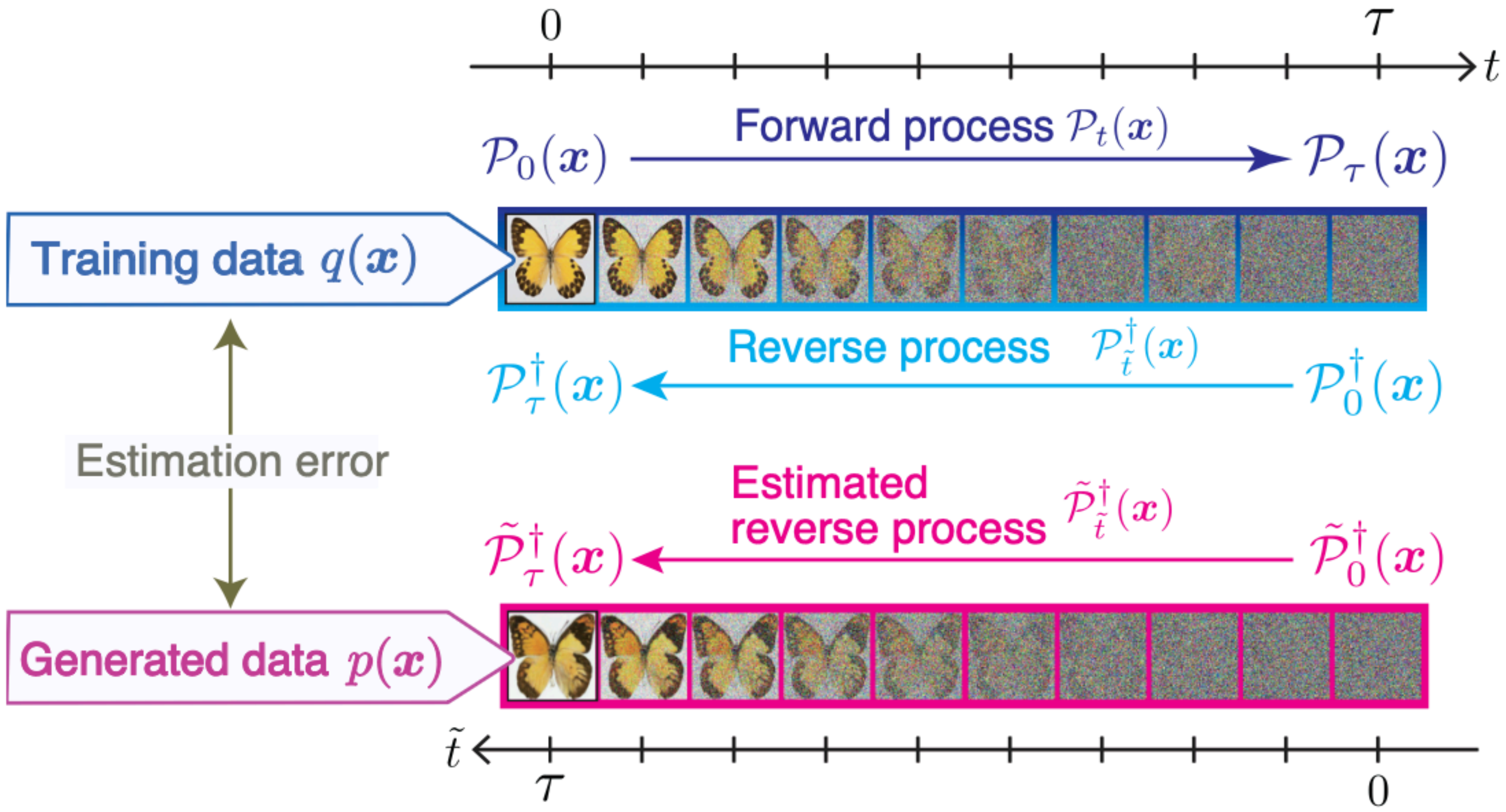
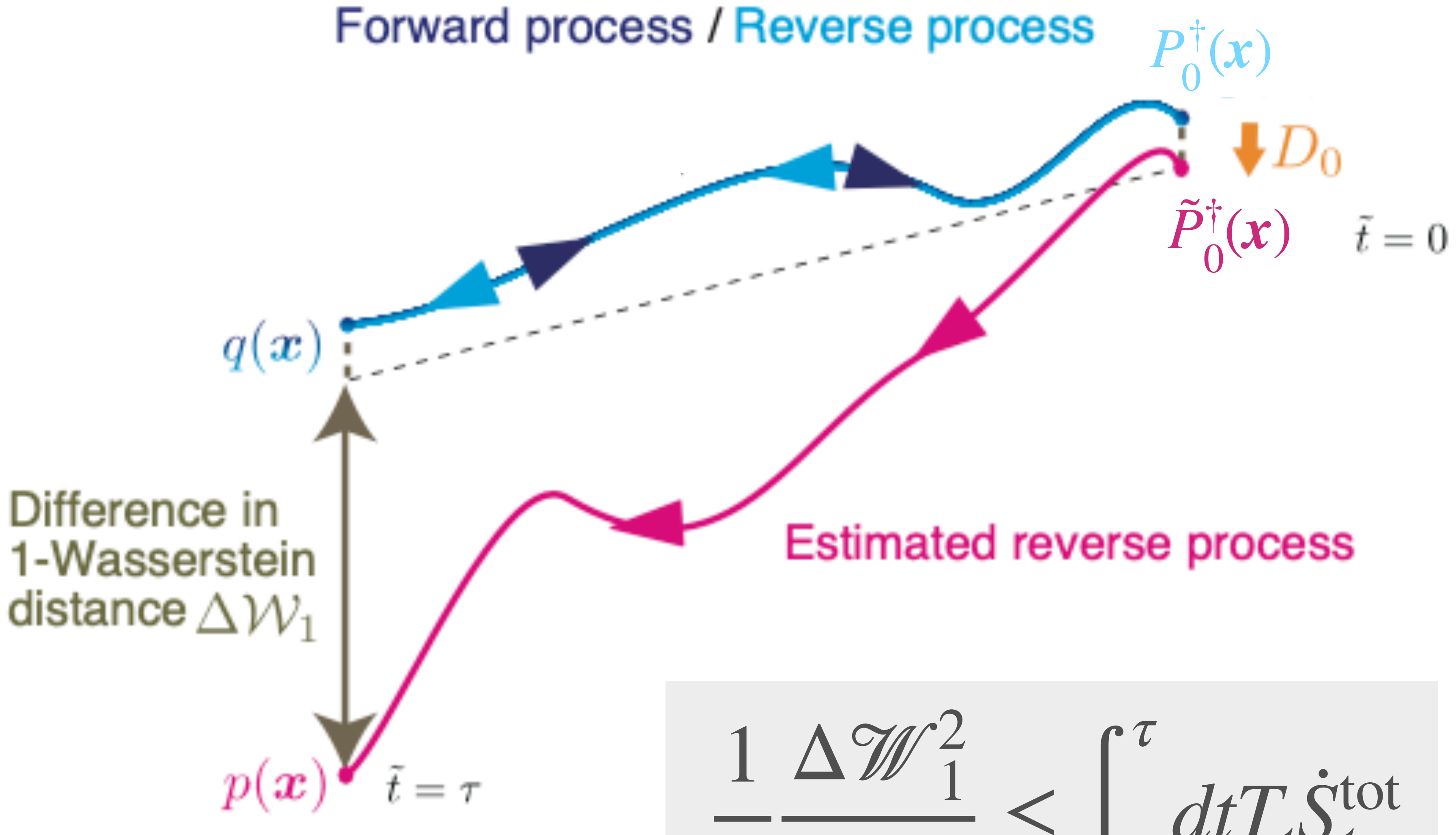
Stochastic thermodynamics for diffusion (generative) models





$$\frac{1}{\tau}\frac{\Delta\mathscr{W}_1^2}{D_0} \leq \int_0^\tau dt T_t \dot{S}_t^{\mathrm{tot}}$$

Sosuke Ito, Universal biology institute, the University of Tokyo Statphys 29, Florence, 17th July 2025

K. Ikeda, T. Uda, D. Okanohara & SI *arXiv:2407.04495* (to appear in Physical Review X)

UTokyo

Universal Biology Institute

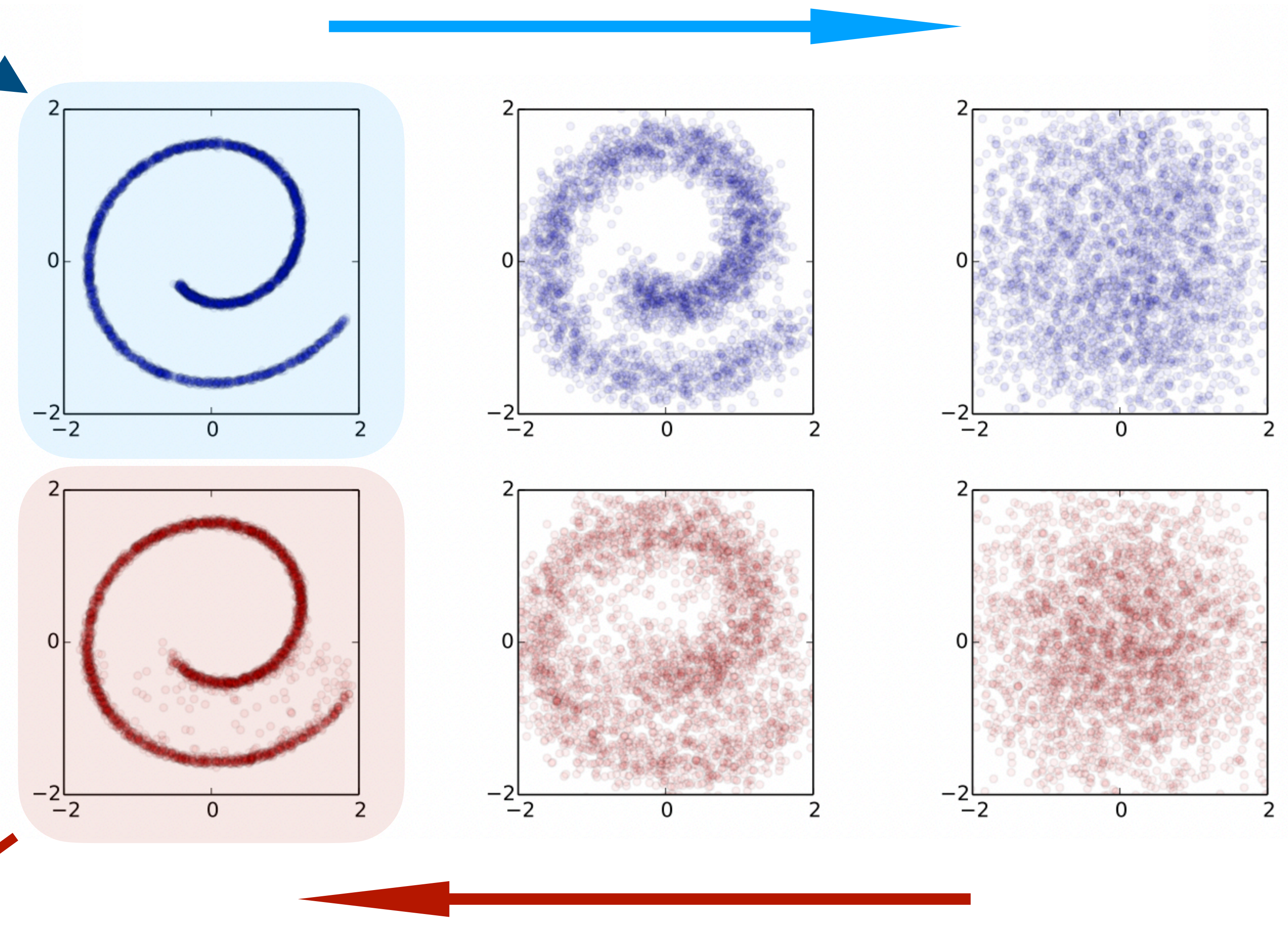
Diffusion models

- A method for generative artificial intelligence inspired by nonequilibrium thermodynamics

Deep Unsupervised Learning using Nonequilibrium Thermodynamics

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, Surya Ganguli *Proceedings of the 32nd International Conference on Machine Learning*, PMLR 37:2256-2265, 2015.

Training data



Forward diffusion process [Learning]

Reverse diffusion process [Data generation]

Generated data

Research Question

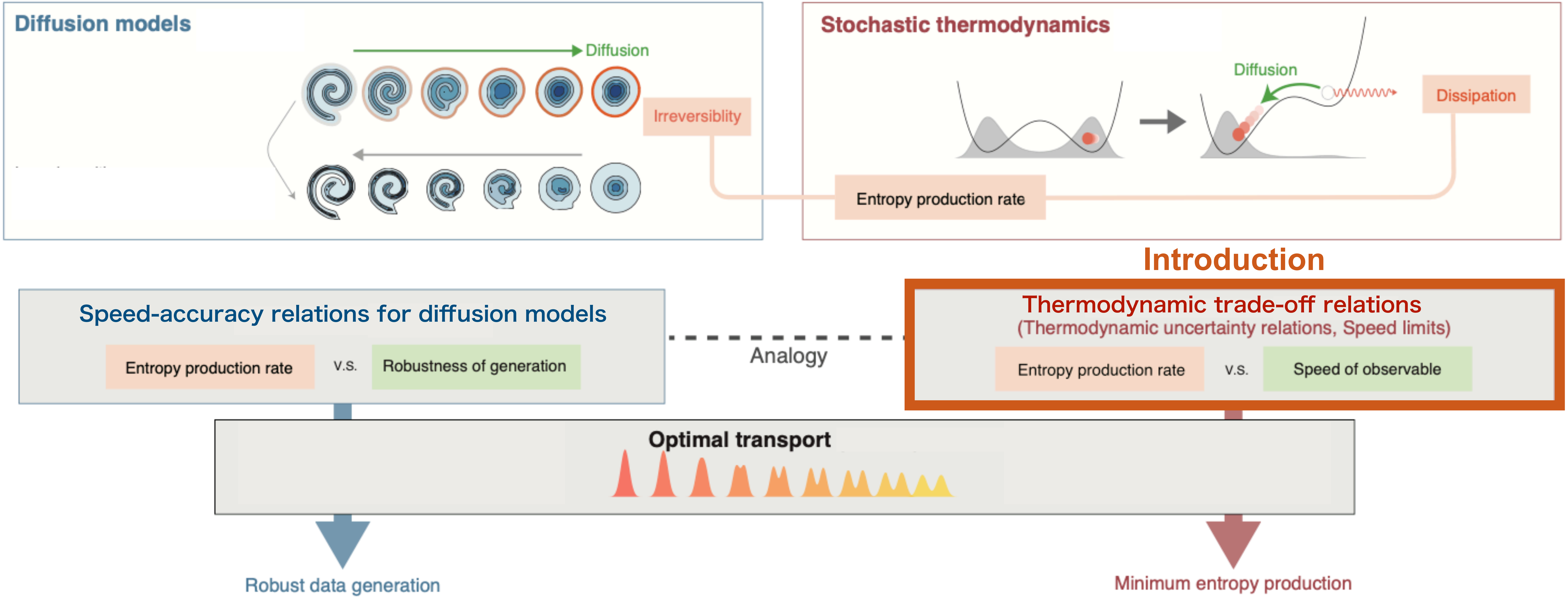
Is stochastic thermodynamics useful for understanding how accurately diffusion models generate data?

A: Yes.

Analogous to thermodynamic trade-off relations, we establish a relationship between thermodynamic dissipation and robust data generation.
The most robust data generation occurs through optimal transport dynamics in diffusion processes.

Our strategy

K. Ikeda, T. Uda, D. Okanohara & SI arXiv:2407.04495 (to appear in Physical Review X)



Stochastic thermodynamics and optimal transport: Lower bound on the entropy production rate

For information geometry and optimal transport in stochastic thermodynamics, see SI, *Information* Geometry 7.Suppl 1, 441-483 (2024).

Fokker-Planck equation

$$\partial_t P_t(\boldsymbol{x}) = -\nabla \cdot (\boldsymbol{\nu}_t(\boldsymbol{x}) P_t(\boldsymbol{x}))$$

$$\boldsymbol{\nu}_t(\boldsymbol{x}) = \boldsymbol{F}_t(\boldsymbol{x}) - T_t \nabla \ln P_t(\boldsymbol{x})$$

Entropy production rate

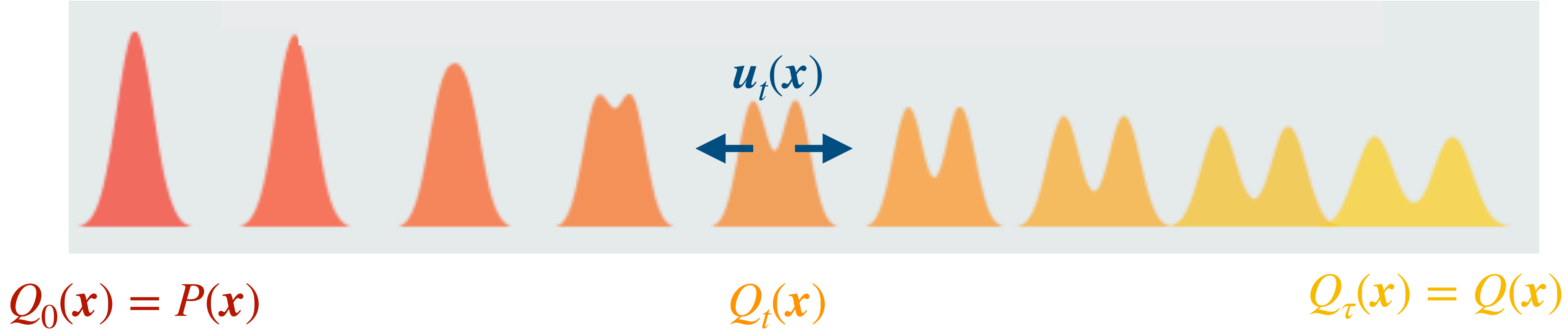
$$\dot{S}_t^{\mathrm{tot}} := \frac{1}{T_t} \int d\boldsymbol{x} P_t(\boldsymbol{x}) \|\boldsymbol{\nu}_t(\boldsymbol{x})\|^2$$

2-Wasserstein distance (Benamou-Brenier formula)

J-D. Benamou, and Y. Brenier. *Numerische Mathematik* 84, 375-393 (2000).

$$\mathcal{W}_2(P, Q) := \sqrt{\min_{(\boldsymbol{\nu}_t, Q_t)_{0 \le t \le \tau}} \tau \int_0^\tau dt \int d\boldsymbol{x} \,||\boldsymbol{u}_t(\boldsymbol{x})||^2 Q_t(\boldsymbol{x})}$$

s.t. $\partial_t Q_t(\boldsymbol{x}) = -\nabla \cdot (\boldsymbol{u}_t(\boldsymbol{x}) Q_t(\boldsymbol{x}))$,
$Q_0(\boldsymbol{x}) = P(\boldsymbol{x}), Q_\tau(\boldsymbol{x}) = Q(\boldsymbol{x})$



Speed in the space of 2-Wasserstein distance

$$v_2(t) = \lim_{\Delta t \to +0} \frac{\mathcal{W}_2(P_{t+\Delta t}, P_t)}{\Delta t}$$

Lower bound on the entropy production rate

M. Nakazato and SI. Phys. Rev. Res. 3, 043093 (2021).

$$\dot{S}_t^{\mathrm{tot}} \ge \frac{[v_2(t)]^2}{T_t}$$

Thermodynamic trade-off relations and minimal dissipation

For information geometry and optimal transport in stochastic thermodynamics, see SI, *Information* Geometry 7.Suppl 1, 441-483 (2024).

Speed of observable $r(\boldsymbol{x})$

$$v_r(t) := \frac{\left|\frac{d}{dt}\mathbb{E}_{P_t}[r]\right|}{\sqrt{\mathbb{E}_{P_t}[\|\nabla r\|^2]}} \qquad \mathbb{E}_{P_t}[r] := \int d\boldsymbol{x} P_t(\boldsymbol{x}) r(\boldsymbol{x})$$

Thermodynamic uncertainty relations

A. Dechant, S-I Sasa and SI. Phys. Rev. Res. 4, L012034 (2022).

$$\dot{S}_t^{\mathrm{tot}} \geq \frac{[v_2(t)]^2}{T_t} \geq \frac{[v_r(t)]^2}{T_t}$$

Speed limits (Minimal dissipation)

E. Aurell, et al. *Journal of statistical physics* 147, 487-505 (2012).
M. Nakazato and SI. Phys. Rev. Res. 3, 043093 (2021).

$$T_t = T(=\mathrm{const.})$$

$$\int_0^\tau dt \dot{S}_t^{\mathrm{tot}} \geq \frac{[\int_0^\tau dt v_2(t)]^2}{T\tau} \geq \frac{[\mathscr{W}_2(P_0, P_\tau)]^2}{T\tau}$$

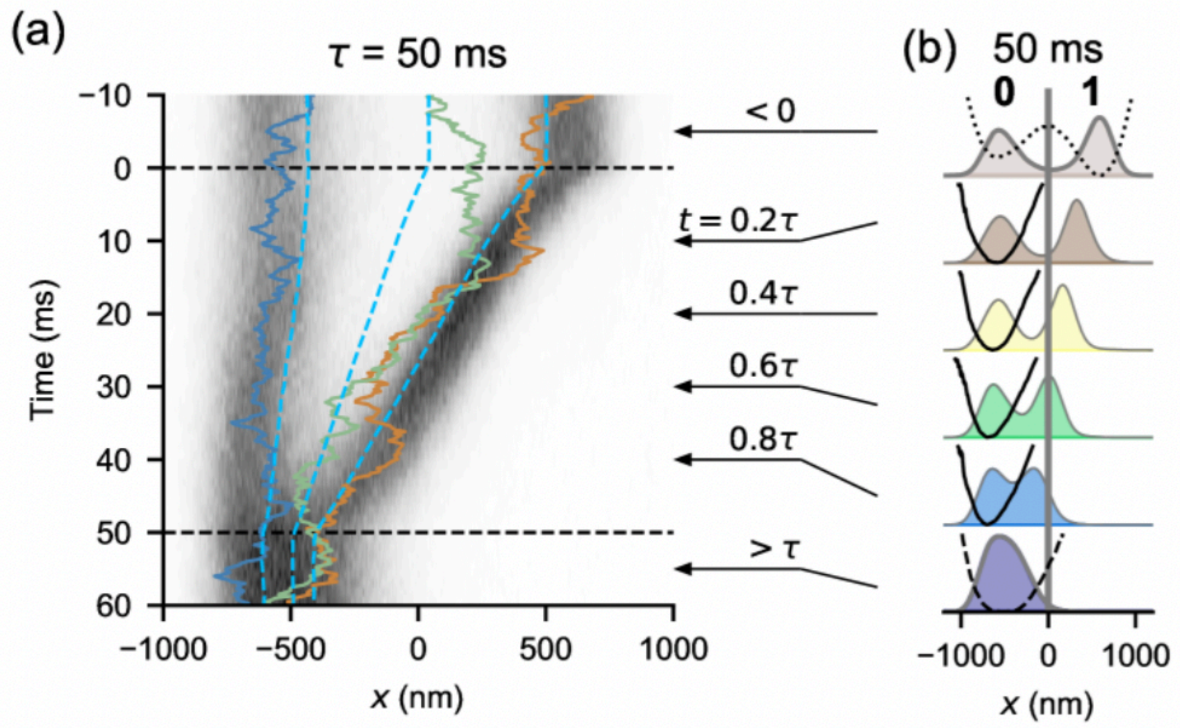
Optimal protocol

1. Conservative force: $F_t(\boldsymbol{x}) = -\nabla U_t(\boldsymbol{x})$
2. Geodesic: $v_2(t) = \dfrac{\mathscr{W}_2(P_0, P_\tau)}{\tau} = \mathrm{const.}$

J-D. Benamou, and Y. Brenier. *Numerische Mathematik* 84, 375-393 (2000).
E. Aurell, et al. Phys. Rev. Lett.,106, 250601 (2011).
M. Nakazato and SI. Phys. Rev. Res. 3, 043093 (2021).

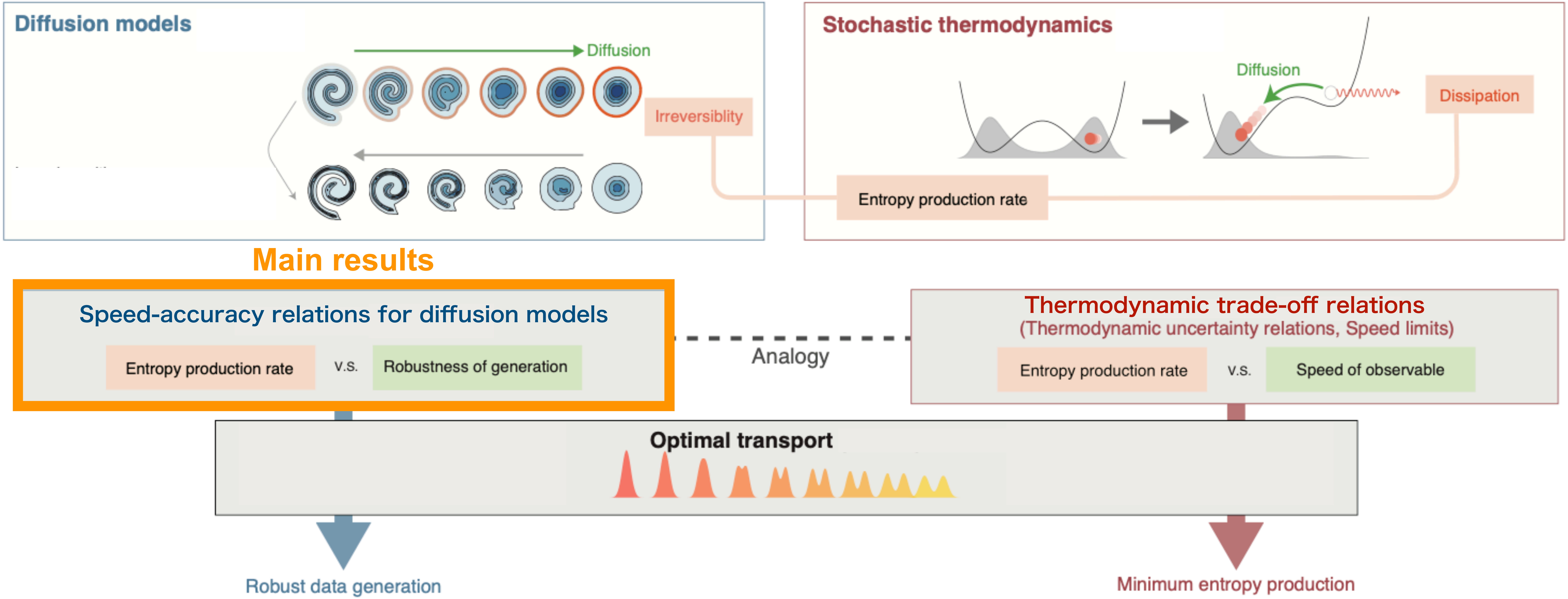
Experiment of minimal dissipation via optimal transport using optical tweezers

S. Oikawa, Y. Nakayama, SI., T. Sagawa, & S. Toyabe, arXiv 2053.01200.

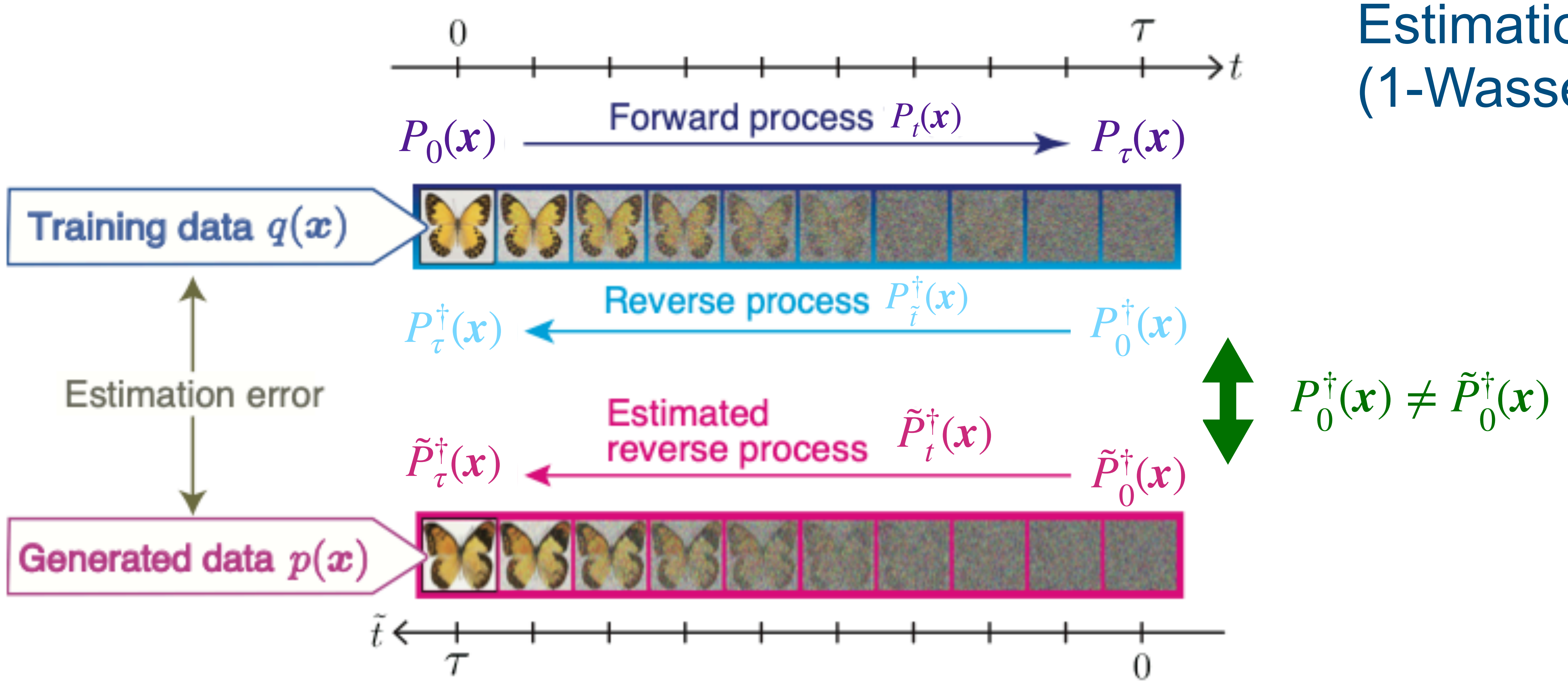


Our strategy

K. Ikeda, T. Uda, D. Okanohara & SI arXiv:2407.04495 (to appear in Physical Review X)



Estimation error and response function in diffusion models



Estimation error
(1-Wasserstein distance between training data and generated data)

$$\mathscr{W}_1(p,q) := \sup_{\psi} \left[\int d\boldsymbol{x} p(\boldsymbol{x})\psi(\boldsymbol{x}) - \int d\boldsymbol{y} q(\boldsymbol{y})\psi(\boldsymbol{y}) \right]$$

s.t. $\|\nabla\psi(\boldsymbol{x})\| \leq 1$

e.g.,) K. Oko, S. Akiyama & T. Suzuki,
In *International Conference on Machine Learning* (pp. 26517-26582). PMLR (2023).

Initial perturbation (χ^2-divergence)

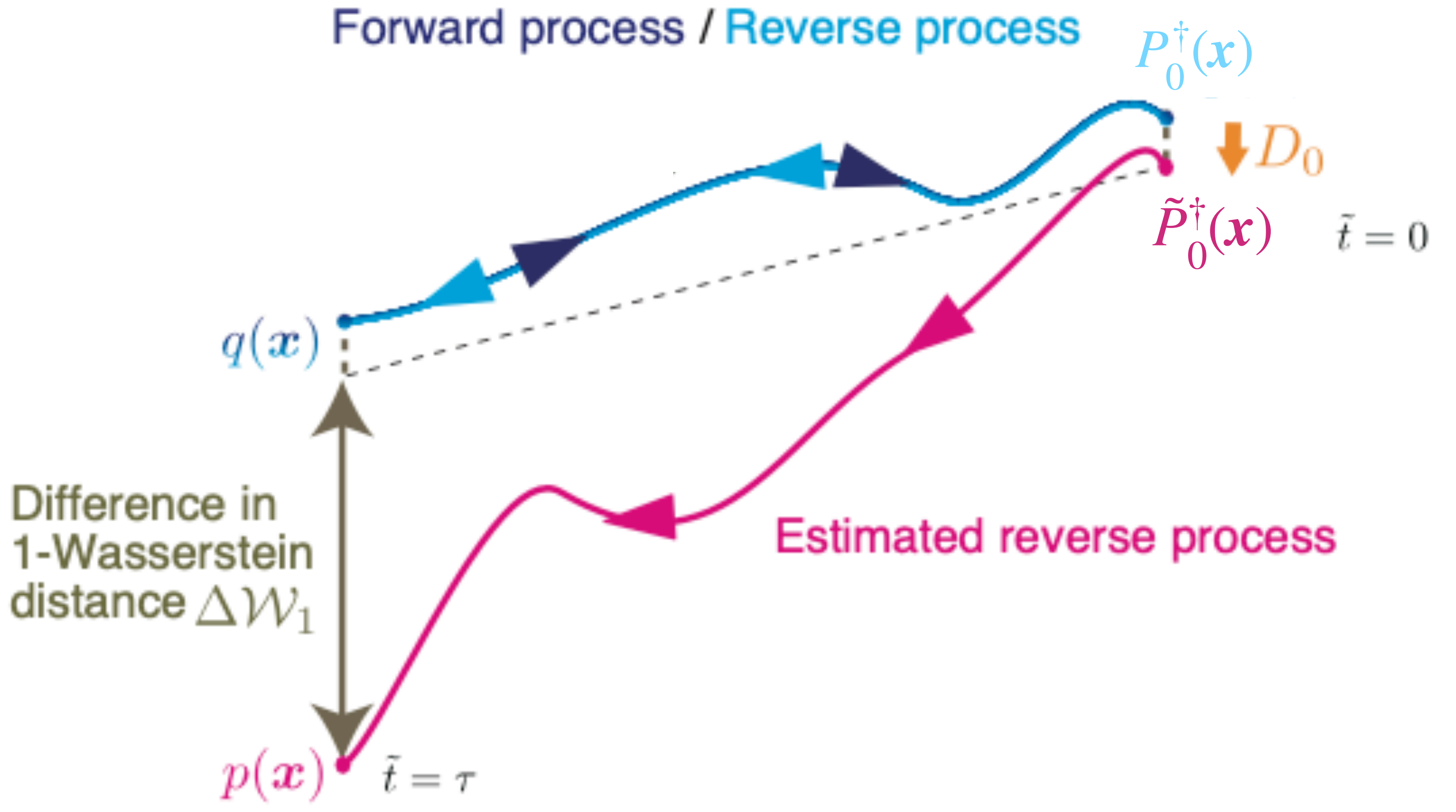
$$D_0 = \int d\boldsymbol{x} \frac{(P^\dagger_0(\boldsymbol{x}) - \tilde{P}^\dagger_0(\boldsymbol{x}))^2}{P^\dagger_0(\boldsymbol{x})}$$

Response function (Sensitivity)

Estimation error

$$\frac{\Delta\mathscr{W}_1^2}{D_0} = \frac{[\mathscr{W}_1(p,q) - \mathscr{W}_1(P^\dagger_0, \tilde{P}^\dagger_0)]^2}{D_0}$$

Perturbation



Response function (Sensitivity) $\frac{\Delta\mathscr{W}_1^2}{D_0}$ is **small**.

$\Rightarrow$ Data generation is **robust** to the initial perturbation.

Main results

Speed-accuracy relations for diffusion models

Response (~ Robust data generation) Dissipation Speed

$$\frac{1}{\tau}\frac{\Delta \mathscr{W}_1^2}{D_0} \leq \frac{\int_0^\tau dt\,|\partial_t \mathscr{W}_1(P^\dagger_{\tau-t}, \tilde{P}^\dagger_{\tau-t})|^2}{D_0} \leq \int_0^\tau dt T_t \dot{S}_t^{\mathrm{tot}} = \int_0^\tau dt[v_2(t)]^2$$

If the force is conservative: $F_t(\boldsymbol{x}) = -\nabla U_t(\boldsymbol{x})$

"Less dissipation ensures robust data generation."

Instantaneous bound

$$[v_{\mathrm{loss}}(t)]^2 := \frac{|\partial_t \mathscr{W}_1(P^\dagger_{\tau-t}, \tilde{P}^\dagger_{\tau-t})|^2}{D_0} \leq T_t \dot{S}_t^{\mathrm{tot}} = [v_2(t)]^2$$

If the force is conservative: $F_t(\boldsymbol{x}) = -\nabla U_t(\boldsymbol{x})$

Optimal protocol for robust data generation

Speed-accuracy relations for diffusion models

Response (~ Robust data generation)

Dissipation

$$\frac{1}{\tau}\frac{\Delta \mathscr{W}_1^2}{D_0} \le \frac{\int_0^\tau dt\,|\partial_t \mathscr{W}_1(P^\dagger_{\tau-t}, \tilde{P}^\dagger_{\tau-t})|^2}{D_0} \le \int_0^\tau dt T_t \dot{S}^{\mathrm{tot}}_t$$

Minimizing the upper bound

cf.) Minimal dissipation via optimal transport

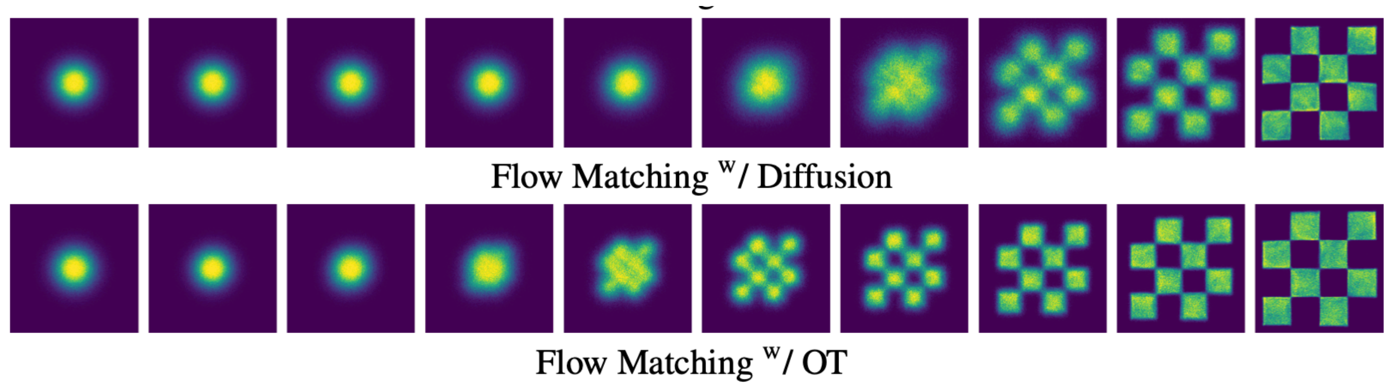
$$\int_0^\tau dt T_t \dot{S}^{\mathrm{tot}}_t \ge \frac{[\int_0^\tau dt v_2(t)]^2}{\tau} \ge \frac{[\mathscr{W}_2(P_0, P_\tau)]^2}{\tau}$$

Optimal protocol

1. **Conservative force:** $F_t(\boldsymbol{x}) = -\nabla U_t(\boldsymbol{x})$
2. **Geodesic:** $v_2(t) = \dfrac{\mathscr{W}_2(P_0, P_\tau)}{\tau} = \mathrm{const}\,.$

Empirical fact: Robust data generation can be achievable if we use approximate optimal transport schedules.

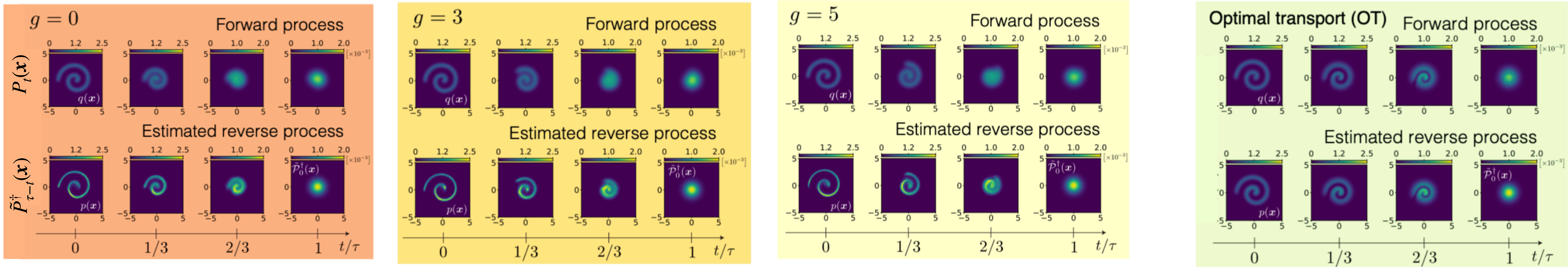
e.g.,) Y. Lipman et al. International Conference on Learning Representations (2022)



Numerical example: 2D Swiss roll data

Non-conservativeness (How far from optimal)

Optimal

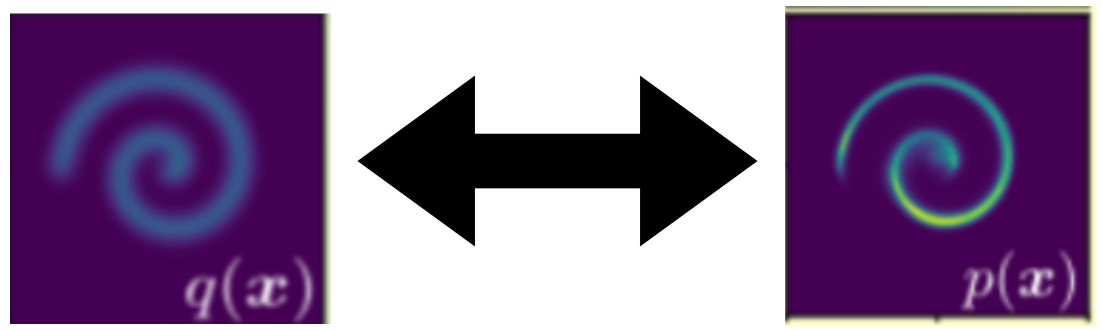


g:Degree of non-conservativeness

The data structure (i.e., a Swiss roll) is well recovered only in the case of optimal transport.

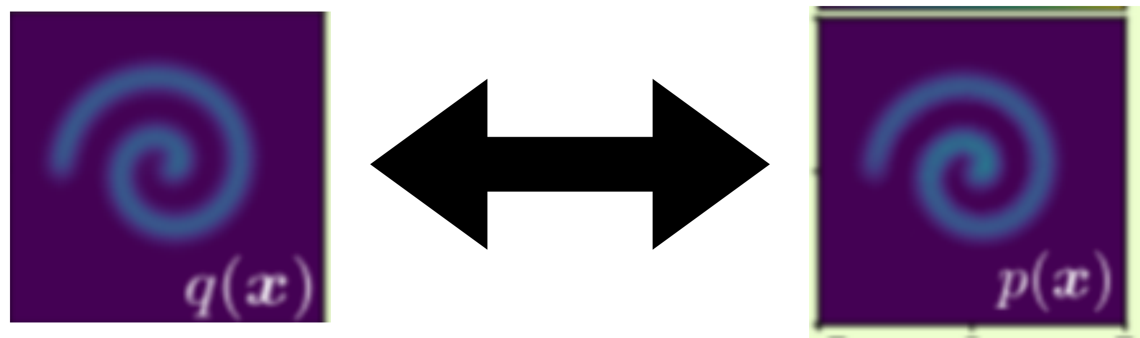
Training data: $q(\boldsymbol{x})$ **Generated data: $p(\boldsymbol{x})$**

$g=5$
(w/ nonconservative force)



Training data: $q(\boldsymbol{x})$ **Generated data: $p(\boldsymbol{x})$**

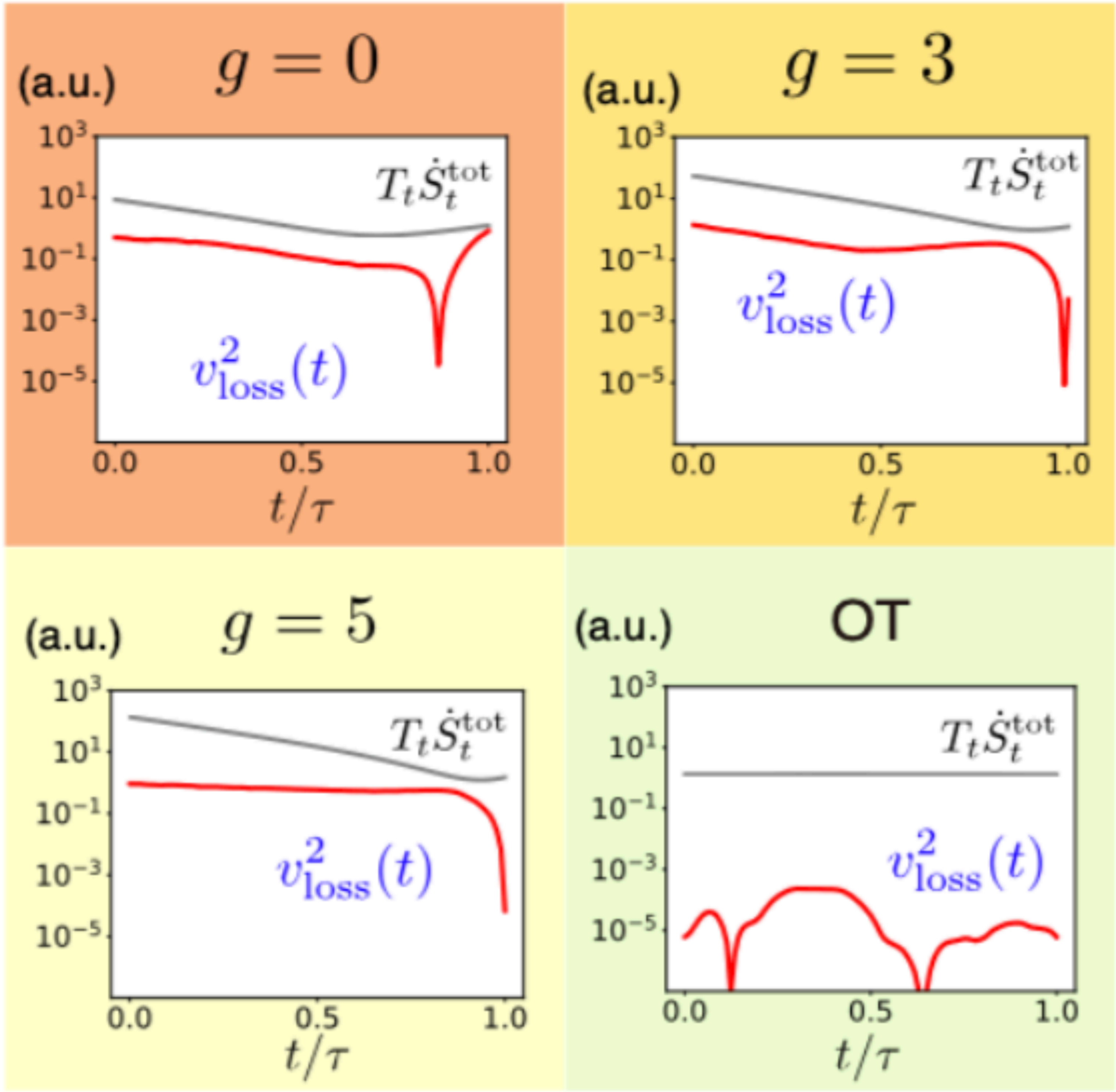
Optimal transport
(w/o nonconservative force)

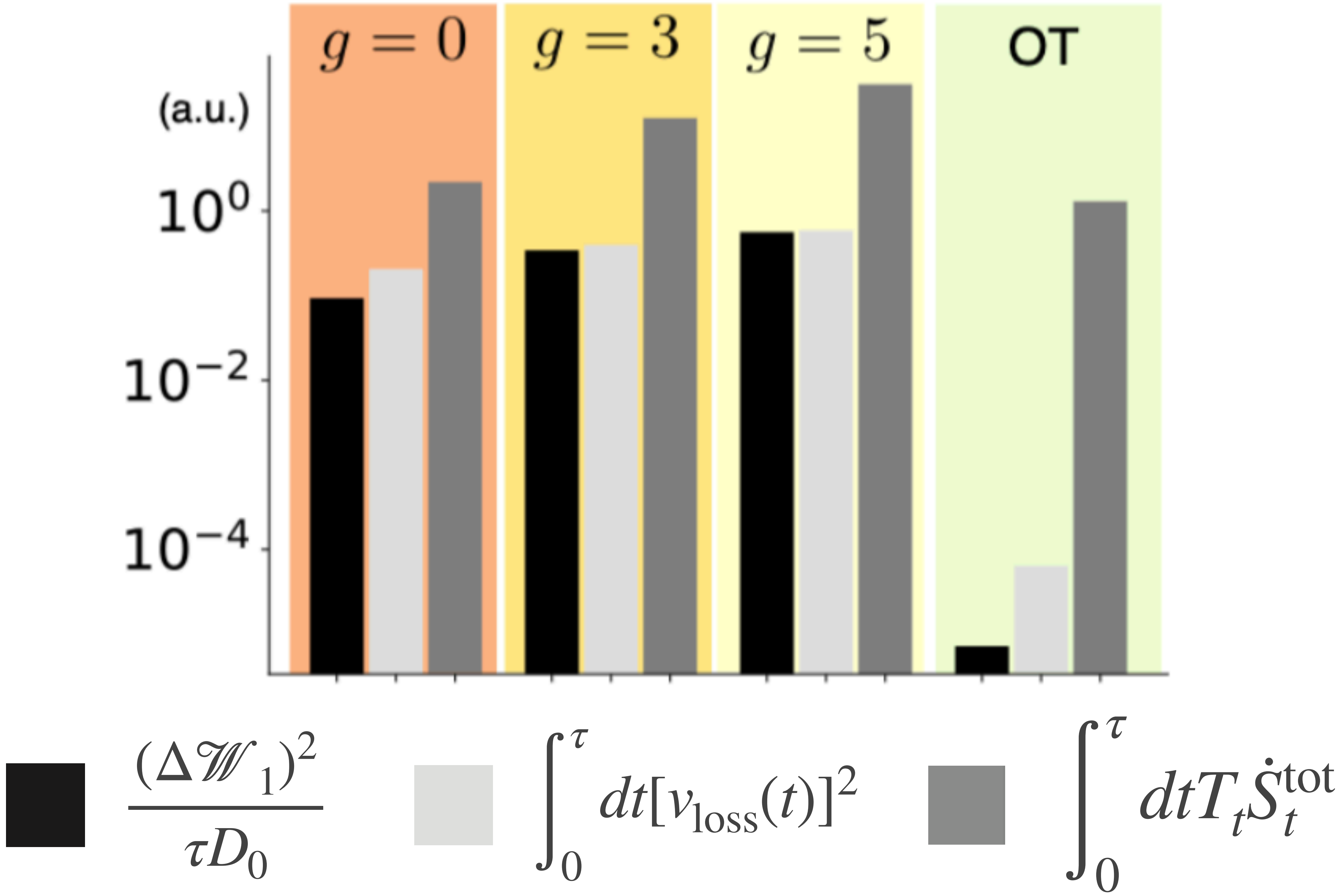


Numerical example: Speed accuracy relations for diffusion models

$$[v_{\mathrm{loss}}(t)]^2 \leq T_t \dot{S}_t^{\mathrm{tot}}$$

$$\frac{(\Delta \mathscr{W}_1)^2}{\tau D_0} \leq \int_0^\tau dt [v_{\mathrm{loss}}(t)]^2 \leq \int_0^\tau dt T_t \dot{S}_t^{\mathrm{tot}}$$



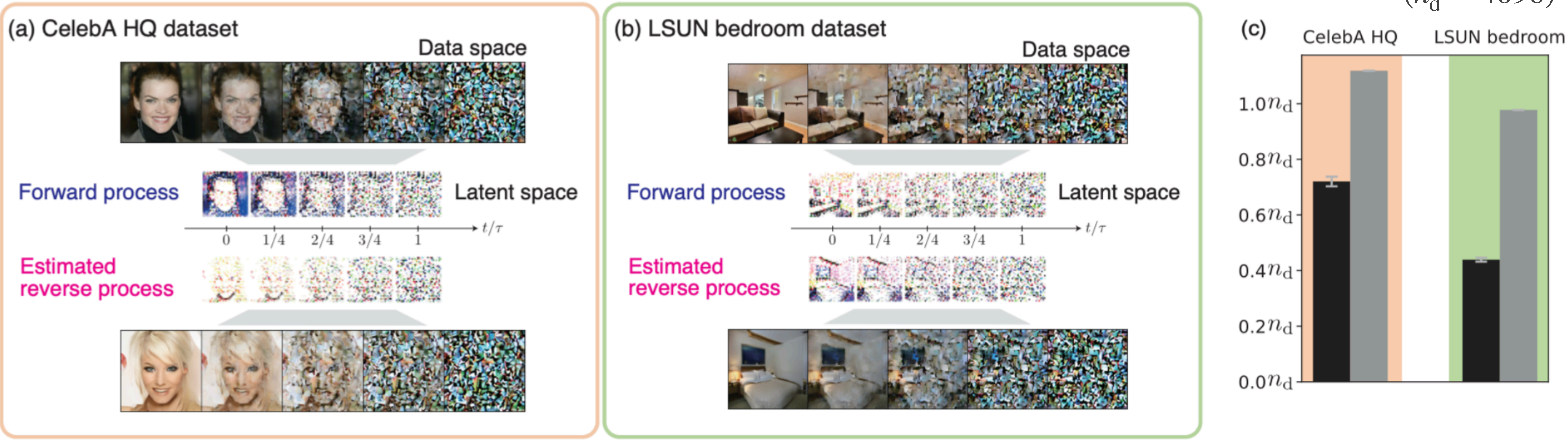


Our results are valid. The optimal transport provides the most robust data generation.

As the degree of non-conservativeness g increases, data generation becomes less robust.

Image data generation: Real-world image dataset in 4096-dimensional latent space

$(n_{\mathrm{d}} = 4096)$



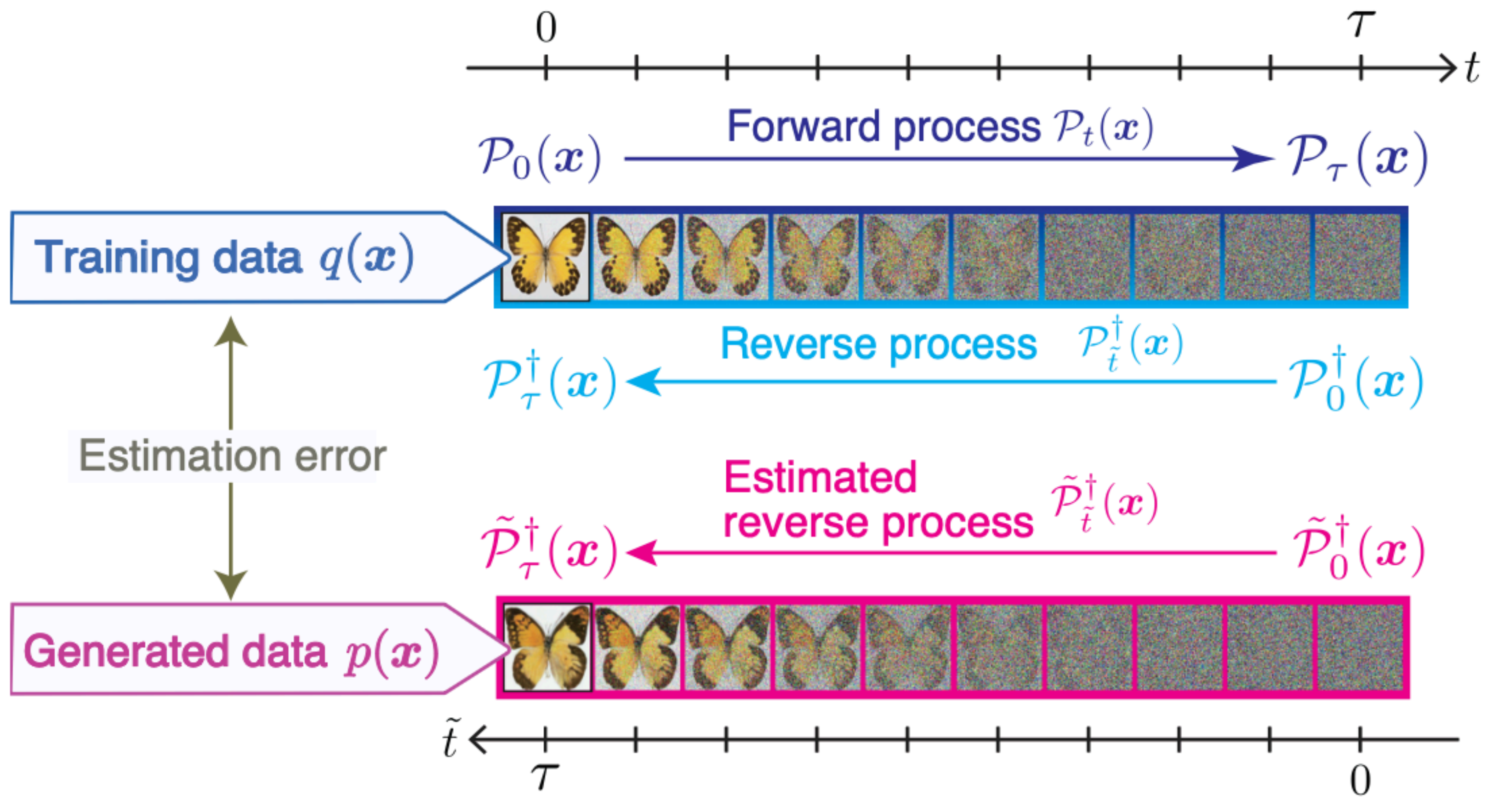
$$\frac{(\Delta \mathscr{W}_1)^2}{\tau D_0} \leq \int_0^\tau dt T_t \dot{S}_t^{\mathrm{tot}}$$

The bound is still computable and valid for practical situations.

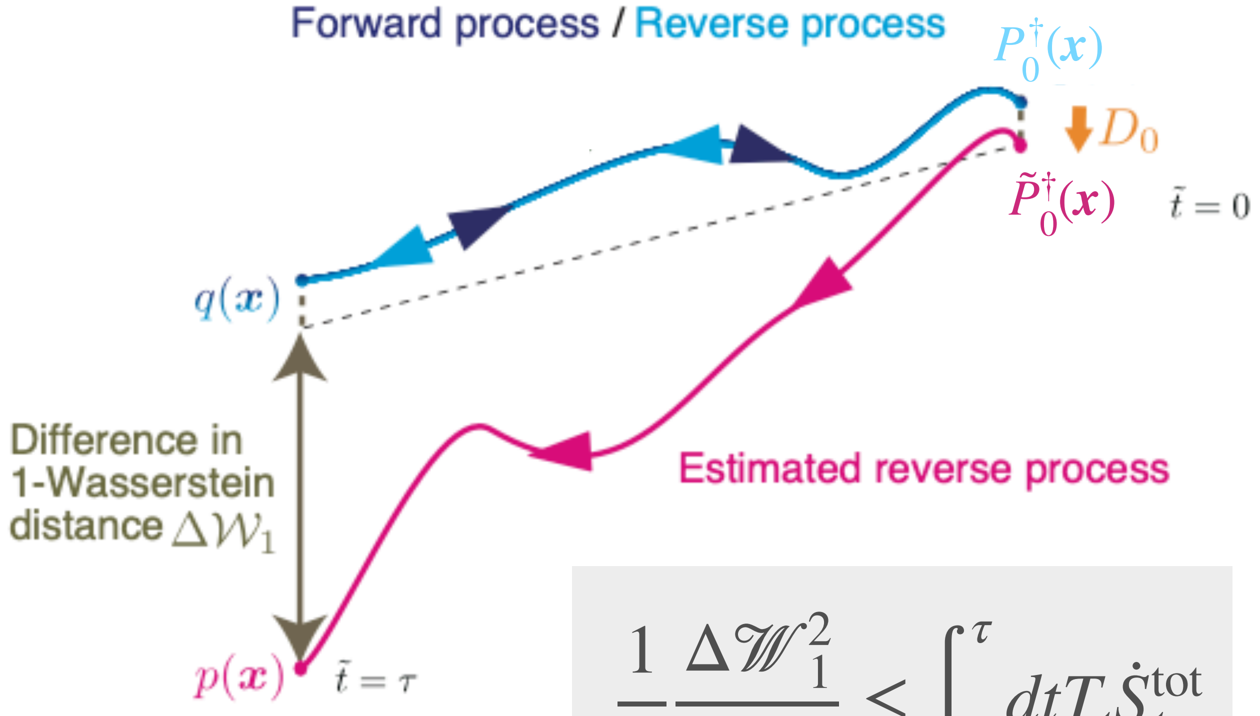
The response $(\Delta \mathscr{W}_1)^2/(\tau D_0)$ has the same order of magnitude as the upper bound $\int_0^\tau dt T_t \dot{S}_t^{\mathrm{tot}}$.

■ $\frac{(\Delta \mathscr{W}_1)^2}{\tau D_0}$ ■ $\int_0^\tau dt T_t \dot{S}_t^{\mathrm{tot}}$

Summary



- We employed stochastic thermodynamics and optimal transport techniques to examine robust data generation in diffusion models.
- We derived the relationship between the robust data generation and the entropy production rate.
- Based on our results, the optimal forward process is given by optimal transport. This result is analogous to the discussion of minimum entropy production.
- Our results remain applicable to image generation using practical diffusion models.



$$\frac{1}{\tau}\frac{\Delta\mathcal{W}_1^2}{D_0} \le \int_0^\tau dt T_t \dot{S}_t^{\mathrm{tot}}$$

For more information and examples, see
K. Ikeda, T. Uda, D. Okanohara & SI *arXiv:2407.04495* (to appear in Physical Review X.)

Collaborators:  **Kotaro Ikeda (UTokyo, B4)**  Tomoya Uda (UTokyo, B4)  Daisuke Okanohara (Preferred Networks Inc.)

Take-home message: Stochastic thermodynamics is useful for generative AI.