# *Variational Optimality of Föllmer Processes*

## *and related design questions in generative diffusions*

Yifan Chen

UCLA Mathematics

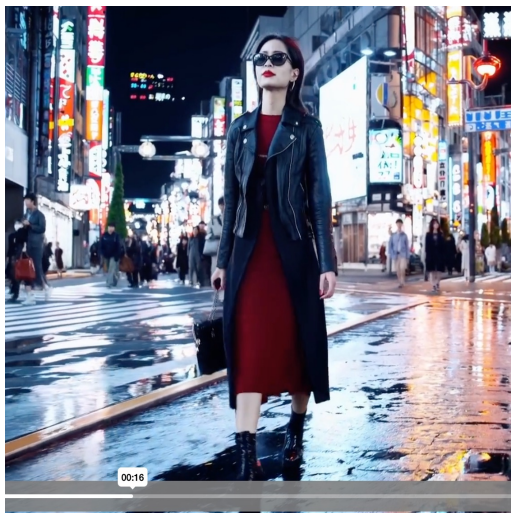
2026

# Success of probabilistic generative modeling techniques

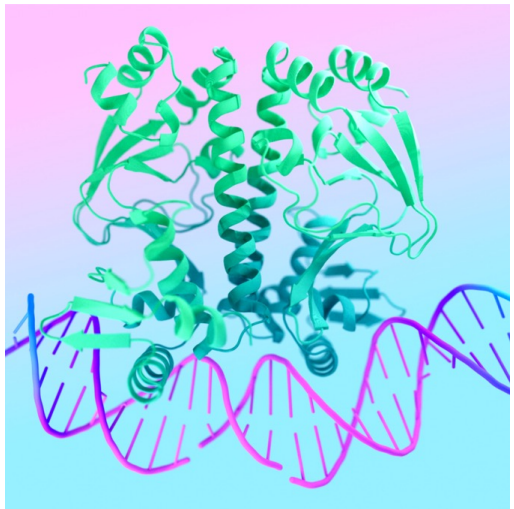
## Generative modeling

Goal: draw new samples from $\pi$, given data $\{x_i\}_{i=1}^N \sim \pi$

DALL·E 3




Sora



Alpha Fold 3

Breakthrough in computer vision and success extended to sciences

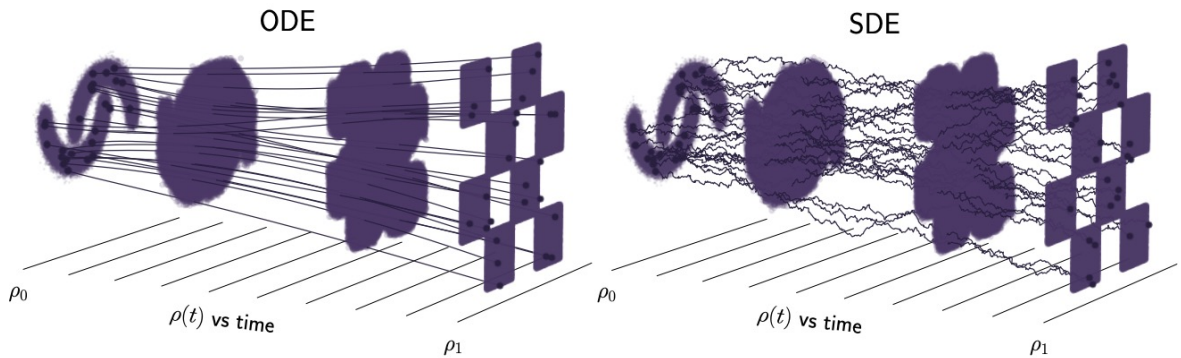
DALL·E 3: `https://openai.com/index/dall-e-3/`
Sora: `https://openai.com/sora/`
Alpha Fold 3: `https://deepmind.google/science/alphafold/`

# Diffusion and flow-based generative modeling: many methodologies

- Time reversal of *prescribed* noising processes

Diffusion models, score-based generative models

- Reproducing *prescribed* marginals in an interpolation path




Flow matching, rectified flow, stochastic interpolants, ...

- Schrödinger bridges, optimal transport, stochastic control, ...

[Sohl-Dickstein et al 2015], [Ho, Jain, Abbeel 2020], [Song et al 2021], [Peluchetti 2021], [De Bortoli et al. 2021], [Liu, Gong, Liu 2022], [Albergo, Vanden-Eijnden, 2022], [Lipman et al 2022], [Albergo, Boffi, Vanden-Eijnden 2023], [Shi et al 2023], etc.

# This talk: connections of these, variational structures, and some implications for applications

We focus on generative diffusions from *a point source* setting

*(generalizable, e.g., to Gaussian base)*

Outline:

1. Probabilistic forecasting motivations
2. Generative diffusions by a prescribed diffusive interpolation
3. Design flexibility of diffusion coefficients
4. Variational optimality: Föllmer processes, time reversal, Schrödinger bridge, and stochastic control interpretation
5. Statistical equivalence of interpolation paths
6. Related numerical design questions of interpolation paths

# Probabilistic forecasting through generative modeling

## A motivating case study: 2d NSE with stochastic forcing

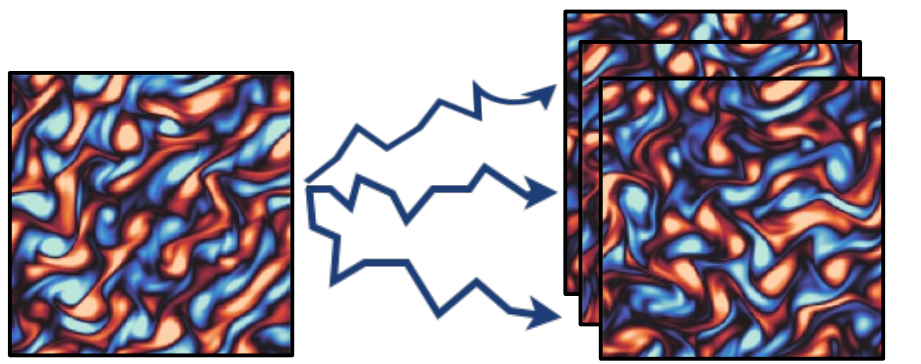
$$\mathrm{d}\omega + v \cdot \nabla\omega \mathrm{d}t = \nu\Delta\omega \mathrm{d}t - \alpha\omega \mathrm{d}t + \epsilon \mathrm{d}\eta \quad \text{on } \mathbb{T}^2$$

- vorticity $\omega$, velocity $v$, and $\mathrm{d}\eta$ is white-in-time random forcing

  Ergodicity: [Hairer, Mattingly, 2006]

**Set-up**: given data pairs $(\omega_t, \omega_{t+\tau})$ at many $t$ under stationarity

**Task**: build a generative model that takes a state $\omega_t$ as input and samples the conditional distribution $\rho^\star(\cdot|\omega_t)$ of $\omega_{t+\tau}|\omega_t$



$x_0$ $\qquad$ $x_1 \sim \rho^\star(\cdot|x_0)$

where we use $x_0 = \omega_t$ and $x_1 = \omega_{t+\tau}$ in the notation

# Construct a generative diffusion from a point source

[**Chen**, Goldstein, Hua, Albergo, Boffi, Vanden-Eijnden 2024]

**Goal**: construct $b_s, g_s$ such that

$$\mathrm{d}X_s = b_s(X_s, x_0)\mathrm{d}s + g_s \mathrm{d}W_s,\ X_s = x_0, \text{ satisfying } X_1 \sim \rho^\star(\cdot|x_0)$$

Let $I_s = \alpha_s x_0 + \beta_s x_1 + \sigma_s W_s$ specify a diffusive interpolation path

- $x_1 \sim \rho^\star(\cdot|x_0)$, and $W$ is a Brownian motion with $W \perp (x_0, x_1)$
- $\alpha_0 = \beta_1 = 1$ and $\alpha_1 = \beta_0 = \sigma_1 = 0$ so that $I_0 = x_0, I_1 = x_1$

**Theorem**: we can take

$$b_s(x, x_0) = \mathbb{E}[\dot\alpha_s x_0 + \dot\beta_s x_1 + \dot\sigma_s W_s | I_s = x, x_0], \quad g_s = \sigma_s$$

so that $\mathrm{Law}(X_s) = \mathrm{Law}(I_s|x_0)$. In particular $X_1 \sim \rho^\star(\cdot|x_0)$

- Why? Itô's formula: $\mathrm{d}I_s = (\dot\alpha_s x_0 + \dot\beta_s x_1 + \dot\sigma_s W_s)\mathrm{d}s + \sigma_s \mathrm{d}W_s$
- Markovization by taking conditional expectation

  Gyöngy's Mimicking lemma [Gyöngy 1986]

# Regression for estimating the drift from data

The drift $b_s(x, x_0) = \mathbb{E}[\dot{\alpha}_s x_0 + \dot{\beta}_s x_1 + \dot{\sigma}_s W_s | I_s = x, x_0]$

- $b_s(x, x_0)$ is the unique minimizer of

$$L_b[\hat{b}_s] = \int_0^1 \mathbb{E}\big[|\hat{b}_s(I_s, x_0) - \dot{\alpha}_s x_0 - \dot{\beta}_s x_1 - \dot{\sigma}_s W_s|^2\big]\mathrm{d}s$$

  given sampled data $(x_0, x_1)$ we can evaluate $L_b$
- Can replace $W_s$ by $\sqrt{s}z$ (simulation-free)
- Parametrize $\hat{b}_s$ and optimize $L_b$ to get final $\hat{b}_s$
- Integrate the SDE with $\hat{b}_s$ for probabilistic forecasting

Note: Tweedie's formula $\nabla \log \rho_s(x|x_0) = -\frac{1}{s\sigma_s}\mathbb{E}[W_s | I_s = x, x_0]$
So score $\nabla \log \rho_s(x|x_0) = A_s b_s(x, x_0) + c_s$ affine-relates to $b_s$

# Experiments for forecasting 2D stochastically forced NSE

Generation results with accurate conditional statistics

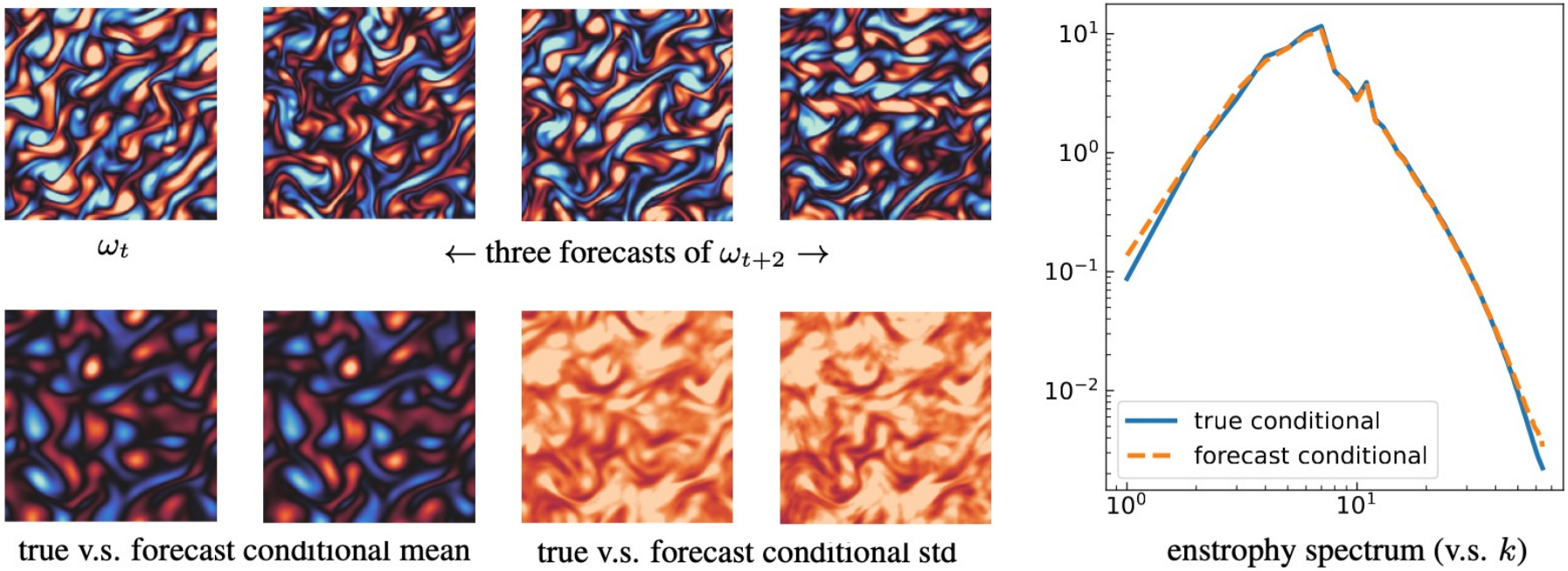



Figure: Lag $\tau = 2$ (autocorrelation $10\%$). Resolution $128 \times 128$, using $20K$ data pairs for training a 2M-parameter-Unet model of the drift

Methodology proven effective across scientific domains, such as weather prediction [Kossaif et al., NVIDIA 2026]

# Design flexibility of diffusion coefficients

**Theorem**: It holds that Law$(X_s)$ = Law$(X_s^g)$ for

$$\mathrm{d}X_s^g = b_s^g(X_s^g, x_0)\mathrm{d}s + g_s \mathrm{d}W_s$$

with $b_s^g(x, x_0) = b_s(x, x_0) + \frac{1}{2}(g_s^2 - \sigma_s^2)\nabla \log \rho_s(x|x_0)$

- Fact due to Fokker-Planck equations and $\nabla \cdot (\rho \nabla \log \rho) = \Delta \rho$
- With estimated $\widehat{\text{score}}$ obtained using $\hat{b}_s$, we get

  New "learned" drift: $\hat{b}_s^g = \hat{b}_s + \frac{1}{2}(g_s^2 - \sigma_s^2)\widehat{\text{score}}$

*Need subtle conditions on $g_s$ for non-singularity of the drift

see discussions in [**Chen**, Vanden-Eijnden 2026]

Variational question: principled selection of $g$?

# KL divergence over path measures to optimize diffusion coefficients

**Theorem**: Let $\mathbb{P}^{X^g}$ and $\mathbb{P}^{\hat{X}^g}$ denote the path measures of

- the truth SDE solution $X^g = (X_s^g)_{s\in[0,1]}$ with drift $b^g$
- the approximation $\hat{X}^g = (\hat{X}_s^g)_{s\in[0,1]}$ with learned $\hat{b}^g$

Then, the path-level KL optimization

$$\min_g \mathrm{KL}[\mathbb{P}^{X^g} \| \mathbb{P}^{\hat{X}^g}]$$

has an explicit solution $g = g^{\mathsf{F}}$ with

$$g_s^{\mathsf{F}} = \left| 2s\sigma_s^2 \frac{\mathrm{d}}{\mathrm{d}s} \log \frac{\beta_s}{\sqrt{s}\sigma_s} \right|^{1/2}$$

Note: $\frac{\beta_s}{\sqrt{s}\sigma_s}$ is

$\sim$ "signal-to-noise ratio"

since by definition

$$I_s = \alpha_s x_0 + \beta_s x_1 + \sigma_s W_s$$

Empirically $g^{\mathsf{F}}$ observed better estimation of certain observables

Related derivations in Gauss-source diffusions and memoryless schedule in fine-tuning diffusions [Ma et al. 2024] [Domingo-Enric et al. 2024]

# The special role of the optimal $g^{\mathrm{F}}$

Let $x_0 = 0$ for simplicity of presentation

**Theorem**: The optimal $X^{\mathrm{F}} := X^{g^{\mathrm{F}}}$ is an Föllmer process

- Solution to Schrödinger bridge when one endpoint is point mass

$$X^{\mathrm{F}} = \operatorname*{argmin}_{X} \mathrm{KL}[\mathbb{P}^{X} \| \mathbb{P}^{X^{\mathrm{prior}}}] \ \text{ s.t. } \ X_1 \sim \rho^\star$$

where $X^{\mathrm{prior}} = Y$ satisfies

$$\mathrm{d}Y_s = a_s Y_s \mathrm{d}s + g_s^{\mathrm{F}} \mathrm{d}W_s, \quad Y_0 = 0, \quad a_s = \frac{\mathrm{d}}{\mathrm{d}s} \log \frac{\beta_s^2 + s\sigma_s^2}{\beta_s}$$

Chain rule of KL divergence

$$\mathrm{KL}[\mathbb{P}^{X} \| \mathbb{P}^{Y}] = \int \mathrm{KL}[\mathbb{P}^{X|X_1=x} \| \mathbb{P}^{Y|Y_1=x}] \rho^\star(x) \mathrm{d}x + \mathrm{KL}[\mathbb{P}^{X_1} \| \mathbb{P}^{Y_1}]$$

Optimal $X$ obtained by time reversal of $Y_s$ while initializing with $\rho^\star$

**Interpretation:** $g^{\mathrm{F}}$ is a "Bayes"/control/time-reversal solution

# Discussion of the result

Comparisons of modeling perspectives

**"Bayes"/control/time-reversal perspective**: specify and parametrize a prior process (or its time reversal)

**Interpolation perspective**: specify and parametrize a marginal interpolation path

For linear prior process and interpolation, they are equivalent in modeling expressivity, by allowing changing diffusion coefficients

Parameterizing interpolation paths directly leads to explicit simulation-free regression for learning drifts

see other technical discussions in [**Chen**, Vanden-Eijnden 2026]

# Statistical efficiency of a particular interpolation path?

**Theorem**: After KL-optimal tuning of the diffusion coefficient, the minimized path-space KL divergence takes the form

$$\mathrm{KL}^\star = 2\int_0^\infty r\,\mathbb{E}\Big[\big|\nabla \log q_r(x_\star + rz) - \hat{s}_r(x_\star + rz)\big|^2\Big]\,\mathrm{d}r$$

where $q_r$ is the density of $\rho^\star * \mathcal{N}(0, r^2\mathrm{I})$ and $\hat{s}_r$ is estimated score at noise level $r$, obtained by affine transformation of estimated drift

- In particular, $\mathrm{KL}^\star$ is independent of scalar schedules $(\beta_t, \sigma_t)$
- All are statistically **the same as** score-based diffusions, by re-parametrization and tuning diffusion coefficients
- To break equivalence, necessary to go beyond scalar schedules in the interpolation paths

# Numerical efficiency and design questions

Due to equivalence, let us focus on ODE models from Gaussian source

Solving flow matching ODE $\dot{X}_t = b_t(X_t), 0 \le t \le 1$ by RK4 scheme

$$b_t(x) = \mathbb{E}[x_1 - z \mid (1-t)z + tx_1 = x]$$

**Setting:** $z$ white noise and $x_1 \sim \mathrm{N}(0, C_1)$ with $C_1 = (-\Delta + I)^{-3}$

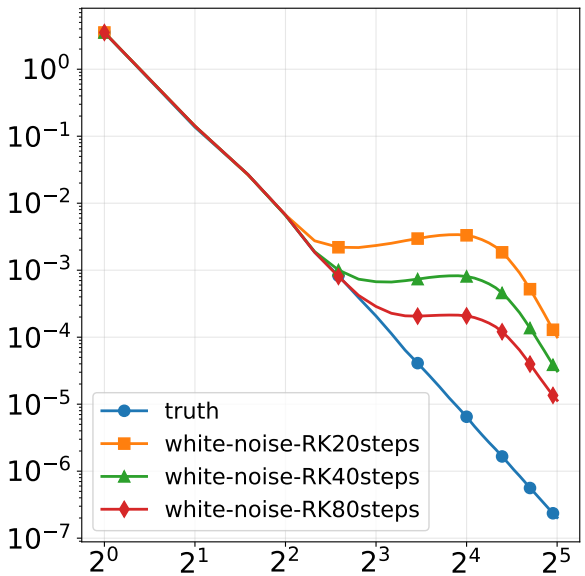



Figure: Fourier spectra of truth and generated with various RK4 steps

- Much more costs when resolution increases
- Many advanced integration methods can help. Fundamentally, the numerical challenge remains when resolution is very fine

# Optimal transport approach to improve straightness

## Minimal kinetic energy

$$\min_{b_t} \ \mathbb{E}[\|b_t(X_t)\|_2^2]$$
$$\text{s.t.} \ \ \dot{X}_t = b_t(X_t), X_0 \sim \mathrm{N}(0, \mathrm{I}), X_1 \sim \rho^*$$

- Benamou-Brenier formula [Benamou, Brenier 2000]
- Trajectories are straight lines: one step integration is exact
- However, $b_t(x)$ can be spatially highly irregular

  [Tsimpos, Ren, Zech, Marzouk 2025]

Widely pursued in generative models by rectification, mini-batch coupling, etc. [Liu, Gong, Liu 2022], [Pooladian et al. 2023], etc.

Entropy regularized OT (Schrödinger's bridges)

Efficient algorithm in generative modeling: [Bortoli, Thornton, Heng, Doucet 2021] [Shi, Bortoli, Campbell, Doucet 2023] [Pooladian, Niles-Weed 2024], etc.

# Alternative approach to improve numerical integration efficiency

**Minimal Lipschitz energy** [**Chen**, Vanden-Eijnden, Xu 2025]

$$\min_{b_t} \ \mathbb{E}[\|\nabla b_t(X_t)\|_2^2]$$
$$\text{s.t.} \ \ \dot{X}_t = b_t(X_t), X_0 \sim \mathrm{N}(0, \mathrm{I}), X_1 \sim \rho^*$$

Constrained optimization in the class of dynamics

$$b_t(x) = \mathbb{E}[\dot{I}_t | I_t = x], \quad I_t = \alpha_t z + \beta_t x_1$$

- Noise $z \sim \mathrm{N}(0, \mathrm{I}) \perp x_1 \sim \rho^*$ the data distribution
  $\alpha_0 = \beta_1 = 1, \alpha_1 = \beta_0 = 1$ are schedules to be optimized
- For any such $\alpha_t, \beta_t$, using the drift $b_t$ yields $X_1 \sim \rho^*$ [Gyöngy 1986]

# Analytically optimized schedules and exponentially improved Lipschitz

**Gaussian targets:** $x_1 \sim \mathrm{N}(0, C) \perp z \sim \mathrm{N}(0, \mathrm{I})$ in $d$ dims. Let eigenvalues of $C$ be $1 \geq \lambda^{(1)} \geq \ldots \geq \lambda^{(d)} > 0$. Denote $M^\star = 1/\lambda^{(d)}$

**Theorem:** For the common linear schedule $\alpha_t = 1 - t, \beta_t = t$

$$\int_0^1 \mathbb{E}[\|\nabla b_t(X_t)\|_2^2]\mathrm{d}t = \Omega(\sqrt{M^\star}), \quad \max_{t,x} \|\nabla b_t(x)\|_2 = \Omega(M^\star)$$

If we optimize Lipschitz energy over all possible linear stochastic interpolants $I_t$ with scalar schedules, then

$$\alpha_t = \sqrt{\frac{(M^\star)^{1-t} - 1}{M^\star - 1}}, \beta_t = \sqrt{\frac{M^\star - (M^\star)^{1-t}}{M^\star - 1}}$$

For the optimal solution, $\|\nabla b_t(x)\|_2 = \frac{1}{2} \log M^\star$ for any $t, x$

- Other analytic results on Gaussian mixtures using Euler-Lagrange equation

# Optimized schedule: performance for Gaussian measures

Target $\rho^* = \mathrm{N}(0, C_1)$, where $C_1 = (-\Delta + I)^{-3}$. Noise is white

- Discretize on $N \times N$ grid points
- Compare optimized to standard schedule $\alpha_t = 1 - t, \beta_t = t$

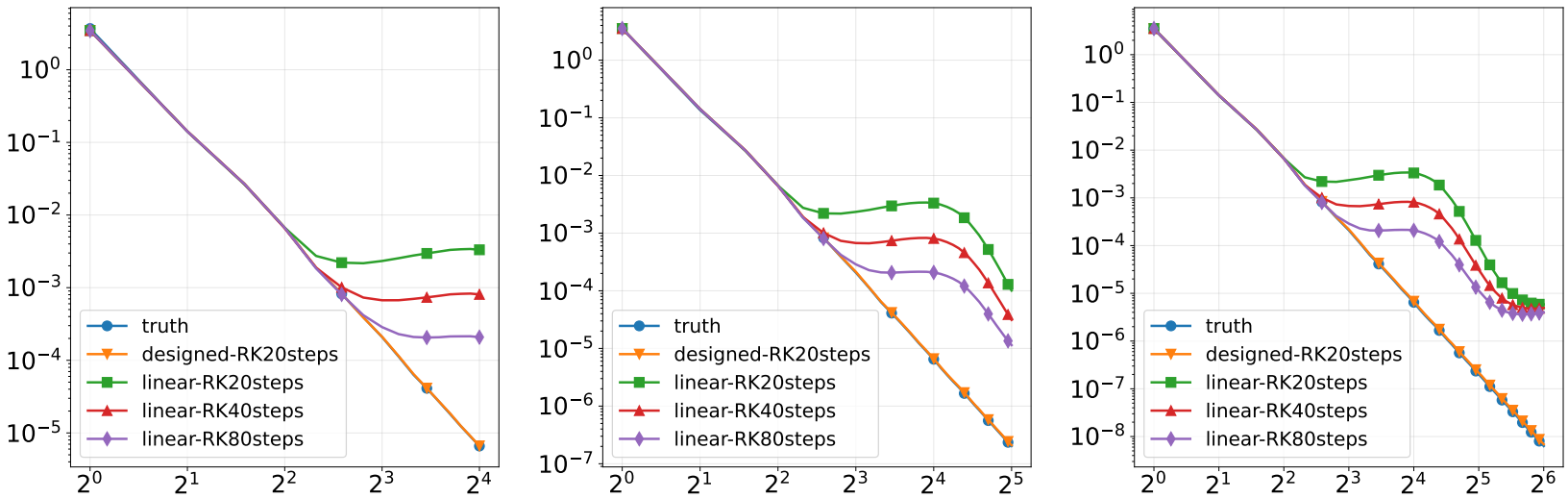



Figure: Gaussian measure example. Linear schedule versus optimized schedules. Left: $32 \times 32$; middle: $64 \times 64$; right: $128 \times 128$

Resolution robust performance with the same integration steps

Note: using colored/designed noise works well too

# Performance for invariant distribution to stochastic Allen-Cahn

Target $\rho^*(u) \propto \exp\left(-\int_0^1 \frac{1}{2}(\partial_x u(x))^2 + (1-u^2(x))^2 \mathrm{d}x\right)$

- Invariant distribution to stochastic Allen-Cahn
- Discretize on $N$ grid points. Same setting

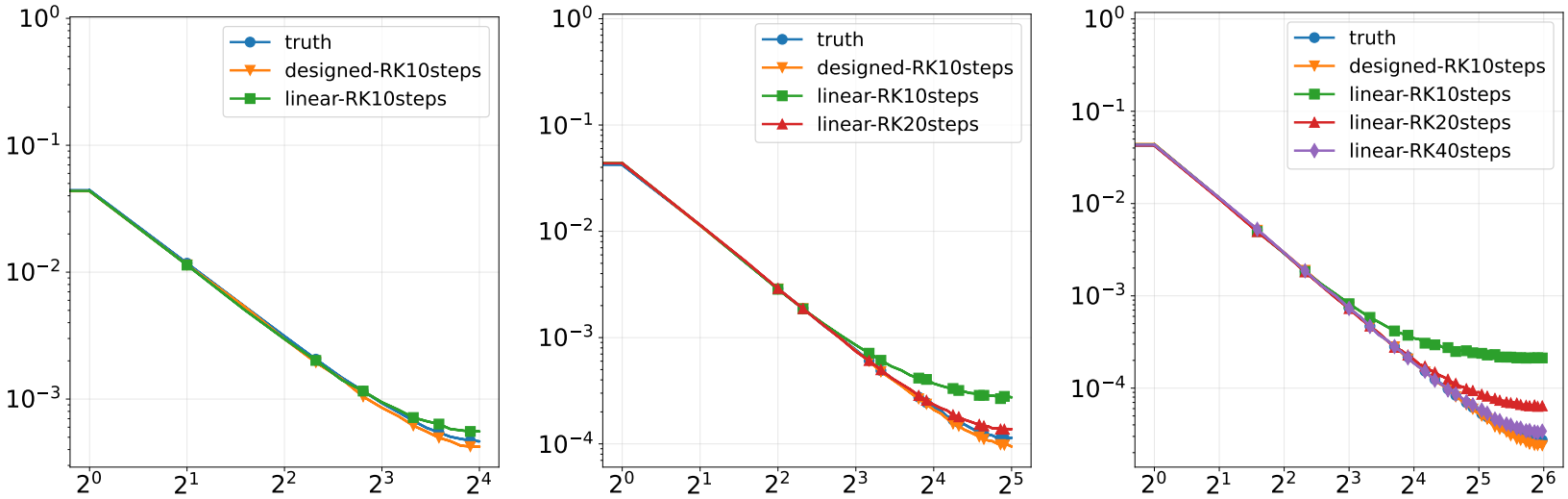



Figure: Stochastic Allen-Cahn example. Linear schedule versus optimized schedules. Left: $N = 32$; middle: $N = 64$; right: $N = 128$

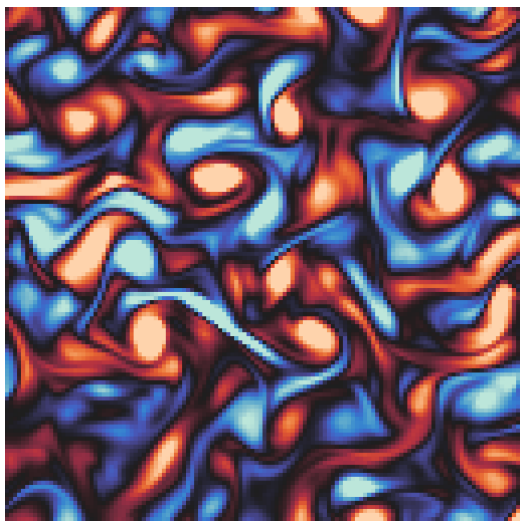
All experiments are done using 2M-parameter-Unet to train $b_t$

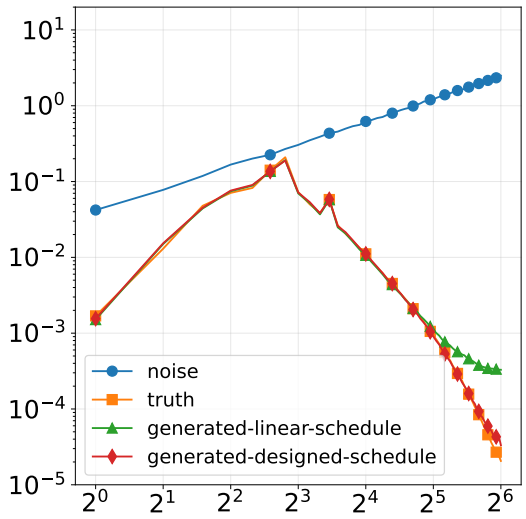
Again robust performance with the same integration steps

## Case study: 2d NSE with stochastic forcing

$$\mathrm{d}\omega + v \cdot \nabla \omega \mathrm{d}t = \nu \Delta \omega \mathrm{d}t - \alpha \omega \mathrm{d}t + \epsilon \mathrm{d}\eta \quad \text{on } \mathbb{T}^2$$

- vorticity $\omega$, velocity $v$, and $\mathrm{d}\eta$ forcing
- $\nu = 10^{-3}$, $\mathrm{d}\eta$ random forcing acts on a few Fourier modes





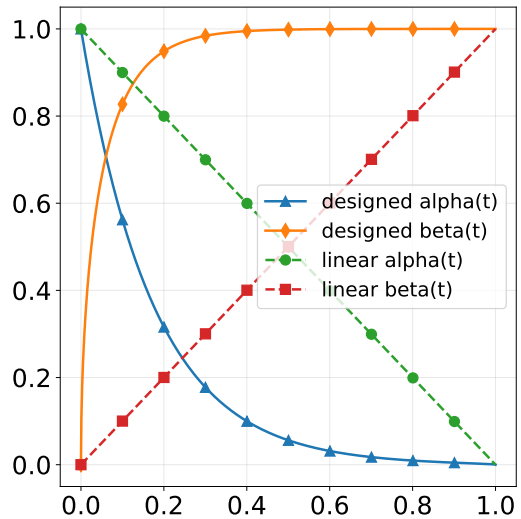




Figure: We use white noise and 10 RK4 integration steps. $128 \times 128$

- Left: generated samples w/ optimized schedule
- Middle: enstrophy spectra of truth, noise, and generations
- Right: linear versus optimized schedule ($M^* = 10^5$)

# Summary

- Generative diffusions from a point source, via reproducing time-marginals of a prescribed diffusive interpolation path
- Variational structure: optimal diffusion coefficients under path KL lead to Föllmer process, which can be constructed via time reversal from the target to the point source. Interpretation as Bayes, Schrödinger bridge, and stochastic control solutions
- These “scalar linear interpolation” models are equivalent in expressivity, and (path-level) statistical efficiency, to standard score based diffusion models. Breaking this equivalence requires design beyond “linear scalar interpolation”
- Numerical efficiency differs in “linear scalar interpolation”. Lipschitz criteria can design more regular generative diffusions