Lipschitz Guided Dynamical Transport

for generative flows of ill-conditioned distributions

Yifan Chen

UCLA Mathematics

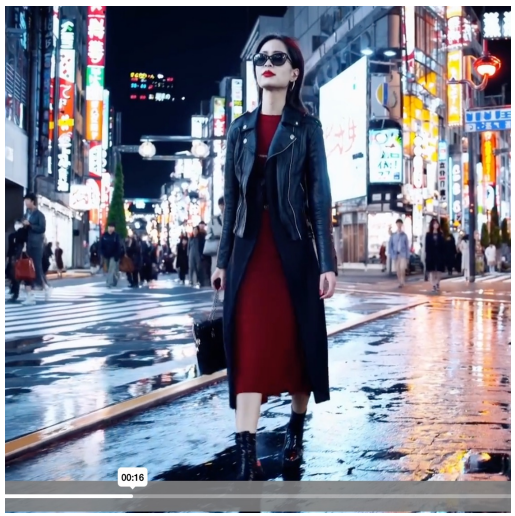
Joint work with Eric Vanden-Eijnden (NYU), Jiawei Xu (Maryland)

Success of generative modeling

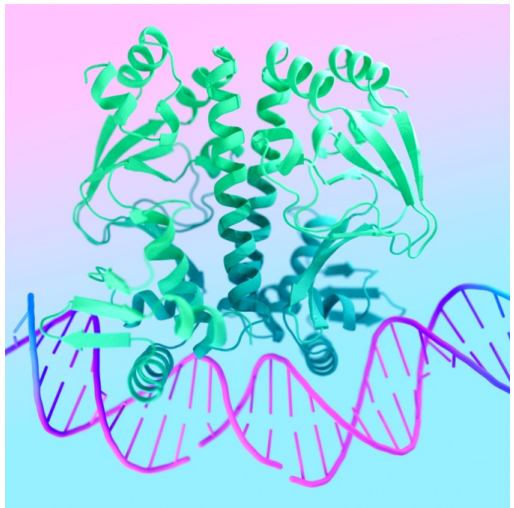
Generative modeling

Goal: draw new samples from π, 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/`

Challenge: field data with a wide range of Fourier spectra

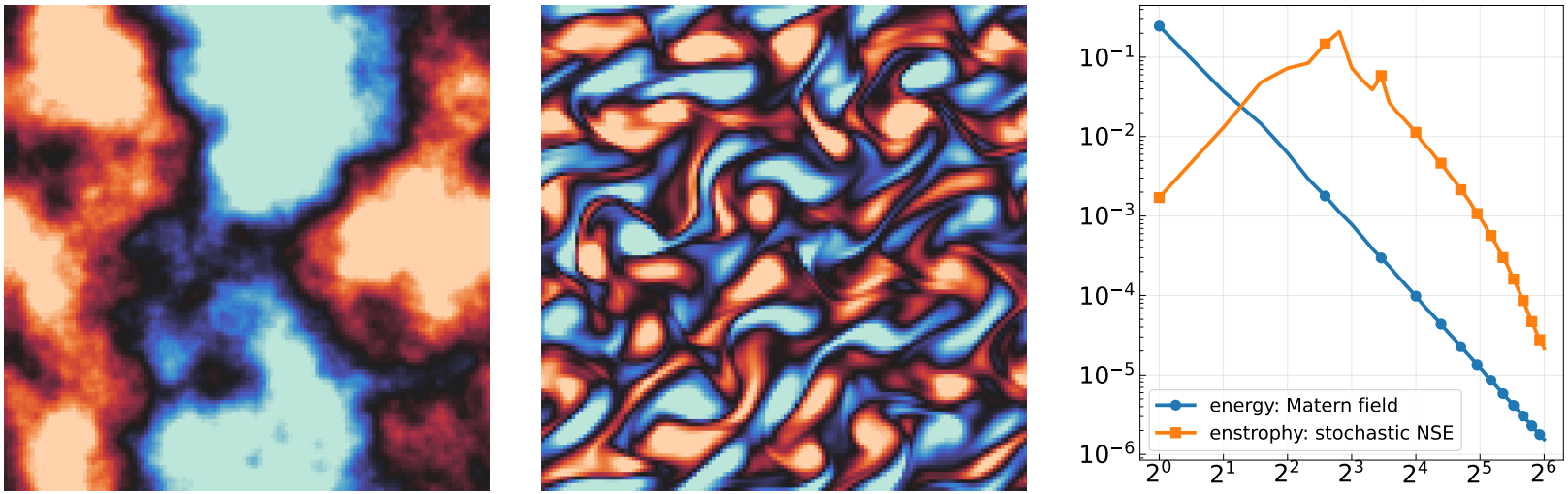


Figure: Examples of data samples from Matérn Gaussian processes (left panel) and invariant measure of stochastically forced Navier-Stokes (middle panel). The right panel shows their energy and enstrophy spectra

- Precise fine scale accuracy is numerically challenging
- W_2 criteria cannot guarantee; KL struggle in optimization
- Existing function space framework often aims for a different goal of coarse scale stability (stable under resolution refinement)

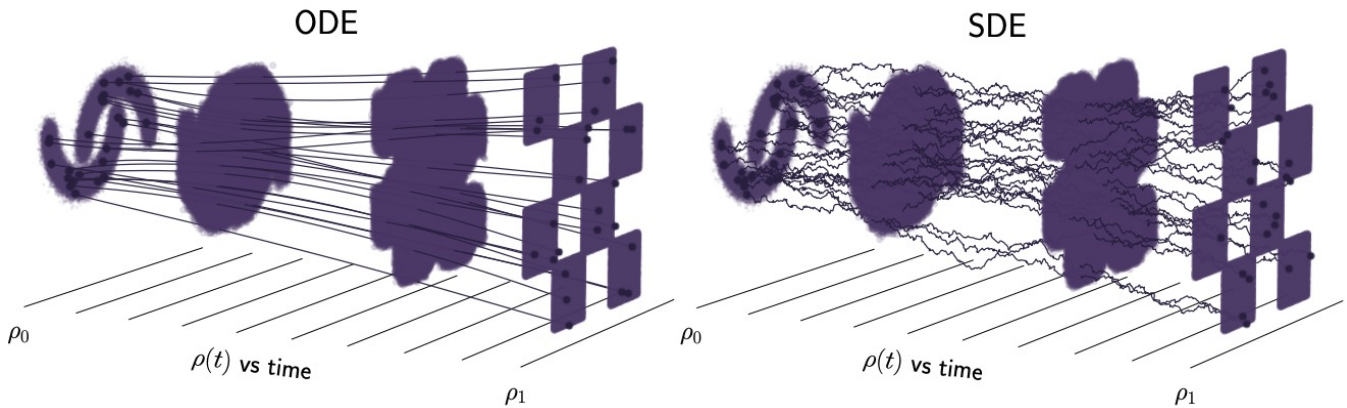
Function space generative models [Lim et al 2023], [Hagemann, Ruthotto, Steidl, Yang 2023], [Pidstrigach, Marzouk, Reich, Wang 2023], [Kerrigan, Migliorini, Smyth 2023], etc.

Wavelets and multiscale generative models [Guth, Coste, Bortoli, Mallat 2022], etc.

State of the art methodology: flow and diffusion

Recent advances in generative modeling driven by building dynamics of measures that iteratively refine the generation to the desired

Diffusion models, score-based generative models



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

[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.

Simple summary of methodology in one slide

- Corruption path via interpolation between data and noise

$$I_t = \alpha_t z + \beta_t x_1, \quad \alpha_0 = \beta_1 = 1, \alpha_1 = \beta_0 = 1$$

where, noise $z \sim \mathrm{N}(0, \mathrm{I}) \perp x_1 \sim \rho^*$ the data distribution

- Generation dynamics via numerically solving

$$\mathrm{d}X_t = b_t(X_t)\mathrm{d}t, \quad b_t(x) = \mathbb{E}[\dot{I}_t | I_t = x]$$

Thm: for such b_t, it holds $X_1 \sim \rho^*$ the target [Gyöngy 1986]

- b_t can be learned from data using the objective

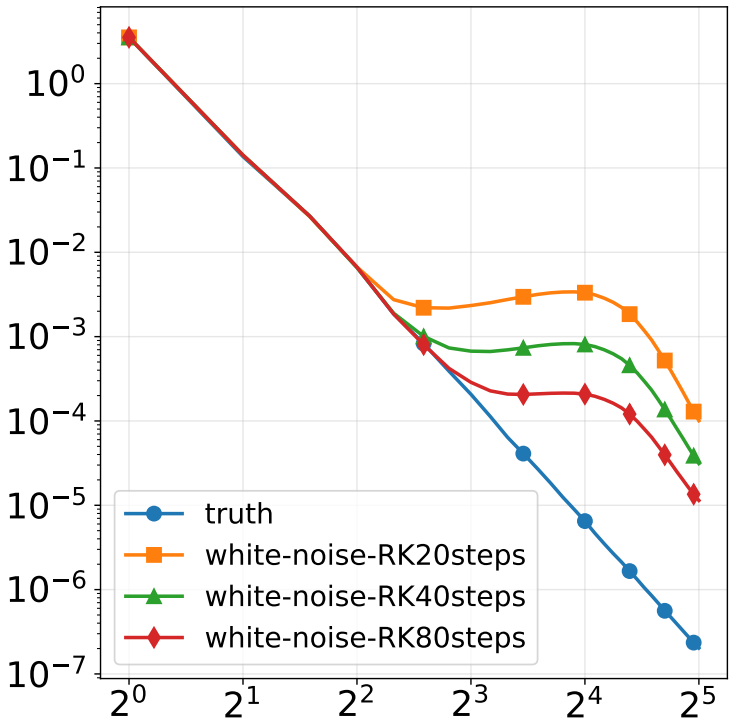
$$\min_{\hat{b}} L(\hat{b}) = \int_0^1 \mathbb{E}[\|\hat{b}_t(I_t) - \dot{I}_t\|_2^2] \, \mathrm{d}t$$

where the expectation is replaced by empirical averages

[Liu, Gong, Liu 2022], [Albergo, Vanden-Eijnden, 2022], [Lipman, Chen, Ben-Hamu, Nickel, Le 2022], [Albergo, Boffi, Vanden-Eijnden 2023], etc.

Gaussian measure example

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



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

Optimal transport approach and its numerical challenge

Minimal kinetic energy in optimal transport approaches

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

- 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 discussed and pursued in generative models [Liu, Gong, Liu 2022], [Albergo, Vanden-Eijnden, 2022], etc.

Entropy regularized OT (a.k.a. Schrödinger's bridges) [Léonard 2014]

Efficient algorithm in generative modeling: [Bortoli, Thornton, Heng, Doucet 2021] [Shi, Bortoli, Campbell, Doucet 2023], [**Chen**, Goldstein, Hua, Albergo, Boffi, Vanden-Eijnden 2024], [Pooladian, Niles-Weed 2024], etc.

How about dynamical transport with minimal Lipschitz energy?

Minimal kinetic energy in optimal transport approaches

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

- Trajectories are straight lines: one step integration is exact
- However, $b_t(x)$ can be spatially highly irregular

Minimal Lipschitz energy

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

- Directly lead to desired b_t for numerical integration
- Hard to solve optimal b_t in general

Searching for minimal Lipschitz energy in linear stochastic interpolation

Practical strategy: selecting noise z and α_t, β_t in $I_t = \alpha_t z + \beta_t x_1$ to make $b_t(x) = \mathbb{E}[\dot{I}_t | I_t = x]$ has (near-)minimal energy in this class

What would be *permissible* choices of noise?

Searching for minimal Lipschitz energy in linear stochastic interpolation

Practical strategy: selecting noise z and α_t, β_t in $I_t = \alpha_t z + \beta_t x_1$ to make $b_t(x) = \mathbb{E}[\dot{I}_t | I_t = x]$ has (near-)minimal energy in this class

What would be *permissible* choices of noise?

Gaussian: Consider $z \sim \mathrm{N}(0, C_0)$ and $x_1 \sim \mathrm{N}(0, C_1)$ are drawn from Gaussian measures supported on Hilbert space H and $z \perp x_1$

- Let $C_0 = \sigma_0^2(-\Delta + \tau_0^2 I)^{-s_0}$ and $C_1 = \sigma_1^2(-\Delta + \tau_1^2 I)^{-s_1}$
- **Thm:** $b_t, t < 1$ is bounded and Lipschitz if and only if $s_0 \leq s_1$

General: Consider $z \sim \mathrm{N}(0, C_0)$ and $x_1 \sim \rho^*$ compactly supported (or smoothed version) in the Cameron-Martin space of the noise

- **Thm:** $b_t, t < 1$ is bounded and Lipschitz

Searching for minimal Lipschitz energy in linear stochastic interpolation

Practical strategy: selecting noise z and α_t, β_t in $I_t = \alpha_t z + \beta_t x_1$ to make $b_t(x) = \mathbb{E}[\dot{I}_t | I_t = x]$ has (near-)minimal energy in this class

What would be *permissible* choices of noise?

Gaussian: Consider $z \sim \mathrm{N}(0, C_0)$ and $x_1 \sim \mathrm{N}(0, C_1)$ are drawn from Gaussian measures supported on Hilbert space H and $z \perp x_1$

- Let $C_0 = \sigma_0^2(-\Delta + \tau_0^2 I)^{-s_0}$ and $C_1 = \sigma_1^2(-\Delta + \tau_1^2 I)^{-s_1}$
- **Thm:** $b_t, t < 1$ is bounded and Lipschitz if and only if $s_0 \leq s_1$

General: Consider $z \sim \mathrm{N}(0, C_0)$ and $x_1 \sim \rho^*$ compactly supported (or smoothed version) in the Cameron-Martin space of the noise

- **Thm:** $b_t, t < 1$ is bounded and Lipschitz

Wellposedness: noise should be rougher, or at least as rough as, data

Matching smoothness (spectrum noise) works for Gaussian

Target $\rho^* = \mathrm{N}(0, C_1)$, where $C_1 = \sigma_1^2(-\Delta + \tau_1^2 I)^{-s_1}, s_1 = 3$

- Discretize on $N \times N$ grid points
- Choose noise to be either white or spectrum noise
- Standard schedule $\alpha_t = 1 - t, \beta_t = t$

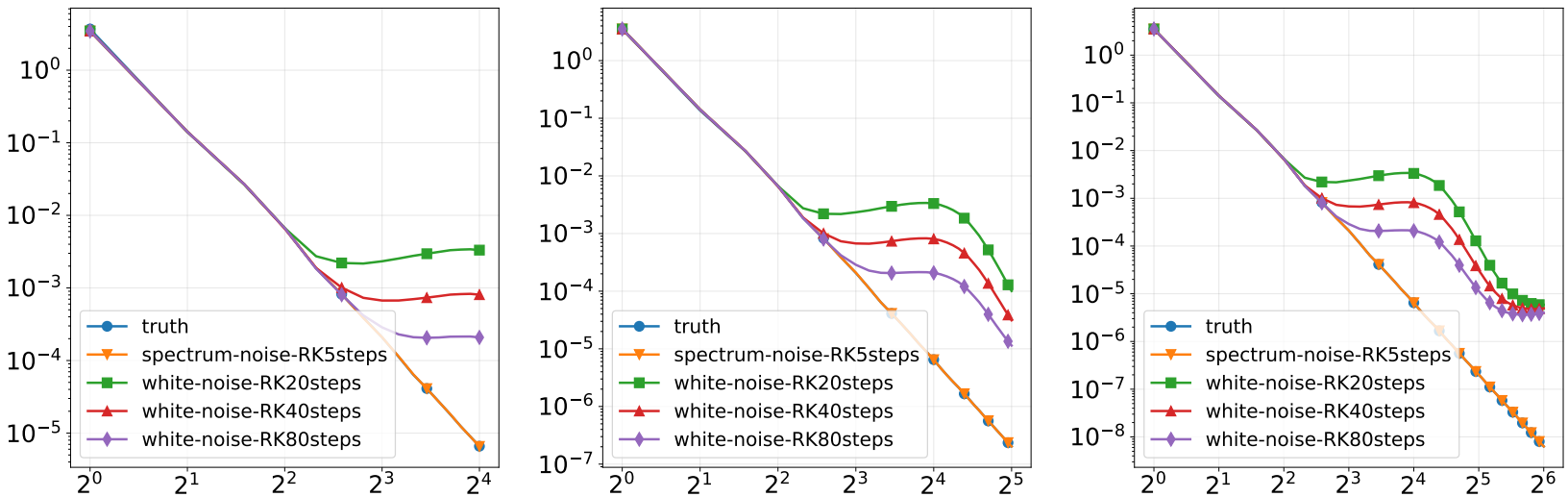


Figure: Energy spectra of Gaussian fields: comparison between ground truth, generation using Gaussian spectrum noise, and generation using white noise. Left: 32×32; middle: 64×64; right: 128×128

- Spectrum noise: Lipschitz bound independent of resolution

Matching smoothness (spectrum noise) works for near-Gaussian

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
- Standard schedule $\alpha_t = 1 - t, \beta_t = t$

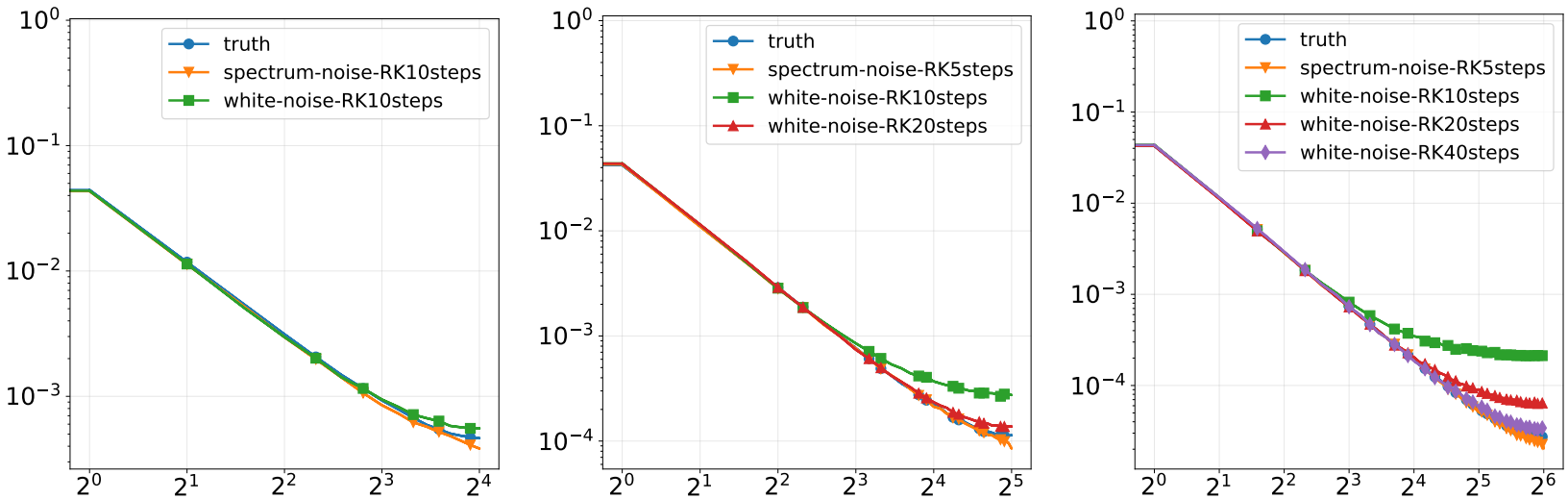


Figure: Energy spectra comparison between ground truth, generation using Gaussian spectrum noise and using white noise. Left: $N = 32$; middle: $N = 64$; right: $N = 128$

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

Spectrum noise struggles for stochastic Navier-Stokes

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 ω, velocity v, and $\mathrm{d}\eta$ forcing Ergodicity: [Hairer, Mattingly, 2006]
- $\nu = 10^3$, $\mathrm{d}\eta$ random forcing acts on a few Fourier modes

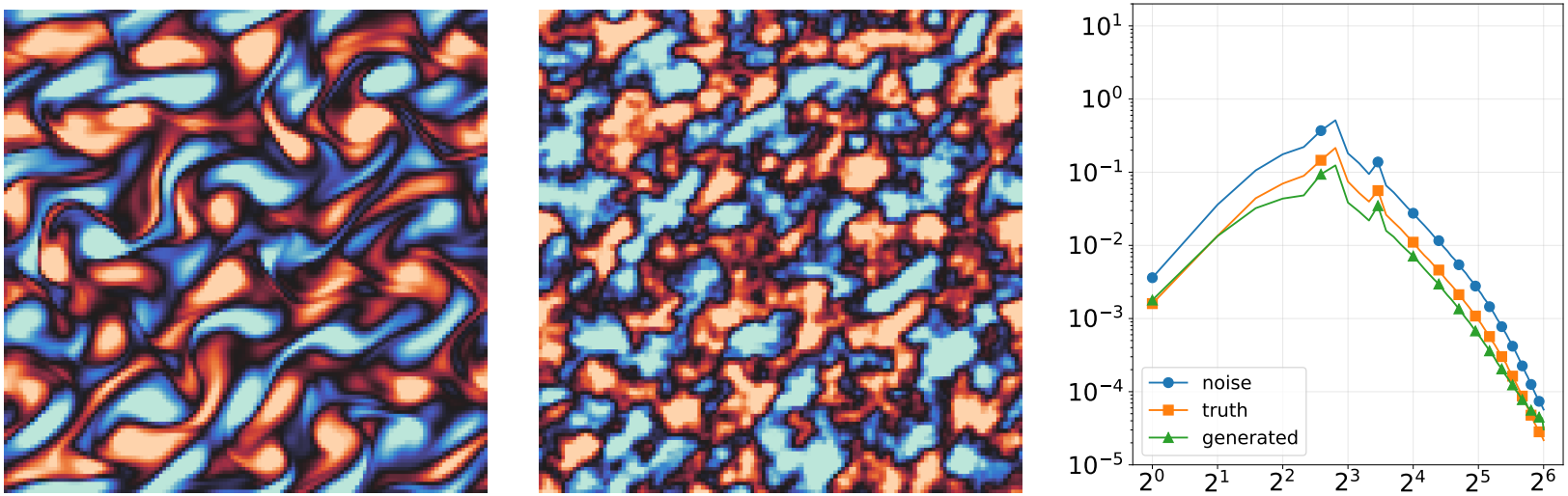


Figure: Left: sample from invariant distribution; middle: sample from Gaussian spectrum noise; right: enstrophy spectra of the noise, truth, and generation. Resolution 128×128. 10 steps of RK4 are used

Rougher noise works better

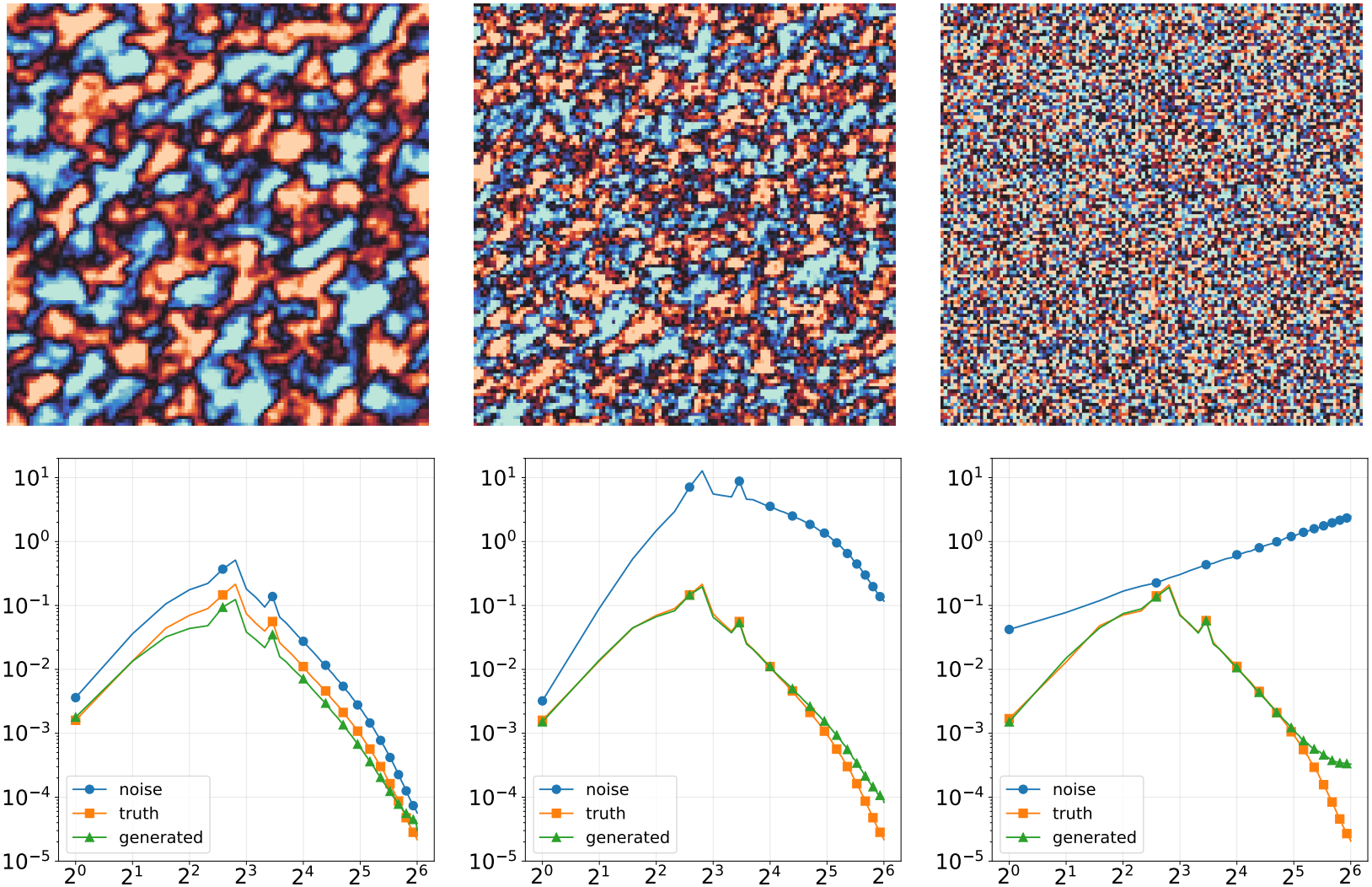


Figure: Three types of noises used for constructing generative models: Gaussian spectrum noise, Gaussian with a rougher spectrum, and white noise. Resolution 128×128. 10 steps of RK4 are used

Rougher noise: large Lipschitz energy, and optimized schedules

Gaussian: $x_1 \sim \mathrm{N}(0, C) \perp z \sim \mathrm{N}(0, \mathrm{I})$ in d dims. Let eigenvalues of C be $1 \geq \lambda^{(1)} \geq \lambda^{(2)} \geq \ldots \geq \lambda^{(d)} > 0$. Denote $M^\star = 1/\lambda^{(d)}$

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

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

Thm: 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 and general distributions using Beltrami Identity

Rougher noise w/ min-Lip schedule works for (near)-Gaussian too

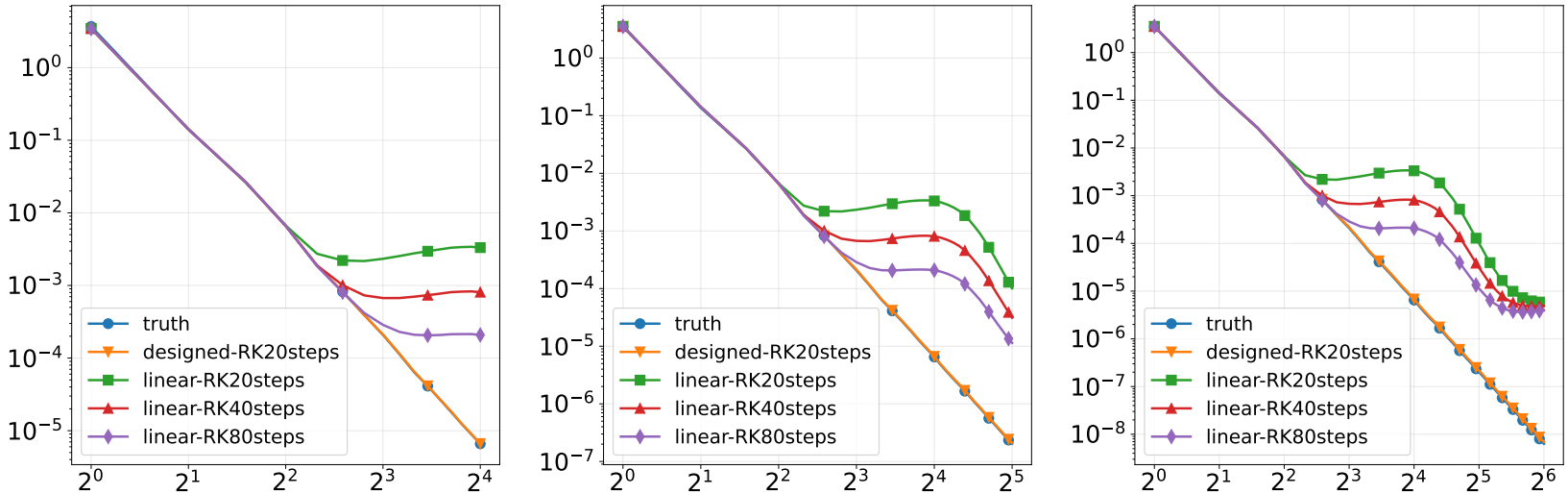


Figure: Gaussian measure example. Linear schedule versus optimized schedules. Left: 32×32; middle: 64×64; right: 128×128

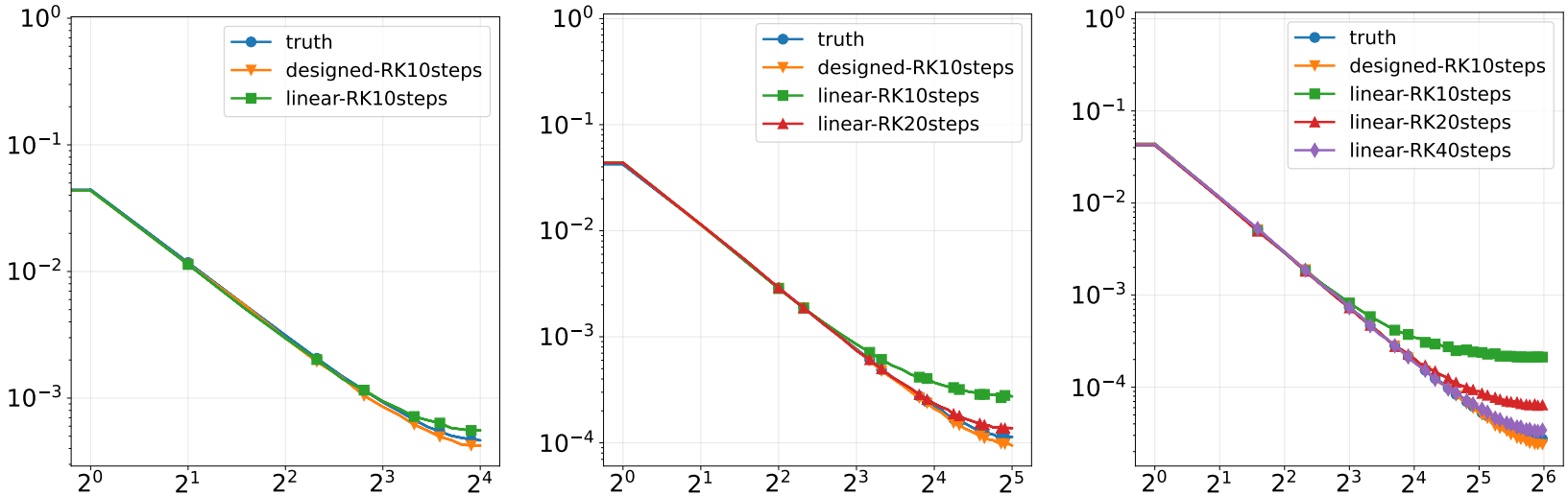


Figure: Stochastic Allen-Cahn example. Linear schedule versus optimized schedules. Left: $N = 32$; middle: $N = 64$; right: $N = 128$

Rougher noise w/ min-Lip schedule improves stochastic NS

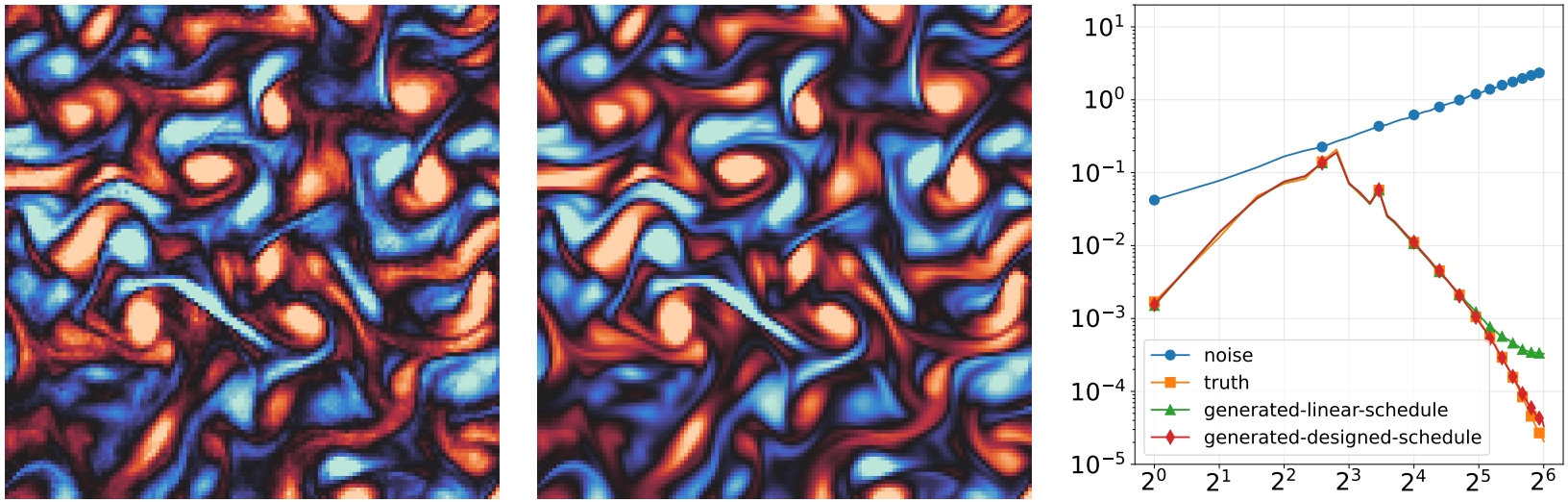


Figure: Experiments on stochastically forced Navier-Stokes using white noise. In all cases, we use 10 RK4 integration steps. Resolution: 128×128

- Left: generated samples w/ linear schedule
- Middle: generated samples w/ optimized schedule ($M^* = 10^5$)

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

- Right: enstrophy spectra of truth, noise, and generations

Case study for forecasting stochastic fluids

Forecasting 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 ω, velocity v, and $\mathrm{d}\eta$ is white-in-time random forcing
- Goal: given data pairs $(\omega_t, \omega_{t+\tau})$, forecast $\omega_{t+\tau}|\omega_t$ for new ω_t

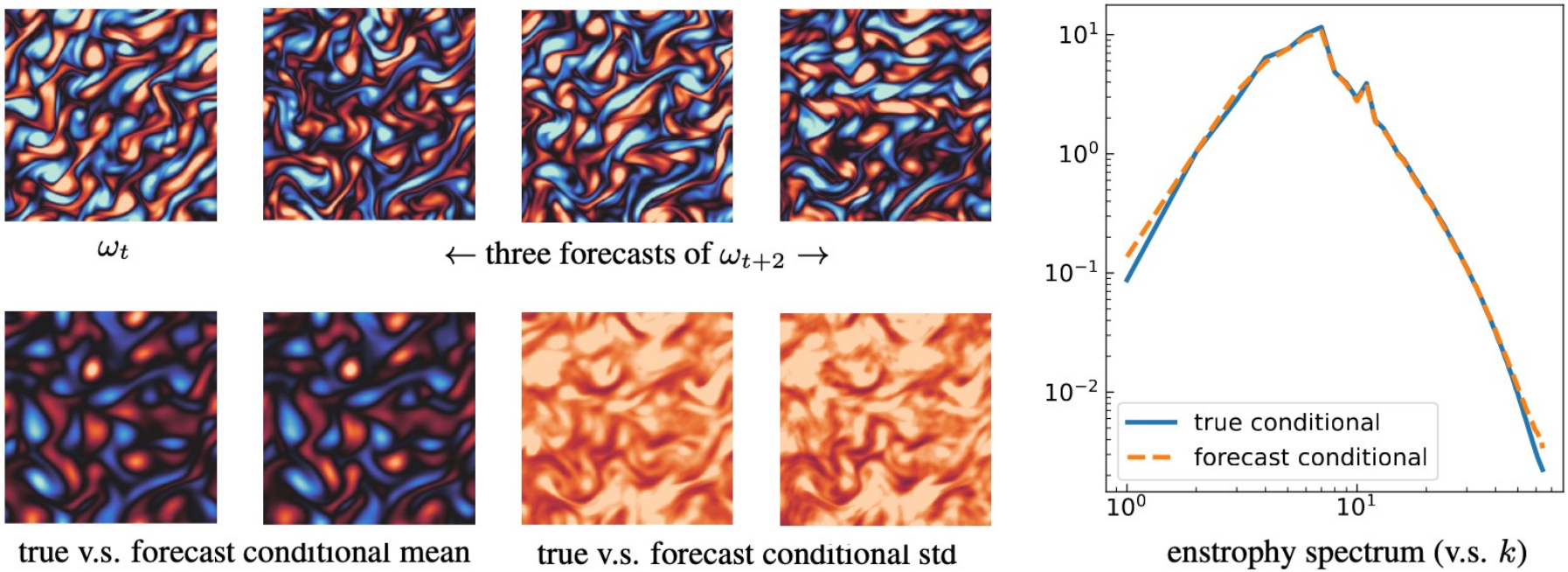


Figure: Results for forecasting with lag $\tau = 2$ (autocorrelation 10%). Resolution 128×128

Case study for combined forecasting and superresolution

Let ω_t be of 32×32 while $\omega_{t+\tau}$ is of 128×128

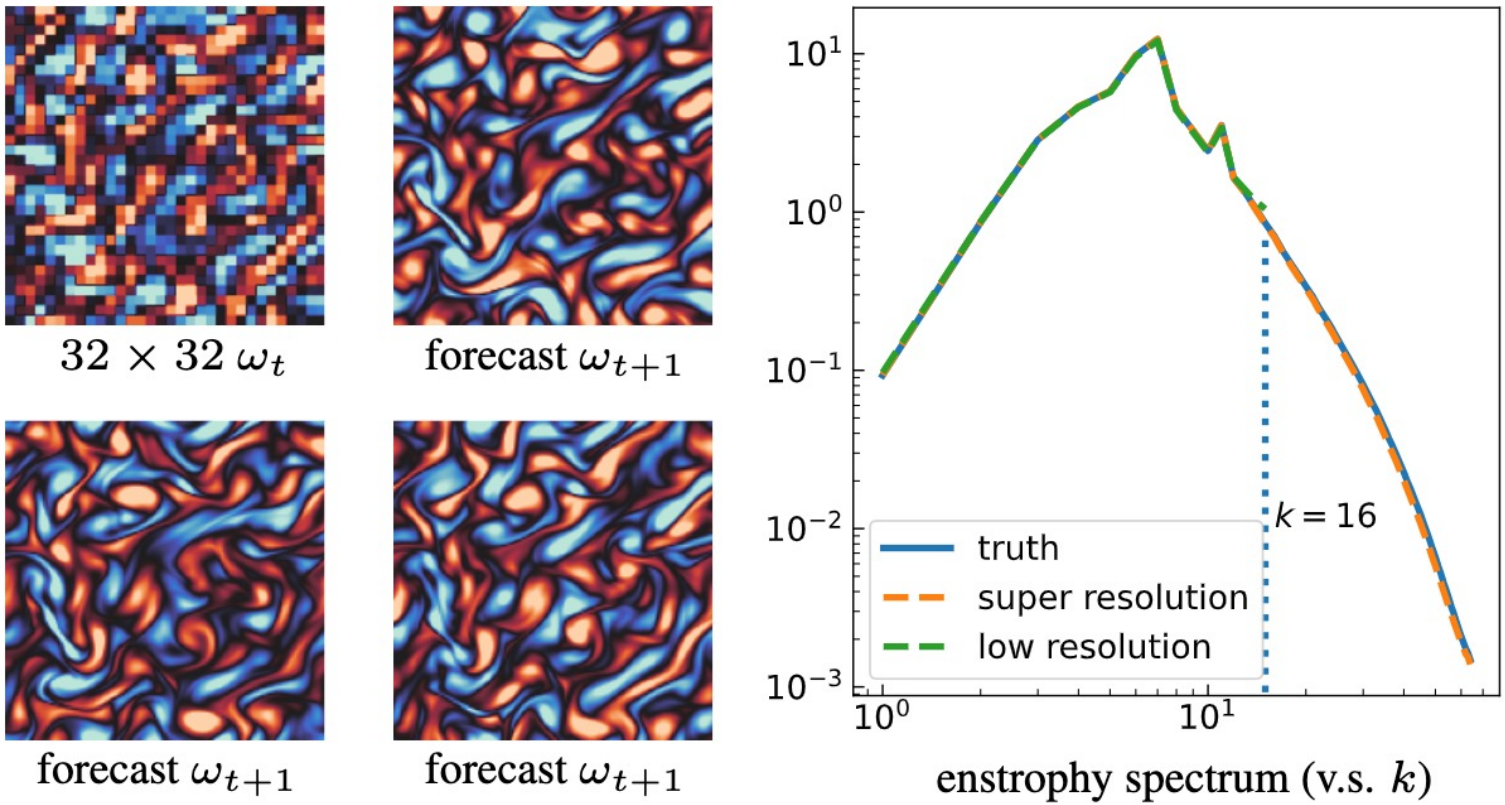


Figure: Results for probabilistic forecasting with low resolution input

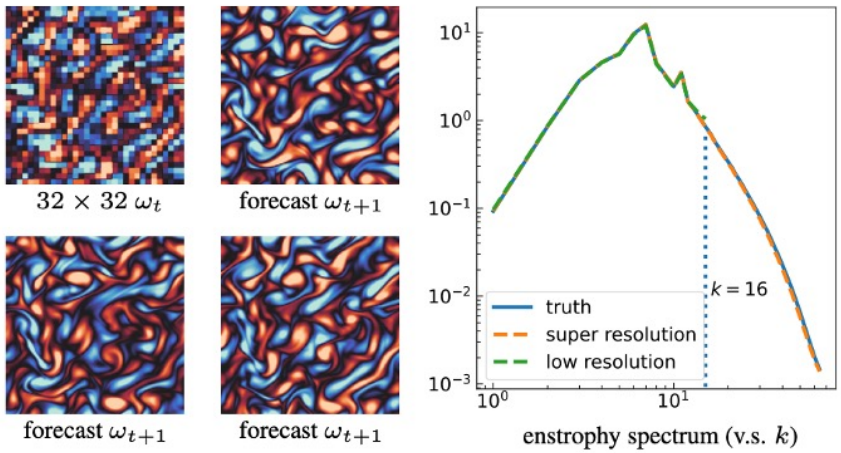
Conclusion

Design dynamical transport for generative flows of scientific data

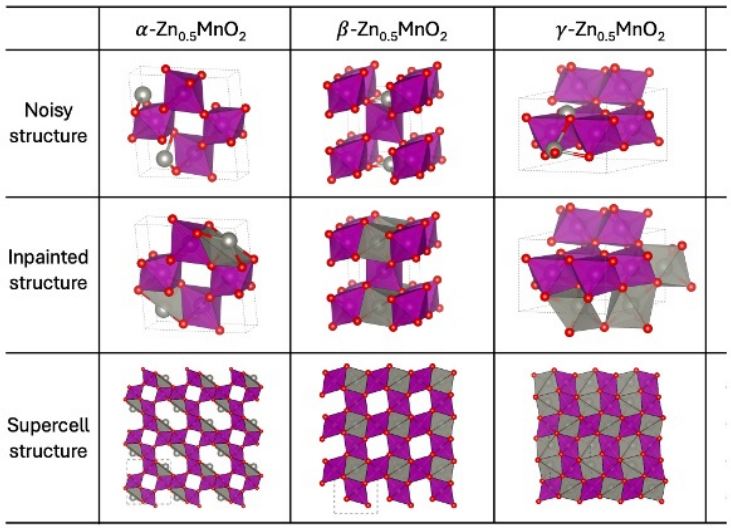
- **Strategy:** optimize Lipschitz energy in the construction of dynamical transport, alternative to optimal kinetic energy in optimal transport
- We discuss detailed theoretical and numerical instantiation of this strategy in the class of linear stochastic interpolants
- **Advantage:** resolution robust performance with small costs

General goal: structure/physics preserving techniques for generative models with improved statistical/numerical performance

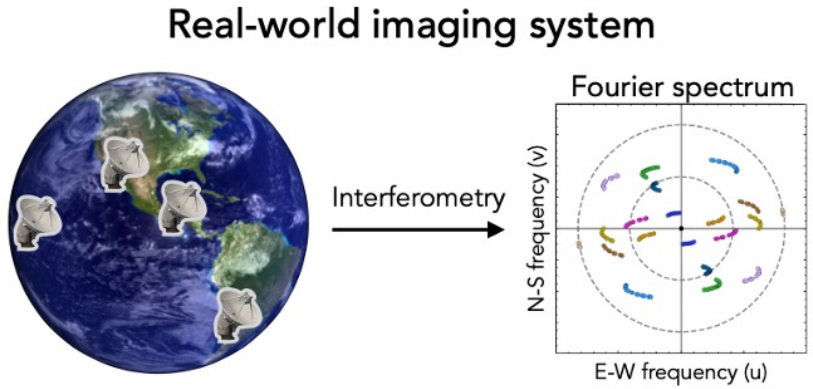
Insights and towards more applications for scientific tasks

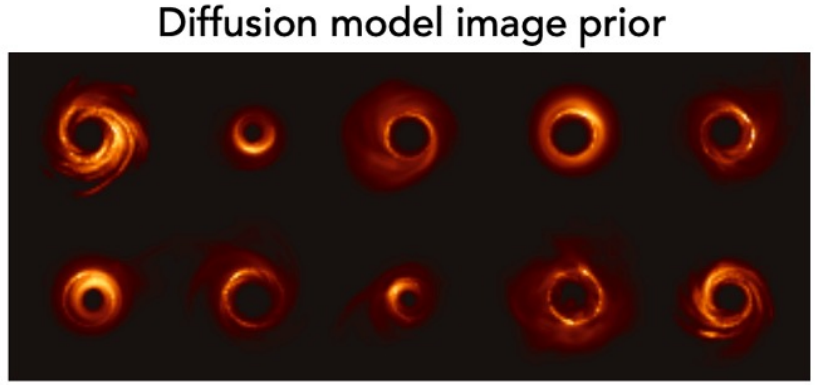


[Chen, Goldstein, Hua, Albergo, Boffi, Vanden-Eijnden 2024]



[Dai, Zhong, Deng, Chen, Ceder 2024]





[Sun, Wu, Chen, Feng, Bouman 2023]
[Wu, Sun, Chen, Zhang, Yue, Bouman 2024]

References

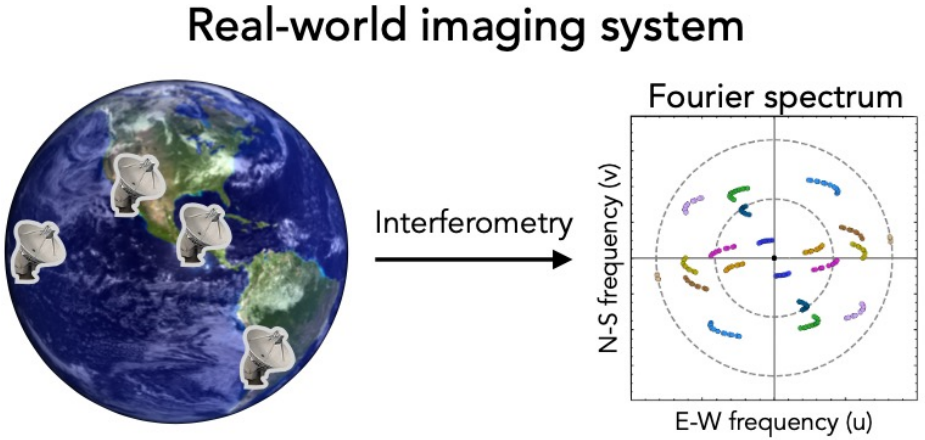
- Y. Chen, E. Vanden-Eijnden. Scale-Adaptive Generative Flows for Multiscale Scientific Data. arXiv:2509.02971, 2025
- Y. Chen, E. Vanden-Eijnden, J Xu. Lipschitz-Guided Design of Interpolation Schedules in Generative Models. arXiv:2509.01629, 2025
- Y. Chen, M. Goldstein, M. Hua, M. Albergo, N. Boffi, E. Vanden-Eijnden. Probabilistic Forecasting with Stochastic Interpolants and Föllmer Processes. ICML 2024

Thank you!

Probabilistic imaging
(real data black hole imaging)

Black hole imaging: Combining generative models and MCMC

[Sun, Wu, **Chen**, Feng, Bouman 2023], [Wu, Sun, **Chen**, Zhang, Yue, Bouman 2024]



As a Bayes inverse problem

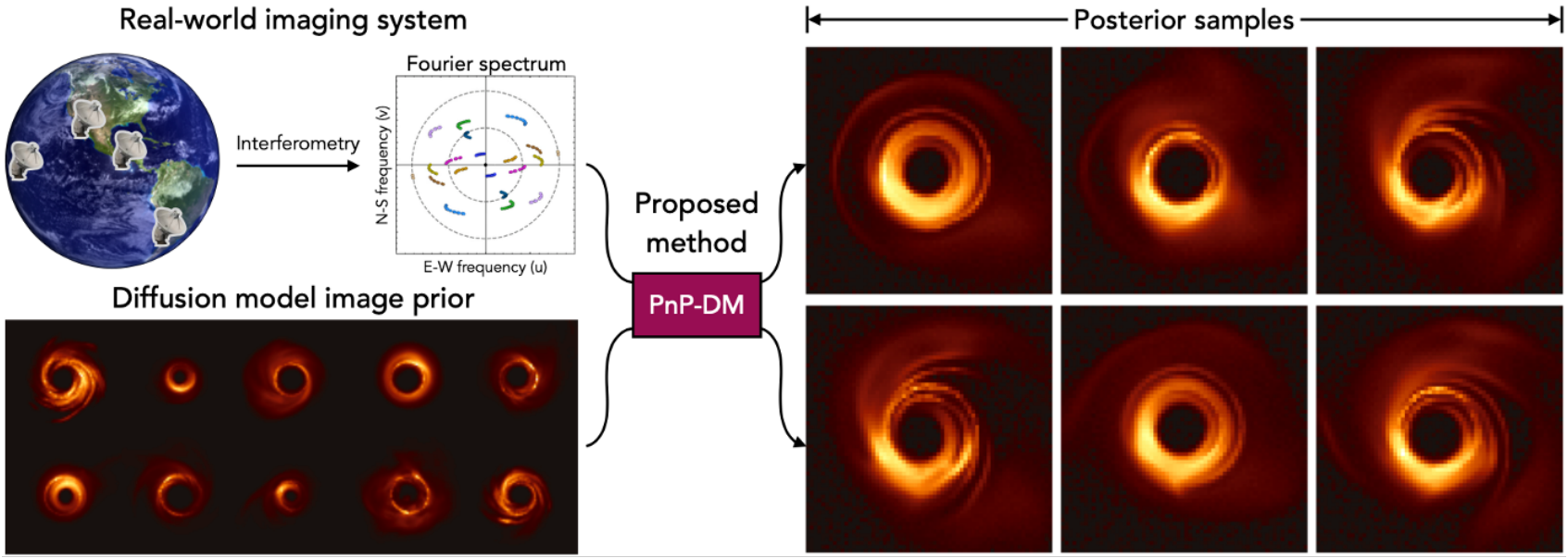
- **Data:** nonlinear functions of Fourier components of the image (very sparse and with strong noise)
- **Prior:** black holes simulated based on General Relativistic Magnetohydrodynamics (GRMHD)

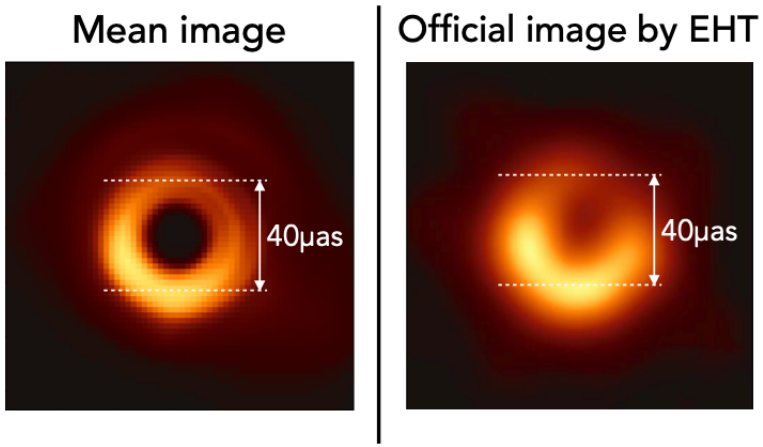
Goal: sample $\rho_{\text{post}} \propto \rho_{\text{prior}} \times L_{\text{likelihood}}$

Approach: learn ρ_{prior} using generative dynamics and combine with designed MCMC dynamics to sample ρ_{post}

Experiments with real data: PnP-DM (plug-and-play diffusion models)

PnP-DM uses split-Gibbs (alternating prior and likelihood update)





** Experiment is performed with real data for the M87 black hole*

Black hole imaging We adopted the same BHI setup as in [59, 61]. The relationship between the black hole image and each interferometric measurement, or so-called *visibility*, is given by

$$V_{a,b}^t = g_a^t g_b^t \cdot e^{-i(\phi_a^t - \phi_b^t)} \cdot \boldsymbol{F}_{a,b}^t(\boldsymbol{x}) + \eta_{a,b} \in \mathbb{C}, \tag{14}$$

where a and b denote a pair of telescopes, t represents the time of measurement acquisition, and $\boldsymbol{F}_{a,b}^t(\boldsymbol{x})$ is the Fourier component of the image $\boldsymbol{x}$ corresponding to the baseline between telescopes a and b at time t. In practice, there are three main sources of noise in (14): gain error g_a and g_b at the telescopes, phase error ϕ_a^t and ϕ_b^t, and baseline-based additive white Gaussian noise $\eta_{a,b}$. The gain and phase errors stem from atmospheric turbulence and instrument miscalibration and often cannot be ignored. To correct for these two errors, multiple noisy visibilities can be combined into data products that are invariant to these errors, which are called *closure phase* and *log closure amplitude* measurements [11]

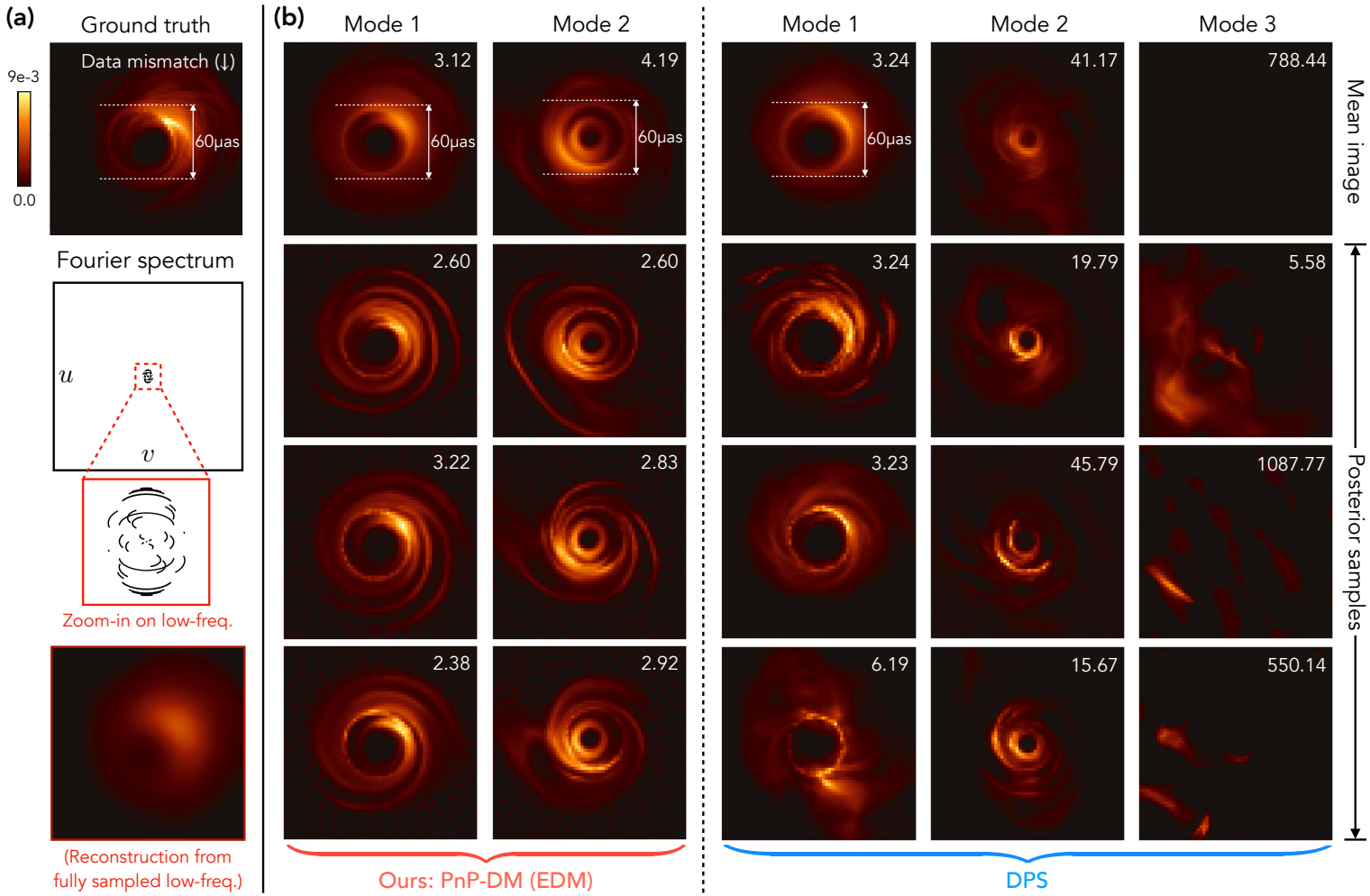
$$\begin{aligned}
\boldsymbol{y}_{t,(a,b,c)}^{\mathsf{cph}} &= \angle(V_{a,b} V_{b,c} V_{a,c}) := \mathcal{A}_{t,(a,b,c)}^{\mathsf{cph}}(\boldsymbol{x}), \\
\boldsymbol{y}_{t,(a,b,c,d)}^{\mathsf{logcamp}} &= \log\left(\frac{|V_{a,b}^t||V_{c,d}^t|}{|V_{a,c}||V_{b,d}^t|}\right) := \mathcal{A}_{t,(a,b,c,d)}^{\mathsf{logcamp}}(\boldsymbol{x}),
\end{aligned}$$

where $\angle$ computes the angle of a complex number. Given a total of M telescopes, there are in total $\frac{(M-1)(M-2)}{2}$ closure phase and $\frac{M(M-3)}{2}$ log closure amplitude measurements at time t, after eliminating repetitive measurements. In our experiments, we used a 9-telescope array ($M = 9$) from the Event Horizon Telescope (EHT) and constructed the data likelihood term based on these nonlinear closure quantities. Additionally, because the closure quantities do not constrain the total flux (i.e. summation of the pixel values) of the underlying black hole image, we added a constraint on the total flux in the likelihood term. The overall potential function of the likelihood is given by

$$f(\boldsymbol{x}; \boldsymbol{y}) = \sum_{t,\mathsf{c}} \frac{\|\mathcal{A}_{t,\mathsf{c}}^{\mathsf{cph}}(\boldsymbol{x}) - \boldsymbol{y}_{t,\mathsf{c}}^{\mathsf{cph}}\|_2^2}{2\sigma_{\mathsf{cph}}^2} + \sum_{t,\mathsf{d}} \frac{\|\mathcal{A}_{t,\mathsf{d}}^{\mathsf{logcamp}}(\boldsymbol{x}) - \boldsymbol{y}_{t,\mathsf{d}}^{\mathsf{logcamp}}\|_2^2}{2\sigma_{\mathsf{logcamp}}^2} + \frac{\left\|\sum_i \boldsymbol{x}_i - \boldsymbol{y}^{\mathsf{flux}}\right\|_2^2}{2\sigma_{\mathsf{flux}}^2}. \tag{15}$$

In this equation, $\boldsymbol{y}^{\mathsf{flux}}$ is the total flux of the underlying black hole, which can be accurately measured. We use $\boldsymbol{y} := (\boldsymbol{y}^{\mathsf{cph}}, \boldsymbol{y}^{\mathsf{logcamp}}, \boldsymbol{y}^{\mathsf{flux}})$ to denote all the measurements and c, d as the indices for

Black hole imaging: experiments with two modal synthetic data



- DPS: existing benchmark [Chung et al 2022]
- Ours: PnP-DM (plug-and-play diffusion models) using split Gibbs, with mathematical consistency guarantee

Technical Details on Choosing Noise

The choice of noise: initial time well-posedness for discretized GPs

- Consider $D = [0, 1]$ and Gaussian process $\xi \sim \mathcal{GP}(0, k)$ with

$$k(y, z) = \exp\left(-\frac{\|y - z\|_2}{2l^2}\right)$$

- Let $x_1 \in \mathbb{R}^N$ be a discretization of ξ with stepsize $h = 1/N$
- Taking $z \sim \mathrm{N}(0, \mathrm{I}_N)$, which preserves variance since

$$\mathbb{E}[\|z\|_2^2] = N = \mathbb{E}[\|x_1\|_2^2]$$

- With this z and $I_t = \alpha_t z + \beta_t x_1$, it holds

$$b_t(x) = \mathbb{E}[\dot{I}_t | I_t = x] = B_N(t)x$$

- We have $\lim_{t\to 0} \lim_{N\to\infty} \|B_N(t)\|_2 = \infty$
- Thus: drift $b_t(x)$ diverges in this limit

Need $z \sim \mathrm{N}(0, N\mathrm{I}_N)$ converging to non-trivial white noise

The choice of noise: initial time well-posedness for function space GPs

- Consider $z \sim \mathrm{N}(0, C_0)$ and $x_1 \sim \mathrm{N}(0, C_1)$ are drawn from Gaussian measures supported on Hilbert space H and $z \perp x_1$
- Given $I_t = \alpha_t z + \beta_t x_1$, it holds

$$b_t(x) = \mathbb{E}[\dot{I}_t | I_t = x] = B(t)x$$

where $B(t)$ is the linear operator defined as

$$B(t) = (\dot{\alpha}_t \alpha_t C_0 + \dot{\beta}_t \beta_t C_1)(\alpha_t^2 C_0 + \beta_t^2 C_1)^{-1}$$

- $\lim_{t \to 0} \|B(t)\|_H = \infty$ if $C_1 C_0^{-1}$ is an unbounded operator

Let $C_0 = \sigma_0^2(-\Delta + \tau_0^2 I)^{-s_0}$ and $C_1 = \sigma_1^2(-\Delta + \tau_1^2 I)^{-s_1}$.

- $C_1 C_0^{-1}$ is bounded if and only if $s_0 \leq s_1$
- noise $\mathrm{N}(0, C_0)$ is rougher, or at least as rough as, data $\mathrm{N}(0, C_1)$

Generalizing to non-Gaussian measures: b_t is bounded near initial time when data are in the Cameron-Martin space of noise

Probabilistic forecasting (benchmarking Navier-Stokes)

Goal: Build a generative dynamics $X_{0\leq s\leq 1}$ from x_0 to $x_1 \sim \rho^\star(\cdot|x_0)$

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

Methodology: Construct the stochastic process

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

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

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

$$\mathrm{d}X_s = b_s(X_s, x_0)\mathrm{d}s + \sigma_s \mathrm{d}W_s,\ X_{s=0} = x_0$$

It holds $\mathrm{Law}(X_s) = \mathrm{Law}(I_s|x_0)$. In particular $X_{s=1} \sim \rho^\star(\cdot|x_0)$

- Why? Itô's formula: $\mathrm{d}I_s = (\dot\alpha_s x_0 + \dot\beta_s x_1 + \dot\sigma_s W_s)\mathrm{d}s + \sigma_s \mathrm{d}W_s$
- Replacing drift by $\mathbb{E}[\dot\alpha_s x_0 + \dot\beta_s x_1 + \dot\sigma_s W_s | I_s, x_0]$ makes the SDE Markovian while keeping time-marginals unchanged

  Mimicking lemma, Markov projection [Gyöngy 1986]

Learning the generative dynamics from data

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

- **Fact:** the drift $b_s(x, x_0)$ is the unique minimizer of

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

  with sampled data (x_0, x_1) we can evaluate L_b

- **Algorithm:** parametrize $\hat{b}_s$ by neural nets, optimize L_b

- **Generative model:** for any x_0, integrate to $s = 1$ the SDE

$$\mathrm{d}\hat{X}_s = \hat{b}_s(\hat{X}_s, x_0)\mathrm{d}s + \sigma_s \mathrm{d}W_s, \ \hat{X}_{s=0} = x_0$$

  This will approximately sample $\rho^\star(\cdot|x_0)$ if $\hat{b}_s \approx b_s$

Conditional generation and applications in probabilistic forecasting

A benchmark case study: 2d NSE with stochastic forcing

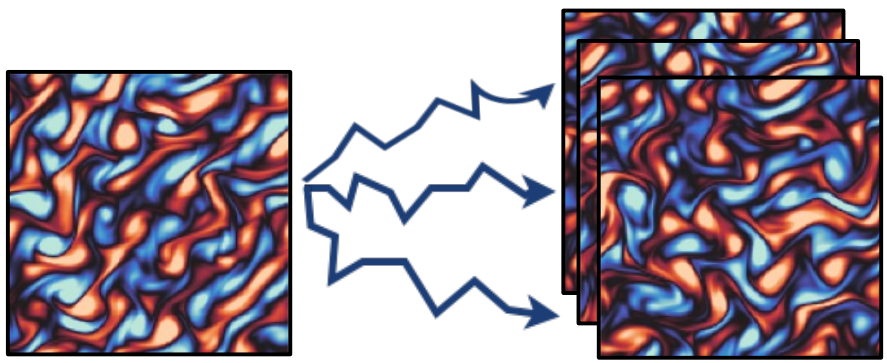
$$\mathrm{d}\omega + v \cdot \nabla\omega \mathrm{d}t = \nu\Delta\omega \mathrm{d}t - \alpha\omega \mathrm{d}t + \epsilon \mathrm{d}\eta \quad \text{on } \mathbb{T}^2$$

- vorticity ω, 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 ω_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

Experiments: Forecasting 2D stochastically forced NSE

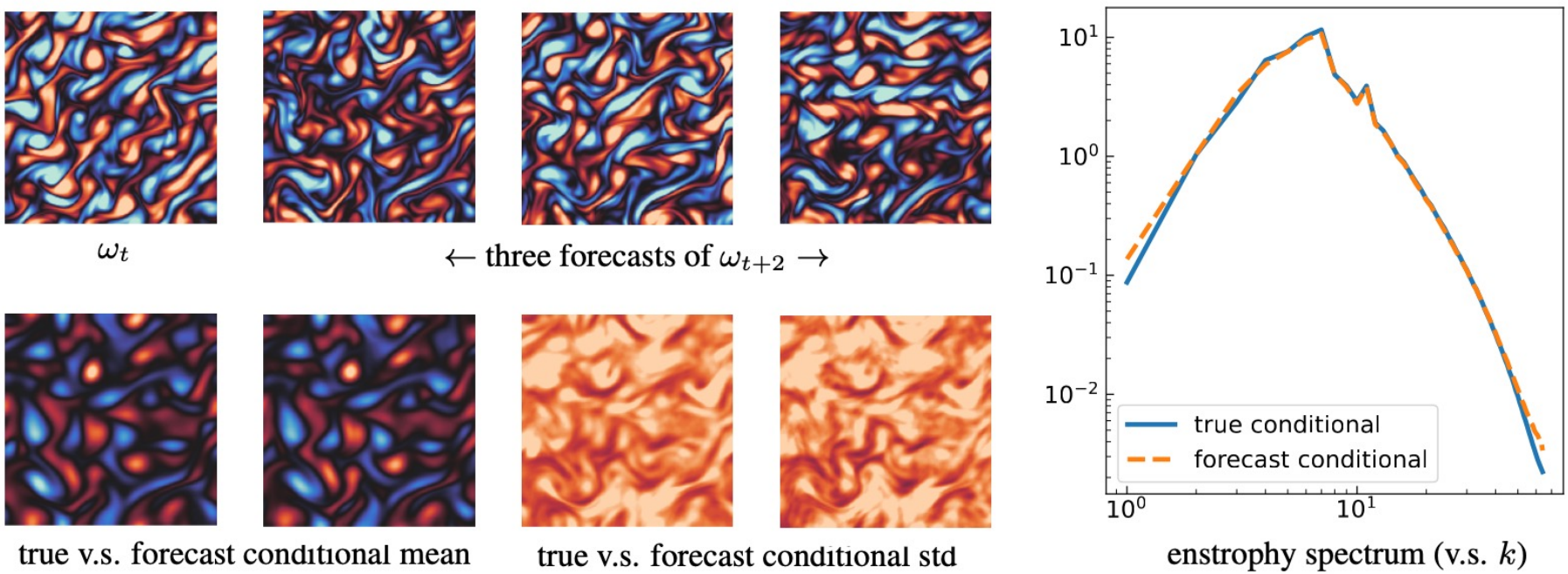


Figure: Lag $\tau = 2$ (autocorrelation 10%). Resolution 128×128, using $200K$ data pairs for training 2M-parameter-Unet

- Can be viewed as a surrogate model: for this example, 100 times faster than running the stochastic PDE simulation

A family of SDEs can be used. Which to choose?

Fact: 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 $\nabla \cdot (\rho \nabla \log \rho) = \Delta \rho$
- $\nabla \log \rho_s(x|x_0)$ is the score, with $\widehat{\text{score}}$ an estimator

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

A family of SDEs can be used. Which to choose?

Fact: 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 $\nabla \cdot (\rho \nabla \log \rho) = \Delta \rho$
- $\nabla \log \rho_s(x|x_0)$ is the score, with $\widehat{\text{score}}$ an estimator

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

Many existing studies on how to choose g in generative models

- ODEs versus SDEs, numerical schemes, perturbation analysis

  [Song et al 2021], [Song, Meng, Ermon 2021], [Karras, Aittala, Aila, Laine 2022], [Zhang, Tao, Chen 2023], [Albergo, Boffi, Vanden-Eijnden 2023], [Cao, Chen, Luo, Zhou 2024]

Answer to this question would depend on the choice of "metric"

KL divergence over path measures as the "metric": theory and practice

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

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

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

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

since by definition

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

KL divergence over path measures as the "metric": theory and practice

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

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

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

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

since by definition

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

SDE with $\sigma_s \mathrm{d}W_s$	SDE with $g_s^{\mathrm{F}} \mathrm{d}W_s$	ODE with Gaussian base
8.49e-3±1.57e-3	2.79e-3±9.19e-4	4.63e-3±9.63e-4

Empirical end-point KL err (total enstrophy of truth v.s. generated samples)

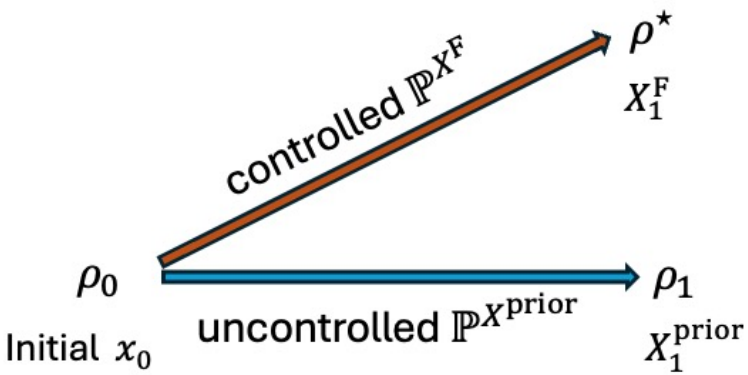
Further insights: What is special about this g^{F}?

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

- Solution to Schrödinger bridge when one endpoint is point mass

$$X^{\mathrm{F}} = \operatorname*{argmin}_{X} \mathrm{KL}[\mathbb{P}^{X} \| \mathbb{P}^{X^{\mathrm{prior}}}] \ \text{ s.t. } \ X_1 \sim \rho^{\star}(\cdot|x_0)$$

Standard Föllmer: X^{prior} is a Brownian motion
In our algorithm: X^{prior} is induced by the choices of $\alpha_s, \beta_s, \sigma_s$




Schrödinger


Föllmer

Interpretation: such optimal g^{F} is a "Bayes"/control solution!

[Schrödinger 1932]. Föllmer process [Föllmer 1986] wide applications in functional inequality [Lehec 2013] and in sampling [Zhang, Chen 2021], [Huang et al 2021], [Vargas et al 2023], etc

Further insights: What is special about this g^{F}?

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

- Solution to Schrödinger bridge when one endpoint is point mass

$$X^{\mathrm{F}} = \operatorname*{argmin}_{X} \mathrm{KL}[\mathbb{P}^{X} \| \mathbb{P}^{X^{\mathrm{prior}}}] \ \text{ s.t. } \ X_1 \sim \rho^{\star}(\cdot|x_0)$$

Standard Föllmer: X^{prior} is a Brownian motion
In our algorithm: X^{prior} is induced by the choices of $\alpha_s, \beta_s, \sigma_s$

Outlook: Design physically motivated X^{prior} (ongoing and future work)

- Multiscale interpolation I_s, connected to renormalization group

  e.g., [Bauerschmidt, Bodineau, Dagallier 2023]

- Function space noise with spectrum decay

  e.g., [Lim et al 2023], [Pidstrigach, Marzouk, Reich, and Wang 2023]

- Improved design choices for better numerical performance

  e.g., [Lim, Wang, Yu, Hart, Mahoney, Li, Erichson 2024]

Forecasting videos: CLEVER datasets

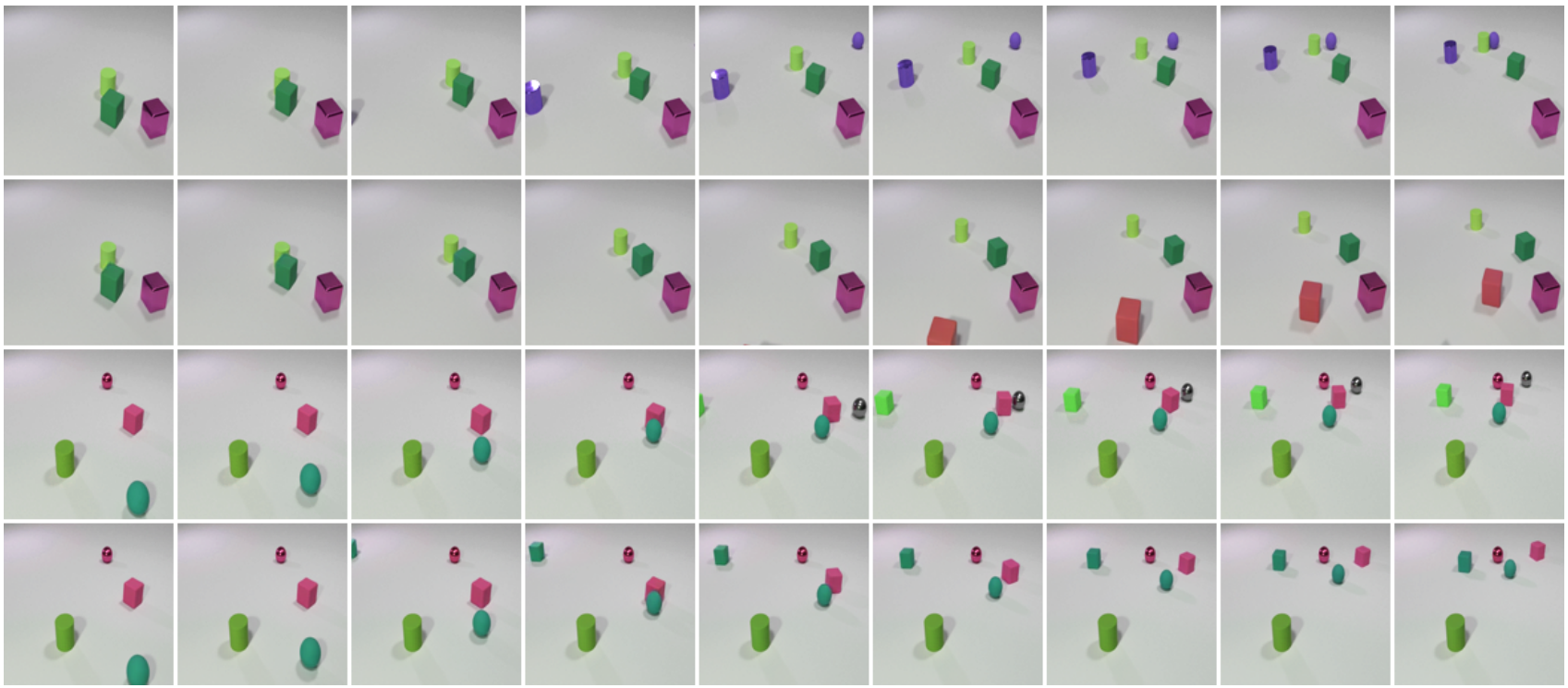


Figure: **Top row:** Real trajectory. **Second row:** Generated trajectory. A new, red cube enters the scene. **Third row:** Real trajectory. **Fourth row:** Generated trajectory. A new green cube enters the scene, and collision physics is respected (green ball hits red cube).

Forecasting videos: quantitative results

	KTH		*CLEVRER*	
Method	100k	250k	100k	250k
RIVER	46.69	41.88	60.40	48.96
PFI (ours)	44.38	39.13	54.7	39.31
Auto-enc.	33.45	33.45	2.79	2.79

Table: FVD computed on 256 test set videos, with the model generating 100 completions for each video. Results are reported for 100k grad steps and 250k. The auto-enc represents the FVD of the pretrained encoder-decoder vs the real data. It serves as a bound on the possible model performance, as the modeling is done in the latent space of a pre-trained VQGAN.

RIVER [Davtyan, Sameni, Favaro 2023]