When Do Local Score Models Extrapolate Across Size? A Diagnostic Theory and Benchmark

Anonymous authors
Paper under double-blind review

Abstract

Scientific generative modeling often requires size transfer, where models trained on small systems are evaluated on larger ones. While translation-invariant architectures enable this evaluation, we show that architectural locality alone does not guarantee stable size extrapolation. Instead, stable extrapolation depends on how far the Gaussian-smoothed score responds to a perturbation. Even when clean interactions are local, posterior correlations can make the smoothed score respond to distant perturbations. We formalize when this response remains sufficiently local for fixed-receptive-field models to transfer across system sizes, and prove a system-size-independent error bound for generated distributions on fixed local regions. We also introduce Finite-Depth Local Flow (FDLF), a white-box diagnostic benchmark with exact clean scores, densities, and controllable response ranges. Across 2D and 3D FDLF experiments, models transfer stably when the teacher response lies within their receptive fields. A related size-dependent pattern emerges near criticality in the standard 2D Ising model. In a 2D Ising test, the center-score change error of a fixed local model grows with system size near criticality but not away from it.

1 Introduction

Many scientific generative modeling problems require size transfer: a model is trained on small, accessible systems but is expected to operate on substantially larger ones. Translation-invariant architectures such as local CNNs and graph neural networks allow the same model to run on larger domains. Related architectures have been widely applied in many fields, from particle and molecular systems to partial differential equations (PDEs) (Satorras et al., 2021; Hoogeboom et al., 2022; Xu et al., 2022; Jing et al., 2022; Li et al., 2021; Kovachki et al., 2023; Lu et al., 2021; Behler and Parrinello, 2007; Mills et al., 2019; Schütt et al., 2017; Batzner et al., 2022). For diffusion models, a fundamental question remains: when does a learned local score remain the correct physical rule after the domain grows?

Our answer is that stable size extrapolation depends on how far the Gaussian-smoothed score responds to a perturbation. Denoising score matching learns Gaussian-smoothed scores Hyvärinen (2005); Vincent (2011); Song et al. (2021). Even when the clean density has finite-range structure, smoothing rewrites the score as a posterior expectation through Tweedie's formula Robbins (1992); Efron (2011). Far-away perturbations can influence a local score component through posterior covariance. When correlations in the noisy posterior decay with distance (Dobrushin, 1968; Georgii, 2011; Weitz, 2006), a condition known as spatial mixing, this influence also decays and a finite receptive field can capture the correct local score. Related work shows that small local denoisers can fail over a narrow range of noise levels when denoising requires information over long distances (Hu et al., 2025).

For size extrapolation, this insight leads to a testable diagnostic: a local score model should stably transfer across sizes when its receptive field covers the effective response range of the target smoothed score, and the extrapolation becomes unreliable when it does not. This diagnostic is stronger than merely checking whether an architecture can be evaluated on larger systems. It asks whether the learned rule has captured the spatial response that actually determines the local reverse-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

diffusion dynamics. Existing score-based analyses connect accurate score estimates to sampling accuracy (Lee et al., 2022; Chen et al., 2023). Dimension-free bounds are also available for smooth observables of reverse diffusion (de Bortoli et al., 2025). Static locality results bound fixed marginals using score differences between two fixed distributions (Cui et al., 2024). We study how score errors at different sites influence a fixed patch during generation as the system size grows. Testing this mechanism directly requires exact scores and a known data-generating rule.

To address these challenges, we make three main contributions. First, we use the relation between posterior covariance and smoothed-score response to formulate a response-tail criterion for local score extrapolation (Gottwald et al., 2025). Second, we prove a system-size-independent error bound for generated distributions on fixed local regions. Third, we introduce Finite-Depth Local Flow (FDLF), a white-box diagnostic benchmark with exact clean scores and controllable response ranges. This enables us to compare, under the same training and evaluation protocol, stable transfer when the response lies within the receptive field and failure when substantial response remains outside it.

On 2D continuous FDLF teachers, the local model extrapolates stably when its receptive field covers the teacher response range. For a long-range teacher, the local model's clean-score error grows sharply with system size. The same stable and failing regimes appear on periodic 3D FDLF. Related work finds that fixed-range score error grows with system size in a critical 2D Gaussian field (Del Bono et al., 2026). We test a fixed local model trained at one size on the standard 2D Ising model near criticality (Fisher and Barber, 1972; Privman, 1990). Under a perturbation outside the receptive field, the true center-score change increasingly differs from the prediction of a fixed-radius model as system size grows beyond the training size. This mismatch does not grow away from criticality.

2 Local Score Extrapolation

Let Λ_L be a finite spatial domain and let $x = (x_i)_{i\in\Lambda_L}$ with $x_i \in \mathbb{R}^d$. For each system size L, let p_L be a density on $(\mathbb{R}^d)^{|\Lambda_L|}$. In the size-extrapolation setting, a score model is trained on samples from $p_{L_{\mathrm{train}}}$ and then evaluated without retraining on larger systems $p_{L_{\mathrm{test}}}$, where $L_{\mathrm{test}} > L_{\mathrm{train}}$. The assumption is that $\{p_L\}_L$ is generated by a shared size-uniform mechanism.

We focus on the case in which the underlying rule is shared across sizes, yet the learned local score may still be wrong because the score component at a site depends on a spatial context larger than the model can see. We therefore evaluate extrapolation locally: a model succeeds if, on every fixed-size patch, the score components and generated distribution remain accurate as L increases.

For score-based diffusion learning (Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Ho et al., 2020; Song et al., 2021), let $X_0 \sim p_L$ and

$$X_\sigma = X_0 + \sigma\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I). \tag{1}$$

The smoothed density is

$$p_{L,\sigma}(x) = \int p_L(x_0)\,\mathcal{N}(x; x_0, \sigma^2 I)\,dx_0, \tag{2}$$

with target score $s_{L,\sigma}(x) = \nabla_x \log p_{L,\sigma}(x)$. Denoising score matching (Vincent, 2011) trains s_θ by

$$\mathcal{L}_{\mathrm{DSM}}(\theta) = \mathbb{E}_{\sigma,X_0,\varepsilon}\left[\left\| s_\theta(X_0 + \sigma\varepsilon, \sigma) + \frac{\varepsilon}{\sigma}\right\|_2^2\right]. \tag{3}$$

The expected DSM objective is minimized by the smoothed score $s_{L,\sigma}$ (Vincent, 2011; Song et al., 2021).

A local score model represents a size-independent site rule,

$$s_\theta(x, \sigma)_i = F_\theta(x_{B_R(i)}, \sigma), \tag{4}$$

where $B_R(i) = \{j : \mathrm{dist}(i, j) \le R\}$. The same R and θ are used for all system sizes. Thus, the relevant extrapolation question becomes: is the true score component at site i effectively determined by information inside $B_R(i)$?

To make this question operational, define the response tail of a differentiable score field by

$$\mathcal{T}_{L,\sigma}(i;R,x) = \sum_{\mathrm{dist}(i,j)>R} \left\| \frac{\partial s_{L,\sigma}(x)_i}{\partial x_j} \right\|_{\mathrm{op}}. \tag{5}$$

If $\mathcal{T}_{L,\sigma}(i;R,x)$ is small uniformly over relevant x, then variables outside the receptive field have little first-order influence on the local score. If it is large, no radius-R score model can represent the true response without additional global information. The experiments use finite-difference and exact derivative versions of this quantity as observable diagnostics.

2.1 Clean and Smoothed Locality

We relate clean-score locality to local interactions and smoothed-score locality to posterior correlations.

For clean densities, strict locality follows from a local log-density decomposition. Suppose

$$\log p_L(x) = \sum_{j\in\Lambda_L} \phi_j(x_{B_r(j)}) \tag{6}$$

with interaction radius r independent of L. Then $\nabla_{x_i} \log p_L(x)$ only receives contributions from terms whose neighborhoods contain i, and each such term depends on variables at distance at most $2r$ from i. Hence the clean score is finite-range local. This elementary statement is useful but insufficient for diffusion training, which targets smoothed scores.

For $\sigma > 0$, Tweedie's formula (Robbins, 1992; Efron, 2011) gives

$$s_{L,\sigma}(x) = \sigma^{-2}\big(\mathbb{E}[X_0 \mid X_\sigma = x] - x\big), \tag{7}$$

where the noisy posterior over clean configurations is

$$\pi^x_{L,\sigma}(dx_0) \propto p_L(x_0) \exp\left(-\frac{\|x-x_0\|_2^2}{2\sigma^2}\right) dx_0. \tag{8}$$

Consequently, differentiating the mean under this noisy posterior gives the covariance, or linear-response, identity (Georgii, 2011): for $i \neq j$,

$$\frac{\partial s_{L,\sigma}(x)_i}{\partial x_j} = \sigma^{-4} \operatorname{Cov}_{\pi^x_{L,\sigma}}(X_{0,i}, X_{0,j}). \tag{9}$$

Thus the spatial response of the smoothed score is governed by posterior correlations. If the noisy posterior has uniform correlation decay, as in classical spatial-mixing regimes for Gibbs measures (Dobrushin, 1968; Georgii, 2011), the score Jacobian is quasi-local. If spatial mixing fails, the smoothed score can have substantial long-range response even when the clean prior has short-range interactions.

Equation (9) is the conceptual hinge of the paper. It says that the effective receptive field required by a denoising score model is a property of the noisy posterior, not only of the clean interaction graph. Increasing CNN depth should help when it reduces the response tail in Eq. (5). It should not solve cases where the teacher still has large response beyond all tested radii.

2.2 Local Distributions and Size-Uniform Control

We prove an upper bound on the error of the generated distribution on a fixed-size patch. This bound does not grow with the total system size.

For a fixed patch $A \subseteq \Lambda_L$, its local marginal is the distribution of the variables x_A, obtained by integrating out the variables outside A. We consider patches with at most m sites, where m does not grow with the total system size L.

Theorem 1 (Informal size-uniform local control)**.** *Under the assumptions in Appendix A.2, for fixed patch size m and generation time T,*

$$\sup_L \sup_{\substack{A\subset\Lambda_L \\ |A|\le m}} d_{\mathrm{BL}}(\widehat{Q}_{L,A}, P_{L,A}) \le \delta_{0,m} + C_{\mathrm{loc}}(m,T)\eta_m. \tag{10}$$

Here $P_{L,A}$ and $\widehat{Q}_{L,A}$ are the final patch distributions produced using the true and learned scores, respectively. The distance d_{BL} measures their largest difference in expected patch observables bounded in magnitude by one and changing by at most the distance between their inputs.

The bound has three ingredients. The initialization term $\delta_{0,m}$ bounds the final patch error caused by different initial noise distributions. It is zero when both processes start from the same distribution. The factor $C_{\mathrm{loc}}(m, T)$ bounds the strength of the exact reverse process's response, and depends on the dynamics, patch size, and generation time, but not on L. The quantity η_m bounds the accumulated score error along the learned generation process, weighted by each site's influence on the patch and including the known time-dependent coefficient that converts score error into the resulting error in the generation dynamics.

The physical picture is that a patch is affected most strongly by nearby score errors during generation. It is less affected by score errors at sufficiently distant sites. If the total influence and the weighted error remain bounded as L grows, the right-hand side of Eq. (10) remains independent of system size. Appendix A.2 gives the complete assumptions, definitions, and proof, while Appendix A.3 directly measures how a local perturbation spreads through exact reverse processes.

3 FDLF: A Testable Diagnostic Benchmark

3.1 Benchmark Construction

We introduce Finite-Depth Local Flow (FDLF) as a *testable diagnostic benchmark* for local size extrapolation.

FDLF is built around three properties. The teacher exposes exact clean scores, so score error can be measured directly rather than inferred from samples. The same teacher rule is reusable at every L. The teacher's response range is tunable, so positive controls, receptive-field-limited cases, and controlled failures can be compared under one protocol.

Following the normalizing-flow change-of-variables framework (Rezende and Mohamed, 2015; Papamakarios et al., 2021), let $u = (u_i)_{i\in\Lambda_L}$ be a latent field with tractable density q_L, and let T_L be a size-uniform finite-depth invertible map. The target field is

$$x = T_L(u), \qquad p_L = (T_L)_{\#} q_L. \tag{11}$$

Its density and clean score are

$$p_L(x) = q_L(T_L^{-1}x)\left|\det DT_L^{-1}(x)\right|, \qquad s_{L,0}(x) = \nabla_x \log p_L(x). \tag{12}$$

All concrete FDLF teachers used here employ coupling-style invertible constructions (Dinh et al., 2017) whose inverse and log-Jacobian remain local or controlled-range. By varying how p_L is generated, we obtain positive controls with short response ranges, medium-range teachers whose responses are captured only by larger CNNs, and long-range stress teachers. The generation rule for each teacher remains unchanged as L increases. The 2D sweep uses $L = 16, \ldots, 64$ and the 3D sweep uses $L = 16, \ldots, 48$.

The long-range teacher is a stress test for size extrapolation. It shows how a local CNN fails when the teacher score depends on information beyond the receptive field.

In summary, any proposed mechanism for local size extrapolation should pass controlled tests of this kind. It should succeed when the teacher response is inside the receptive field. It should improve when the receptive field is enlarged to cover an intermediate response range. It should fail or degrade when the teacher response remains outside the field. The following sections instantiate these cases.

3.2 2D Continuous Diagnostic

We first test size transfer on periodic 2D continuous lattices with a fixed model architecture. We train only at $L_{\mathrm{train}} = 16$ and evaluate the model at $L = 16, 32, 48, 64$. The model uses one residual block, with $R_{\mathrm{eff}} = 4$.

All score models in this section are pure CNNs with finite receptive fields. A single 3×3 convolution expands the receptive field by one site in each coordinate direction. Our CNN has one input

projection, two 3×3 convolutions per residual block, and one output projection. A model with B residual blocks therefore has effective radius $R_{\text{eff}} = 2B + 2$. The CNN normalization is computed independently at each lattice site. It does not pass information between sites or enlarge this effective radius. The 2D and 3D FDLF students are trained with a σ^2-weighted version of the DSM objective in Eq. (3). We evaluate the model at the smallest noise level, $\sigma_{\min} = 0.005$, and compare it with the exact clean score. We report the relative root-mean-square error (RMSE)

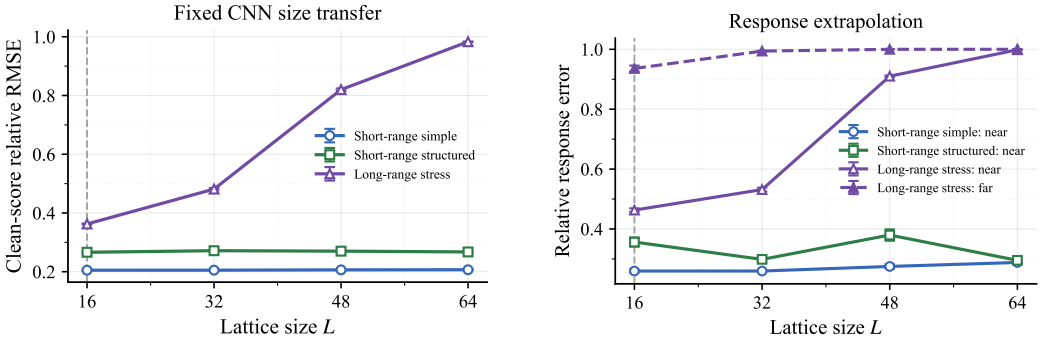
$$\mathrm{RMSE}_{\mathrm{score}}(L) = \left(\frac{\mathbb{E}_{x \sim p_L} \| s_\theta(x, \sigma_{\min}) - s_{L,0}(x) \|_2^2}{\mathbb{E}_{x \sim p_L} \| s_{L,0}(x) \|_2^2} \right)^{1/2}. \tag{13}$$

We also perturb local and distant regions and compare the resulting score changes,

$$\Delta s_{L,0} = s_{L,0}(x + \delta) - s_{L,0}(x), \qquad \Delta s_\theta = s_\theta(x + \delta, \sigma_{\min}) - s_\theta(x, \sigma_{\min}). \tag{14}$$

Near perturbations act on the central 5×5 patch. Far perturbations act outside this same patch for every CNN. The relative response error compares the predicted and teacher center-score changes. It is their root-mean-square difference divided by the teacher's root-mean-square change.

This experiment compares the two short-range teachers to the long-range stress teacher. The short-range structured teacher adds local smoothing and a nonlinear correction to the short-range simple construction.



(a) Clean-score relative RMSE.

(b) Response error.

Figure 1: Fixed-architecture pure-CNN size extrapolation on the 2D FDLF benchmark, aggregated over three independent training seeds. Error bars are standard errors of the mean (s.e.m.) across these seeds. Models use $B = 1$ residual block, hence $R_{\text{eff}} = 4$, are trained only at $L = 16$ with pure DSM, and are evaluated at larger sizes. Normalization is independent at each lattice site. For the short-range simple and short-range structured teachers, student score and near-response errors remain stable across L. For the L-independent long-range stress teacher, these errors increase sharply as the evaluated lattice grows.

Figure 1 establishes the basic size-transfer phenomenon. For the two short-range teachers, the fixed local CNN trained at $L = 16$ maintains nearly constant clean-score relative RMSE as the lattice grows. For the long-range stress teacher, the result is different: under the same L-independent teacher rule, the student score error increases from 0.362 ± 0.002 at $L = 16$ to 0.9830 ± 0.0003 at $L = 64$, and its far-response error remains essentially saturated at 1.0000004 ± 0.0000003 for $L = 64$.

We then compare CNNs with different depths and receptive-field sizes. All models in this comparison are trained at $L_{\text{train}} = 32$ and evaluated at $L = 64$. Adding residual blocks increases the effective radius, depth, and number of parameters. The largest tested radius, $R_{\text{eff}} = 14$, remains below half the training lattice width. This experiment replaces the short-range structured teacher by the medium-range teacher to test radius-dependent improvement.

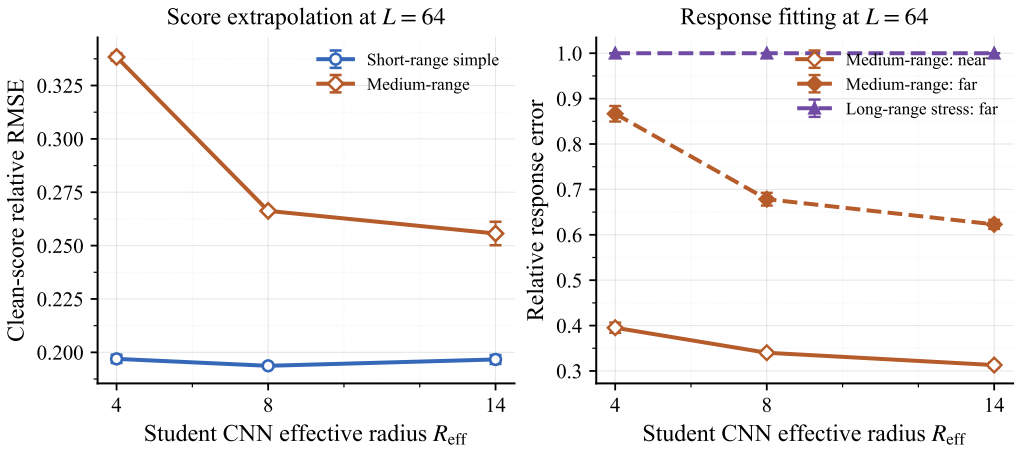


Figure 2: Pure-CNN depth and receptive-field sweep on the 2D FDLF benchmark, aggregated over three independent training seeds. Models are trained at $L = 32$ with pure DSM and evaluated at $L = 64$. The x-axis is the effective CNN radius induced by the number of 3×3 convolutional residual blocks. Left: student clean-score error remains stable across radii for the short-range simple teacher. It decreases substantially for the medium-range teacher as the receptive field grows. Right: student near/far response errors decrease for the medium-range teacher. Far-response error remains high for the L-independent long-range stress teacher. Error bars are s.e.m. across training seeds.

The short-range simple teacher is already captured by the smallest CNN. Its clean-score relative RMSE remains essentially flat at $L = 64$, with 0.197 ± 0.002 at $R_{\text{eff}} = 4$ and 0.197 ± 0.002 at $R_{\text{eff}} = 14$. For the medium-range teacher, clean-score RMSE at $L = 64$ decreases from 0.338 ± 0.002 to 0.256 ± 0.005 as the CNN effective radius increases from 4 to 14.

The response panel measures how much of the teacher response the CNN captures. For the medium-range teacher, near-field response error decreases from 0.395 ± 0.011 to 0.313 ± 0.006, while far-field response error drops from 0.867 ± 0.017 to 0.623 ± 0.010. For the L-independent long-range stress teacher, student clean-score RMSE at $L = 64$ remains near one across $R_{\text{eff}} = 4, 8, 14$, 0.982 ± 0.0002, 0.986 ± 0.001, and 0.982 ± 0.002, and far-response error remains essentially saturated. In conclusion, deeper local models with larger receptive fields capture more of the medium-range response. However, the long-range response remains poorly captured. These results are consistent with the proposed mechanism.

The same stable and failing regimes also appear in 3D (Appendix B.2).

4 Critical Ising Stress Test

Finally, we test size extrapolation near criticality in the 2D Ising model. Here, nearest-neighbor interactions can produce a long-range smoothed-score response. The clean distribution is (Onsager, 1944; Georgii, 2011)

$$p_\beta(x) \propto \exp\left(\beta \sum_{\langle i,j \rangle} x_i x_j\right), \qquad x_i \in \{-1, +1\}. \tag{15}$$

Here $\langle i, j \rangle$ denotes nearest-neighbor pairs, and β controls their interaction strength. We observe the spins through $Y_i = X_i + \sigma \epsilon_i$, with independent standard Gaussian noise ϵ_i. The noisy posterior is

$$\pi_\beta(x \mid y) \propto \exp\left(\beta \sum_{\langle i,j \rangle} x_i x_j + \sigma^{-2} \sum_i y_i x_i\right). \tag{16}$$

At $y = 0$, this posterior is the zero-field Ising model. As β approaches the square-lattice critical value $\beta_c = \frac{1}{2} \log(1 + \sqrt{2})$ (Onsager, 1944), posterior correlations extend over larger distances. By

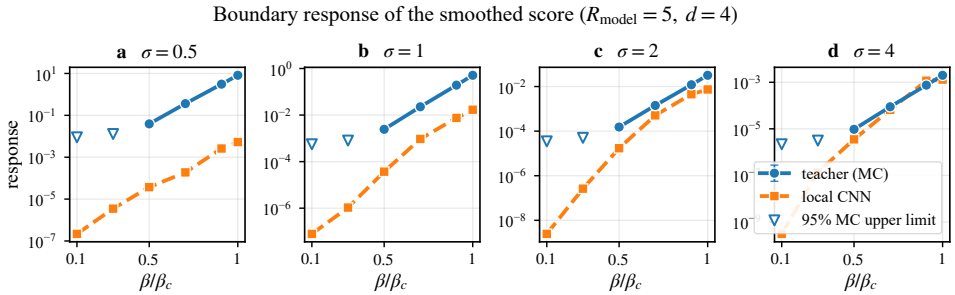


Figure 3: Boundary response of the smoothed score near the Ising critical point. A local CNN with radius $R = 5$ is evaluated at distance $d = 4$, inside its receptive field. Both scores are evaluated at $y = 0$ on a periodic 16×16 lattice. The response is the mean of $|\partial s_i/\partial y_j|$ over the 32 sites at distance four from the center i. Here s is the teacher or CNN score. Distance is the largest coordinate separation after periodic wrapping. This is a derivative, not a finite score change. The teacher response is estimated from eight independent Monte Carlo sampling runs. We use signed covariance estimates in Eq. (9) to estimate the nonnegative teacher derivatives. Filled points show means and standard errors. Hollow triangles show one-sided 95% upper limits when the response is not distinguishable from zero. The detected teacher response grows strongly as $\beta/\beta_c \to 1$.

Eq. (9), the smoothed-score response inherits critical correlations even though the clean model has only nearest-neighbor interactions. In terms of Eq. (5), the response tail grows because posterior covariance decays more slowly. We therefore test whether the score error of a fixed local CNN grows as the system size increases near the critical point.

In Figure 3, we compare the teacher's and the student CNN's center-score responses to a perturbation four lattice sites away. The teacher response is detected for $\beta/\beta_c = 0.5, 0.7, 0.9, 0.99$ and grows as β/β_c approaches one. The student CNN response is smaller at all detected points except $\sigma = 4$, $\beta/\beta_c = 0.9$, where it is larger. Near criticality, the student CNN becomes more sensitive to perturbations near the edge of its receptive field. The teacher responses at $\beta/\beta_c = 0.1$ and 0.3 are nondetections and are shown only as upper limits.

For the size-extrapolation test, we keep the CNN receptive radius fixed at $R = 5$ and the training size at $L = 16$. We perturb only the noisy observations outside the receptive field. The CNN therefore predicts no change in the center score. We estimate the teacher's center score before and after the perturbation and calculate the difference. The magnitude of this difference is the CNN's center-score change error. As shown in Figure 4, at $0.99\beta_c$, the finite score-change error increases from 0.109 with 95% interval $[0.105, 0.114]$ at $L = 16$ to 0.289 with interval $[0.279, 0.299]$ at $L = 64$. At $0.5\beta_c$, it remains of order 10^{-5} and does not grow. Thus, near criticality, the response missed by the CNN model grows with the system size. These observations confirm that the true smoothed score is non-local near criticality.

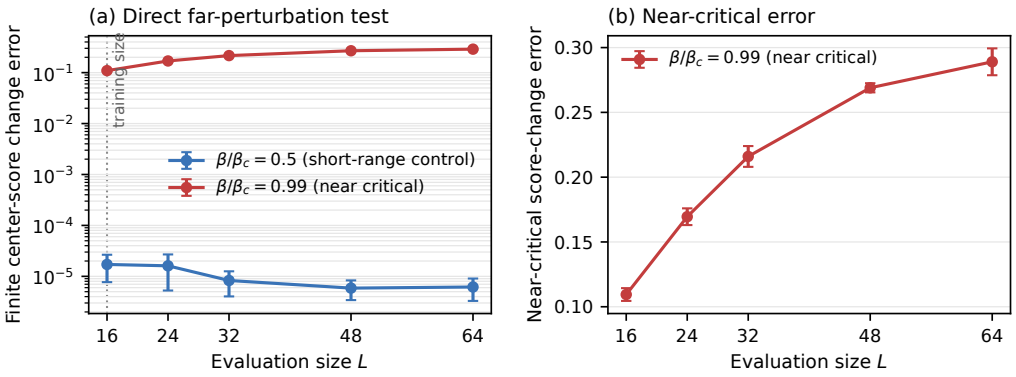


Figure 4: Direct Ising size-extrapolation test. Starting from $y = 0$, we add $\delta/\sqrt{N_{\text{far}}}$ to each noisy observation outside the CNN's receptive field and leave all observations inside unchanged. Here N_{far} is the number of perturbed sites. We use $R = 5$, $\sigma = 1$, and $\delta = 0.02$. The total L_2 norm of the perturbation equals δ at every system size. Left: finite center-score change error for the near-critical and short-range cases on a logarithmic scale. Right: the same near-critical data on a linear scale. Points are means with 95% intervals across eight independent Monte Carlo sampling runs. Sampling details are given in the code README.

5 Conclusion

Our results give a simple picture of local size extrapolation. A local model transfers across system sizes when its receptive field covers the important response of the true smoothed score. This gives a practical standard for evaluating local scientific diffusion models. A claim of size extrapolation should show that the data-generating rule is shared across sizes, the target score response is local or quasi-local, and the model receptive field covers that response. Under these checks, local score rules can transfer across size. When the smoothed-score response remains outside the model's field, local extrapolation should be treated as a measurable failure mode rather than assumed away by architectural locality.

Reproducibility Statement

The code package contains the teacher states, model configurations, training seeds, checkpoints, raw CSV/JSON results, and scripts that generate figures from these results. It also includes checks of student locality and Ising sampling. The code README gives the experiment settings and reproduction commands.

AI Use Statement

We used generative AI tools to assist with code implementation, experimental design and result interpretation, mathematical proofs, and manuscript revision. The authors take responsibility for the final content of this work.

References

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13:2453, 2022.

Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters*, 98:146401, 2007.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *International Conference on Learning Representations*, 2023.

Tiangang Cui, Shuigen Liu, and Xin T. Tong. Stein's method for marginals on large graphical models. *arXiv preprint arXiv:2410.11771*, 2024.

Valentin de Bortoli, Romuald Elie, Anna Kazeykina, Zhenjie Ren, and Jiacheng Zhang. Dimension-free error estimate for diffusion model and optimal scheduling. *arXiv preprint arXiv:2512.01820*, 2025.

Luca Maria Del Bono, Giulio Biroli, Patrick Charbonneau, and Marylou Gabrié. The critical slowing down in diffusion models. *arXiv preprint arXiv:2605.12597*, 2026.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. *International Conference on Learning Representations*, 2017.

R. L. Dobrushin. The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory of Probability and Its Applications*, 13(2):197–224, 1968.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Michael E. Fisher and Michael N. Barber. Scaling theory for finite-size effects in the critical region. *Physical Review Letters*, 28(23):1516–1519, 1972.

Hans-Otto Georgii. *Gibbs Measures and Phase Transitions*. De Gruyter, second edition, 2011.

Georg A. Gottwald, Shuigen Liu, Youssef Marzouk, Sebastian Reich, and Xin T. Tong. Localized diffusion models. *arXiv preprint arXiv:2505.04417*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *International Conference on Machine Learning*, pages 8867–8887, 2022.

Fangjun Hu, Guangkuo Liu, Yifan F. Zhang, and Xun Gao. Local diffusion models and phases of data distributions. *arXiv preprint arXiv:2508.06614*, 2025.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. In *Advances in Neural Information Processing Systems*, 2022.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. In *Advances in Neural Information Processing Systems*, 2022.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *International Conference on Learning Representations*, 2021.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3:218–229, 2021.

Kyle Mills, Kevin Ryczko, Iryna Luchak, Adam Domurad, Chris Beeler, and Isaac Tamblyn. Extensive deep neural networks for transferring small scale learning to large scale systems. *Chemical Science*, 10:4129–4140, 2019.

Lars Onsager. Crystal statistics. I. A two-dimensional model with an order-disorder transition. *Physical Review*, 65(3–4):117–149, 1944.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Vladimir Privman, editor. *Finite Size Scaling and Numerical Simulation of Statistical Systems*. World Scientific, 1990.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538, 2015.

Herbert E. Robbins. An empirical Bayes approach to statistics. In *Breakthroughs in Statistics*, pages 388–394. Springer, 1992.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *International Conference on Machine Learning*, pages 9323–9332, 2021.

Kristof T. Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: a continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems*, 2017.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2021.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Dror Weitz. Counting independent sets up to the tree threshold. In *ACM Symposium on Theory of Computing*, pages 140–149, 2006.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. GeoDiff: A geometric diffusion model for molecular conformation generation. *International Conference on Learning Representations*, 2022.

A Theory Details and Proofs

We derive the spatial response of clean and smoothed scores and prove the size-independent bound on local generation error. This appendix gives the complete statements and proofs behind Section 2.

A.1 Clean and Smoothed Score Locality

Clean-score locality follows from local interactions. Smoothed-score locality depends on posterior correlations.

Proposition 1 (Local log-density implies local clean score)**.** *Here* $\mathrm{dist}(i,j)$ *denotes graph distance, the number of edges in a shortest path between sites, and* $B_r(i) = \{j : \mathrm{dist}(i,j) \le r\}$ *is the radius-*r *neighborhood of site* i*. Suppose that for every* L*,*

$$\log p_L(x) = \sum_{j \in \Lambda_L} \phi_j(x_{B_r(j)}),$$

where r *is independent of* L*. Then the clean score component*

$$s_{L,0}(x)_i = \nabla_{x_i} \log p_L(x)$$

depends only on $x_{B_{2r}(i)}$*, up to the usual boundary convention. In particular, its dependence radius is independent of* L*.*

Proof. Differentiating the local decomposition gives

$$\nabla_{x_i} \log p_L(x) = \sum_{\substack{j \in \Lambda_L \\ i \in B_r(j)}} \nabla_{x_i} \phi_j(x_{B_r(j)}).$$

If $i \in B_r(j)$ and $k \in B_r(j)$, then $\mathrm{dist}(i,k) \le \mathrm{dist}(i,j) + \mathrm{dist}(j,k) \le 2r$. Hence every term that contributes to the score at i depends only on sites in $B_{2r}(i)$. □

Proposition 2 (Tweedie formula and covariance response)**.** *Let* $X \sim p_L$*,* $Y = X + \sigma\varepsilon$*, and* $\varepsilon \sim \mathcal{N}(0, I)$*. Define*

$$p_{L,\sigma}(y) = \int p_L(x) \mathcal{N}(y; x, \sigma^2 I)\, dx$$

and the noisy posterior

$$\pi^y_{L,\sigma}(dx) \propto p_L(x) \exp\left(-\frac{\|y - x\|_2^2}{2\sigma^2}\right) dx.$$

Then

$$s_{L,\sigma}(y) = \nabla_y \log p_{L,\sigma}(y) = \sigma^{-2}\big(\mathbb{E}_{\pi^y_{L,\sigma}}[X] - y\big).$$

Moreover, for vector-valued site variables $X_i \in \mathbb{R}^d$*,*

$$\frac{\partial s_{L,\sigma}(y)_i}{\partial y_j} = \sigma^{-4} \operatorname{Cov}_{\pi^y_{L,\sigma}}(X_i, X_j) - \sigma^{-2} \mathbf{1}_{\{i=j\}} I_d.$$

Here I_d *is the* $d \times d$ *identity matrix, and* $\mathbf{1}_{\{i=j\}}$ *equals one for* $i = j$ *and zero otherwise.*

Proof. Differentiating the Gaussian convolution under the integral gives

$$\nabla_y p_{L,\sigma}(y) = \int p_L(x) \frac{x - y}{\sigma^2} \mathcal{N}(y; x, \sigma^2 I)\, dx.$$

Dividing by $p_{L,\sigma}(y)$ yields Tweedie's formula.

For the response identity, write $m_i(y) = \mathbb{E}_{\pi^y_{L,\sigma}}[X_i]$. The posterior log weight has derivative

$$\frac{\partial}{\partial y_j} \log\left[p_L(x) \exp\left(-\frac{\|y - x\|_2^2}{2\sigma^2}\right)\right] = \sigma^{-2}(x_j - y_j).$$

Using the standard derivative-of-expectation identity,

$$\frac{\partial m_i(y)}{\partial y_j} = \sigma^{-2} \operatorname{Cov}_{\pi^y_{L,\sigma}}(X_i, X_j).$$

Since $s_{L,\sigma}(y)_i = \sigma^{-2}(m_i(y) - y_i)$, differentiating once more gives the stated Jacobian formula. □

Assumption 1 (Uniform noisy-posterior covariance decay)**.** *We first fix the noisy observation* $Y = y$*. This assumption bounds the covariance between clean variables at distant sites. The decay rate does not deteriorate with* L*. The operator norm* $\|\cdot\|_{\mathrm{op}}$ *is the largest factor by which a matrix can increase a vector's length. For a fixed noise level* $\sigma > 0$*, suppose there are constants* $C_\sigma < \infty$ *and* $\mu_\sigma > 0$*, independent of* L*, such that for all* $i, j \in \Lambda_L$ *and all relevant* y*,*

$$\left\| \operatorname{Cov}_{\pi^y_{L,\sigma}}(X_i, X_j) \right\|_{\mathrm{op}} \le C_\sigma e^{-\mu_\sigma \mathrm{dist}(i,j)}.$$

Under Assumption 1, Proposition 2 gives, for $i \neq j$,

$$\left\| \frac{\partial s_{L,\sigma}(y)_i}{\partial y_j} \right\|_{\mathrm{op}} \leq \sigma^{-4} C_\sigma e^{-\mu_\sigma \mathrm{dist}(i,j)}.$$

Thus the smoothed-score response decays with distance at a rate independent of system size.

We next study how a perturbation spreads during generation. The reverse drift b_s^L is the deterministic update field of the exact reverse process. Its Jacobian block $\partial b_{s,i}^L / \partial x_j$ measures how the update at site i responds to a change at site j. If these blocks decay sufficiently fast with distance, a local perturbation has little influence outside an expanding region.

For a patch A and a function $\phi(x_A)$ that reads only this patch, define

$$P_{t,T}^L \phi(x) = \mathbb{E}[\phi(X_T^{t,x})],$$

Here $X_T^{t,x}$ is the state at time T of the exact process started from x at time t. The quantity $P_{t,T}^L \phi(x)$ is the expected final value of the patch observable ϕ.

Proposition 3 (Drift-Jacobian bound on response spreading)**.** *Consider the exact reverse diffusion*

$$dX_s = b_s^L(X_s)\, ds + \sqrt{2}\, dW_s, \qquad t \leq s \leq T,$$

where W_s is standard Brownian noise with independent components, and assume b_s^L is continuously differentiable in x. Suppose there exist $\mu_0 > 0$ and $\alpha_{\mu_0} < \infty$, independent of L, such that

$$\sup_{L,s,x,i} \sum_{j \in \Lambda_L} e^{\mu_0 \mathrm{dist}(i,j)} \left\| \frac{\partial b_{s,i}^L(x)}{\partial x_j} \right\|_{\mathrm{op}} \leq \alpha_{\mu_0}.$$

Then for every patch A, every smooth patch test function $\phi(x_A)$, and every $i \in \Lambda_L$,

$$\|\nabla_i P_{t,T}^L \phi(x)\| \leq |A| \,\mathrm{Lip}(\phi)\, \exp[-\mu_0 \mathrm{dist}(i,A) + \alpha_{\mu_0}(T-t)]\,.$$

The gradient ∇_i differentiates with respect to the variable x_i at site i. Here $\mathrm{dist}(i,A) = \min_{a \in A} \mathrm{dist}(i,a)$ is the distance from site i to the patch. The constant $\mathrm{Lip}(\phi)$ bounds how fast ϕ changes with its input. Let $v = \alpha_{\mu_0}/\mu_0$. The influence is concentrated within a distance of order $v(T-t)$ from the patch and decays exponentially outside it. We call this expanding region the dynamic light cone. It provides the limited-spread condition used in Appendix A.2.

Proof. Fix t, x. The matrix $U_{ai}(s) = \partial X_{s,a}^{t,x} / \partial x_i$ measures how site a at time s responds to an initial perturbation at site i. The noise is additive, so this matrix satisfies

$$U_{ai}(t) = \mathbf{1}_{\{a=i\}} I_d, \qquad \frac{d}{ds} U_{ai}(s) = \sum_{k \in \Lambda_L} \frac{\partial b_{s,a}^L(X_s^{t,x})}{\partial x_k} U_{ki}(s).$$

Let $M_{ai}(s) = \|U_{ai}(s)\|_{\mathrm{op}}$. Taking operator norms gives the standard differential inequality

$$\frac{d}{ds} M_{ai}(s) \leq \sum_k \left\| \frac{\partial b_{s,a}^L(X_s^{t,x})}{\partial x_k} \right\|_{\mathrm{op}} M_{ki}(s)$$

When an ordinary derivative is unavailable, we use the upper-Dini derivative. This is the upper limiting right-hand rate of change. Define the weighted quantity

$$N_i(s) = \sup_a e^{\mu_0 \mathrm{dist}(a,i)} M_{ai}(s).$$

Using the triangle inequality $\mathrm{dist}(a,i) \leq \mathrm{dist}(a,k) + \mathrm{dist}(k,i)$, we obtain

$$\frac{d}{ds} N_i(s) \leq \alpha_{\mu_0} N_i(s), \qquad N_i(t) \leq 1.$$

Gronwall's inequality bounds the growth of $N_i(s)$. It gives

$$M_{ai}(T) \leq \exp[\alpha_{\mu_0}(T-t)] e^{-\mu_0 \mathrm{dist}(a,i)}.$$

Finally,

$$\nabla_i P_{t,T}^L \phi(x) = \mathbb{E}\left[\sum_{a \in A} (\nabla_a \phi)(X_{T,A}^{t,x})\, U_{ai}(T) \right],$$

so

$$\|\nabla_i P_{t,T}^L \phi(x)\| \leq |A| \,\mathrm{Lip}(\phi)\, e^{\alpha_{\mu_0}(T-t)} e^{-\mu_0 \mathrm{dist}(i,A)}.$$

For fixed $|A| \leq m$ and fixed T, the sum of this response bound over all sites remains bounded independently of L on the lattice families considered in the paper. □

A.2 Size-Uniform Control of Fixed-Patch Distributions

We prove the size-independent bound on fixed-patch distribution error. Let Λ_L be a finite graph with graph distance $\mathrm{dist}(\cdot,\cdot)$, and let $x_i \in \mathbb{R}^d$. A patch $A \subset \Lambda_L$ is a fixed set of sites. Its marginal is the distribution of x_A obtained by integrating out the variables outside A.

To compare two patch distributions μ and ν, we use the bounded-Lipschitz distance

$$d_{\mathrm{BL}}(\mu,\nu) = \sup_{\phi \in \mathrm{BL}_1(A)} \left| \int \phi \, d\mu - \int \phi \, d\nu \right|,$$

where

$$\mathrm{BL}_1(A) = \{\phi : \mathbb{R}^{d|A|} \to \mathbb{R} : \|\phi\|_\infty \le 1,\ \mathrm{Lip}(\phi) \le 1\}.$$

Each observable in $\mathrm{BL}_1(A)$ has magnitude at most one. Its change is no larger than the distance between its two inputs. The distance d_{BL} measures the largest difference in its expectation under the two distributions. We extend each observable to the full system by writing $\phi(x) = \phi(x_A)$.

Let X_t and $\widehat{X}_t$ denote the exact and learned reverse processes. Generation starts from noise at $t = 0$ and ends at $t = T$. The deterministic update fields are b_t^L and $\widehat{b}_t^L$. The term W_t denotes standard Brownian noise. The processes satisfy

$$dX_t = b_t^L(X_t)\,dt + \sqrt{2}\,dW_t, \qquad d\widehat{X}_t = \widehat{b}_t^L(\widehat{X}_t)\,dt + \sqrt{2}\,dW_t, \qquad 0 \le t \le T.$$

Their initial distributions are $X_0 \sim \rho_0^L$ and $\widehat{X}_0 \sim \widehat{\rho}_0^L$. For a smooth observable f, the generator describes its instantaneous expected change under the process:

$$L_t^L f = \sum_{i \in \Lambda_L} b_{t,i}^L \cdot \nabla_i f + \sum_{i \in \Lambda_L} \Delta_i f,$$

Here $\Delta_i f = \sum_{a=1}^d \partial^2 f / \partial x_{i,a}^2$ sums second derivatives over the components at site i. We define $\widehat{L}_t^L$ in the same way, with b^L replaced by $\widehat{b}^L$. Hence

$$(\widehat{L}_t^L - L_t^L) f = \sum_{i \in \Lambda_L} (\widehat{b}_{t,i}^L - b_{t,i}^L) \cdot \nabla_i f.$$

The difference $\widehat{b}_t^L - b_t^L$ is the drift error. For Gaussian noising with forward noise variance increasing at rate two, the reverse noise level satisfies $\sigma(t)^2 = \sigma_{\mathrm{start}}^2 - 2t$. Here σ_{start} is the initial noise level. With the $\sqrt{2}$ noise convention above,

$$b_t^L(x) = 2 s_{L,\sigma(t)}(x), \qquad \widehat{b}_t^L(x) - b_t^L(x) = 2\big[s_\theta(x,\sigma(t)) - s_{L,\sigma(t)}(x)\big].$$

Thus the drift error is twice the score error in this convention.

For the exact process started from x at time s, the expected value of an observable at time t is

$$P_{s,t}^L \phi(x) = \mathbb{E}[\phi(X_t^{s,x})],$$

We write $\widehat{\nu}_t^L = \mathrm{Law}(\widehat{X}_t)$ for the distribution of the learned process at time t. The exact and learned final patch distributions are $P_{L,A}$ and $\widehat{Q}_{L,A}$, respectively.

Assumption 2 (Backward regularity)**.** *This assumption ensures that both reverse processes and the expected patch observables used below are well defined. For each finite L, both reverse processes are well posed on $[0,T]$. For every smooth bounded patch test function ϕ, $u(t,x) = P_{t,T}^L \phi(x)$ solves*

$$\partial_t u + L_t^L u = 0, \qquad u(T,x) = \phi(x),$$

and Itô's formula applies to $u(t,\widehat{X}_t)$. This formula is the stochastic version of the chain rule. We assume that its random fluctuation term has mean zero.

Assumption 3 (Consistent initialization on fixed patches)**.** *For every fixed $m < \infty$, there exists $\delta_{0,m} \ge 0$, independent of L, such that for every $A \subset \Lambda_L$ with $|A| \le m$ and every smooth $\phi \in \mathrm{BL}_1(A)$,*

$$\left| \mathbb{E}_{\widehat{\rho}_0^L}[P_{0,T}^L \phi] - \mathbb{E}_{\rho_0^L}[P_{0,T}^L \phi] \right| \le \delta_{0,m}.$$

The quantity $\delta_{0,m}$ bounds the final patch error caused by different initial noise distributions. It can be zero when both processes start from the same distribution.

Assumption 4 (Limited spread of influence (dynamic quasi-locality)). *The gradient $\nabla_i P^L_{t,T}\phi$ measures how the final patch observable responds to the state at site i and time t. We bound this response using an influence weight $\omega^L_{A,t}(i)$. The total weight must remain bounded as L grows. We call this property dynamic quasi-locality. For every fixed $m < \infty$, assume there are weights $\omega^L_{A,t}(i) \geq 0$ and a constant $C_{\mathrm{loc}}(m,T) < \infty$, independent of L, such that*

$$\sup_L \sup_{\substack{A\subset\Lambda_L \\ |A|\leq m}} \sup_{t\in[0,T]} \sum_{i\in\Lambda_L} \omega^L_{A,t}(i) < \infty,$$

and for every smooth $\phi \in \mathrm{BL}_1(A)$,

$$\|\nabla_i P^L_{t,T}\phi(x)\| \leq C_{\mathrm{loc}}(m,T)\, \omega^L_{A,t}(i)$$

for all L, A, t, i, x. A useful example is an exponentially decaying dynamic light cone,

$$\omega^L_{A,t}(i) = \exp[-\mu(\mathrm{dist}(i,A) - v(T-t))_+],$$

Here $(z)_+ = \max(z,0)$. The constant $\mu > 0$ sets the spatial decay rate. The constant v sets how fast the region of influence expands. The weight equals one inside this region and decays exponentially outside it. On the lattices considered here, the total weight stays bounded for fixed patch size and generation time.

Assumption 5 (Influence-weighted drift error along learned trajectories). *The learned process visits states with distribution $\widehat{\nu}^L_t$. We average the drift error over these states and accumulate it over time. Each site's contribution is weighted by its influence on the patch. For every fixed $m < \infty$, assume there is a bound $\eta_m \geq 0$, independent of L, such that for every $A \subset \Lambda_L$ with $|A| \leq m$,*

$$\int_0^T \mathbb{E}_{\widehat{\nu}^L_t}\left[\sum_{i\in\Lambda_L} \omega^L_{A,t}(i)\|\widehat{b}^L_{t,i} - b^L_{t,i}\|\right] dt \leq \eta_m.$$

The following comparison converts the accumulated effect of drift error on every patch observable into a bound on the generated patch distribution.

Proposition 4 (Patch-distribution comparison). *Assume Assumptions 2 and 3. If, for every A with $|A| \leq m$ and every smooth $\phi \in \mathrm{BL}_1(A)$,*

$$\int_0^T \left|\mathbb{E}_{\widehat{\nu}^L_t}\left[(\widehat{L}^L_t - L^L_t)P^L_{t,T}\phi\right]\right| dt \leq \varepsilon_m$$

with ε_m independent of L, then

$$\sup_L \sup_{\substack{A\subset\Lambda_L \\ |A|\leq m}} d_{\mathrm{BL}}(\widehat{Q}_{L,A}, P_{L,A}) \leq \delta_{0,m} + \varepsilon_m.$$

Proof. Fix L, A, and a smooth $\phi \in \mathrm{BL}_1(A)$. Set $u(t,x) = P^L_{t,T}\phi(x)$. By Assumption 2,

$$\partial_t u + L^L_t u = 0, \qquad u(T,x) = \phi(x).$$

Itô's formula along the learned process gives

$$\mathbb{E}[\phi(\widehat{X}_{T,A})] = \mathbb{E}_{\widehat{\rho}^L_0}[P^L_{0,T}\phi] + \int_0^T \mathbb{E}_{\widehat{\nu}^L_t}[(\widehat{L}^L_t - L^L_t)P^L_{t,T}\phi]\, dt.$$

For the exact process,

$$\mathbb{E}[\phi(X_{T,A})] = \mathbb{E}_{\rho^L_0}[P^L_{0,T}\phi].$$

Subtracting these equations and applying the two assumed bounds gives

$$\left|\mathbb{E}[\phi(\widehat{X}_{T,A})] - \mathbb{E}[\phi(X_{T,A})]\right| \leq \delta_{0,m} + \varepsilon_m.$$

Smooth functions can approximate every test function in $\mathrm{BL}_1(A)$. These approximations preserve the bounds on magnitude and rate of change. Dominated convergence then extends the inequality to all $\phi \in \mathrm{BL}_1(A)$. Taking the supremum over ϕ, A, and L proves the claim. □

Theorem 2 (Size-uniform control of fixed-patch distributions)**.** *Under Assumptions 2–5, for every fixed $m < \infty$,*

$$\sup_L \sup_{\substack{A \subset \Lambda_L \\ |A| \le m}} d_{\mathrm{BL}}(\widehat{Q}_{L,A}, P_{L,A}) \le \delta_{0,m} + C_{\mathrm{loc}}(m,T)\eta_m.$$

The right-hand side contains no factor that grows with the total system size L.

Proof. It suffices to verify the comparison condition in Proposition 4. For smooth $\phi \in \mathrm{BL}_1(A)$,

$$(\widehat{L}_t^L - L_t^L)P_{t,T}^L\phi = \sum_{i \in \Lambda_L} (\widehat{b}_{t,i}^L - b_{t,i}^L) \cdot \nabla_i P_{t,T}^L \phi.$$

Therefore, by the limited-spread condition in Assumption 4,

$$\begin{aligned}\left|\mathbb{E}_{\widehat{\nu}_t^L}[(\widehat{L}_t^L - L_t^L)P_{t,T}^L\phi]\right| &\le \mathbb{E}_{\widehat{\nu}_t^L} \sum_i \|\widehat{b}_{t,i}^L - b_{t,i}^L\| \|\nabla_i P_{t,T}^L \phi\| \\ &\le C_{\mathrm{loc}}(m,T)\mathbb{E}_{\widehat{\nu}_t^L} \sum_i \omega_{A,t}^L(i) \|\widehat{b}_{t,i}^L - b_{t,i}^L\|.\end{aligned}$$

After integrating over time, Assumption 5 gives $\varepsilon_m = C_{\mathrm{loc}}(m,T)\eta_m$. Proposition 4 then proves the stated bound. □

The bound depends on the score error that can influence the patch during generation. Distant errors carry less weight. The total system size does not enter as an extra multiplier.

A.3 Exact Reverse-Process Dynamic Response

We measure how a one-site perturbation spreads during an exact reverse process. We use two Gaussian fields on a periodic 2D lattice with $L = 128$. Both fields penalize differences between neighboring sites and deviations from zero. The mass parameter controls the local penalty. A smaller mass produces longer-range correlations. Writing the mass as $m_{\mathrm{G}} > 0$, the clean density is

$$p(x) \propto \exp\left[-\frac{1}{2}\left(m_{\mathrm{G}}^2 \sum_i x_i^2 + \sum_{\langle i,j \rangle} (x_i - x_j)^2\right)\right].$$

Each nearest-neighbor pair $\langle i,j \rangle$ is counted once. Let Q be the matrix in this quadratic form, so that the exponent is $-x^\top Qx/2$. Equivalently, $(Qx)_i = m_{\mathrm{G}}^2 x_i + \sum_{j:\, j \text{ neighbors } i}(x_i - x_j)$. The clean covariance is Q^{-1}, and the smoothed score is $s_\sigma(x) = -(Q^{-1} + \sigma^2 I)^{-1}x$. Here I is the identity matrix on the full lattice.

We compare two trajectories driven by the same noise. They differ initially at one site. The perturbation is applied at noise scale 4. Let $f \in [0,1]$ be the completed fraction of reverse time. The remaining noise variance is $\sigma_f^2 = (1-f)\sigma_{\mathrm{start}}^2$, with $\sigma_{\mathrm{start}} = 4$. The unperturbed initial state follows $\mathcal{N}(0, Q^{-1} + \sigma_{\mathrm{start}}^2 I)$. The same Brownian noise cancels when we subtract the two trajectories. The linear score then gives the exact response matrix

$$J_f = \frac{\partial X_f}{\partial X_{\mathrm{start}}} = (I + \sigma_f^2 Q)(I + \sigma_{\mathrm{start}}^2 Q)^{-1}.$$

Here X_f denotes the state after fraction f of generation. A unit initial change at site i produces the change $(J_f)_{ji}$ at site j. Figure 5 shows its exact response after $1/4$, $1/2$, $3/4$, and all of the reverse process.

At distance r, each curve shows $\max_{j:\, d_\infty(i,j)=r} |(J_f)_{ji}|$. The distance d_∞ is the largest coordinate separation after periodic wrapping. We also report the fraction of total absolute response outside radius R,

$$\frac{\sum_{j:\, d_\infty(i,j)>R} |(J_f)_{ji}|}{\sum_j |(J_f)_{ji}|}.$$

For mass 1.0, the fraction of total absolute response outside radius 4 increases from 0.00038 after $1/4$ of the process to 0.01997 at the endpoint. For mass 0.1, it increases from 0.103 to 0.463. The same initial one-site change can therefore produce very different amounts of distant influence.

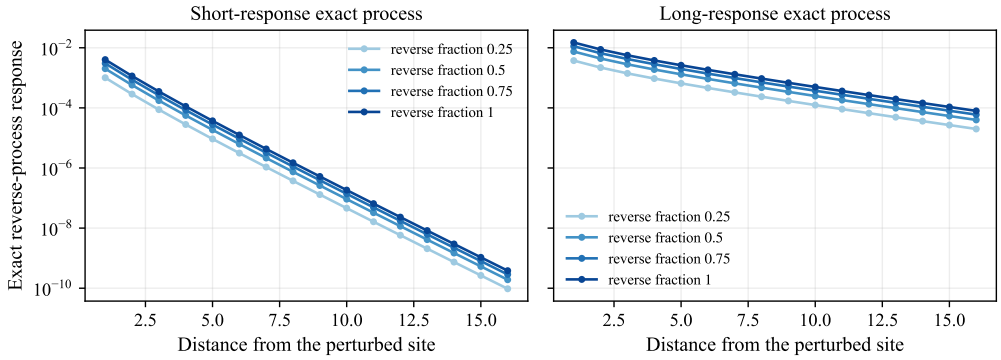


Figure 5: Exact reverse-process response to a one-site perturbation. Curves report the largest absolute response at each distance after the indicated fraction of the reverse process. The short-response process uses mass 1.0, the long-response process uses mass 0.1, and both start at noise scale 4. These curves are exact calculations and therefore have no sampling error bars.

B FDLF Construction and Additional Results

We describe the teacher construction and test size extrapolation in 3D. We also measure teacher responses in 2D.

B.1 Teacher Construction

The 2D and 3D teachers generate a continuous field on a periodic square or cubic lattice. Each site carries one real variable. Opposite edges or faces are identified, so updates wrap around the lattice. We use an even side length L. Starting from independent $u_i \sim \mathcal{N}(0,1)$, they apply the map $x = T_L(u) = S(T_{\text{flow}}(T_{\text{pre}}(u)))$. The checkerboard divides sites into two groups according to whether the sum of their coordinates is even or odd. Each layer uses one group to compute an update to the other.

Scale and shift updates. For one layer, let A be the unchanged group and B the updated group. Write v for the layer input and v' for its output. The masked field v_A equals v on A and zero on B. A CNN reads v_A and the binary mask of B. It returns a log-scale $a_i(v_A)$ and a shift $b_i(v_A)$. The update is

$$v'_i = \begin{cases} v_i, & i \in A, \\ v_i \exp(a_i(v_A)) + b_i(v_A), & i \in B. \end{cases}$$

The CNN uses three periodic convolutions with a 3×3 kernel in 2D or a $3 \times 3 \times 3$ kernel in 3D. Between them, it applies the scalar function $h(z) = z/(1+e^{-z})$ separately at each site and channel. We bound the log-scale by replacing its raw value $\widetilde{a}_i$ with $a_i = 1.5 \tanh(\widetilde{a}_i/1.5)$. The pre-flow and flow stages T_{pre} and T_{flow} each compose a fixed number of these layers with alternating updated groups and separate fixed CNN parameters.

The inverse leaves A unchanged and gives $v_i = (v'_i - b_i(v'_A)) \exp(-a_i(v'_A))$ on B. It uses the same scale and shift because $v'_A = v_A$. If F denotes this layer, then

$$\log |\det DF(v)| = \sum_{i \in B} a_i(v_A).$$

Here DF is the matrix of derivatives of the output with respect to the input. Its determinant is the local volume-change factor. Ordering A before B makes this matrix triangular in blocks, which gives the formula above.

Additive smoother. The smoother S applies additive updates. It again alternates the updated group. A weighted sum of source values gives

$$M_i = \sum_{j \in A} K(i-j) v_j, \qquad v'_i = \begin{cases} v_i, & i \in A, \\ v_i + \alpha \left[M_i + \lambda H_i \right], & i \in B. \end{cases}$$

The kernel K specifies the spatial weights. The constant α sets the update amplitude. The nonlinear correction is $H_i = 2\tanh(g_i(v_A, M, \mathbf{1}_B)/2)$, with $\lambda = 1$. Here $\mathbf{1}_B$ is the binary update mask, and g is a CNN with the same convolution sizes and scalar nonlinearity as above. Its inputs are the masked source field, the weighted field M, and the mask. Both M and H depend only on the unchanged group. The inverse therefore subtracts the same additive update, and this layer has log-Jacobian determinant zero.

Teacher families. The short-range simple teacher uses local coupling layers without a smoother. The short-range structured teacher adds a normalized nearest-neighbor smoother and a local nonlinear correction. For this teacher, $K(i-j) = 1/4$ on the four nearest neighbors in 2D and $1/6$ on the six nearest neighbors in 3D. All other weights are zero. The 2D medium-range teacher instead uses weights proportional to $r^{-0.75}$ for $1 \le r \le 5$, where r is the Euclidean distance after periodic wrapping. Only offsets connecting opposite checkerboard groups are included. We divide these weights by their sum, so $\sum_j K(i-j) = 1$. Both finite-range smoothers use $\alpha = 1$.

The long-range teacher instead uses an exponential kernel $\exp(-r/60)$, where r is the periodic Euclidean distance. This kernel covers the whole torus on opposite-group offsets and is not renormalized as L grows. Its update amplitude α is 0.0015 in 2D and 2.5×10^{-5} in 3D. Each amplitude stays fixed across system sizes. Larger systems include more distant sites, so the total coupling can increase under the same rule.

The simple and structured teachers each use one pre-flow layer and one flow layer. The simple teacher has no smoother, and the structured teacher has one additive layer. The medium-range and long-range teachers each use two pre-flow layers and two flow layers. They use three and six additive layers, respectively. These maps have exact inverses and known Jacobian determinants, so they give exact densities and clean scores. To evaluate the density, we invert the layers in reverse order and sum their inverse log-Jacobian determinants in Eq. (12). Differentiating this log density gives the clean score.

The 2D and 3D experiments use the same teacher roles and student receptive radius. They compare stable and failing regimes across dimensions, not numerical growth rates. CNN widths, fixed teacher parameters, randomization settings, teacher seeds, and student parameter counts are listed in the code README.

B.2 3D Continuous Diagnostic

We test whether the 2D stable and failing regimes also appear in 3D. We train each periodic CNN at $L = 16$ and evaluate it at $L = 16, 24, 32, 40, 48$. Every CNN has one residual block, pointwise normalization, and $R_{\text{eff}} = 4$.

We also measure the teacher's exact clean-score response to Gaussian perturbation directions. Near directions act inside a radius-four cube around the output site, where $d_\infty(i,j) \le 4$. Far directions act outside that cube. A Gaussian direction assigns independent standard normal values to its allowed sites and zero elsewhere. For a configuration x, an output site i, and a direction ζ, the response is the directional derivative

$$D_i(x,\zeta) = \frac{d}{d\epsilon} s_{L,0}(x + \epsilon\zeta)_i \Big|_{\epsilon=0}.$$

It measures the infinitesimal change in the center score per unit perturbation amplitude. We evaluate both checkerboard groups and combine their squared responses. For configuration a and checkerboard group $p \in \{0,1\}$, let $D^{\text{far}}_{a,p}$ and $D^{\text{near}}_{a,p}$ be these derivatives for independent far and near directions. The far-to-near ratio is therefore

$$\left[\frac{\sum_{a=1}^{64}\sum_{p=0}^{1}(D^{\text{far}}_{a,p})^2}{\sum_{a=1}^{64}\sum_{p=0}^{1}(D^{\text{near}}_{a,p})^2}\right]^{1/2}.$$

The ratio is not normalized by the number of perturbed sites. The code README describes the sampling and uncertainty estimates.

Figure 6 shows stable clean-score error for both short-range teachers. For the simple teacher, student RMSE changes from 0.386 ± 0.011 at $L = 16$ to 0.384 ± 0.011 at $L = 48$. For the structured

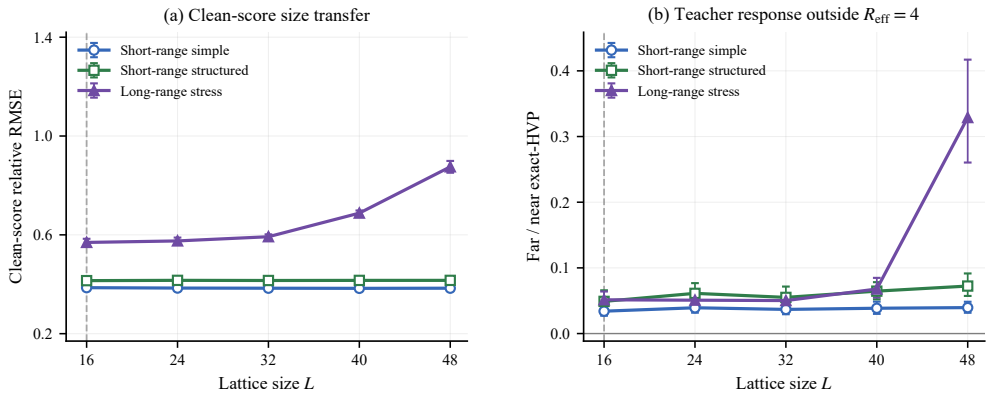


Figure 6: Periodic 3D continuous diagnostic. Models use $R_{\text{eff}} = 4$, are trained at $L = 16$, and are evaluated up to $L = 48$. Left: student clean-score relative RMSE. Error bars are s.e.m. across three independent student training seeds. Right: the ratio of the teacher's score response to perturbations outside and inside $R_{\text{eff}} = 4$. Error bars are 95% intervals from resampling 64 configurations.

teacher, it changes from 0.415 ± 0.001 to 0.416 ± 0.001. For the long-range teacher, it increases from 0.569 ± 0.015 to 0.875 ± 0.024.

The short-range teachers keep far-to-near response ratios below 0.10. The long-range ratio increases from 0.051, with 95% interval $[0.041, 0.064]$, at $L = 16$ to 0.329, with interval $[0.260, 0.417]$, at $L = 48$. The student error rises sharply from $L = 32$, while the response ratios separate clearly at $L = 48$. These measurements show the same stable and failing regimes as in 2D.

B.3 Teacher-Only Response Diagnostics

Figure 7 shows the teacher's clean-score response in 2D. Here we measure finite score changes. Near perturbations act inside the central 5×5 patch, and far perturbations act outside it, as in Figure 1. For a clean sample x, we draw independent standard Gaussian values ζ_j and retain only those in the near or far region. The center-score change is

$$\Delta s_i^q = s_{L,0}(x + \epsilon \zeta^q)_i - s_{L,0}(x)_i, \qquad q \in \{\text{near}, \text{far}\}.$$

The amplitude ϵ is 0.1 times the root mean square of the clean field in each evaluation batch. We calculate the root mean square of Δs_i^q across samples in each batch and average it over batches. For each evaluation run, the ratio divides the resulting far response by the near response. The figure averages these ratios over the three recorded runs. We do not divide by the number of perturbed sites or by the total perturbation norm. The short-range teachers have small far-to-near response ratios. The long-range teacher's ratio grows strongly with system size.

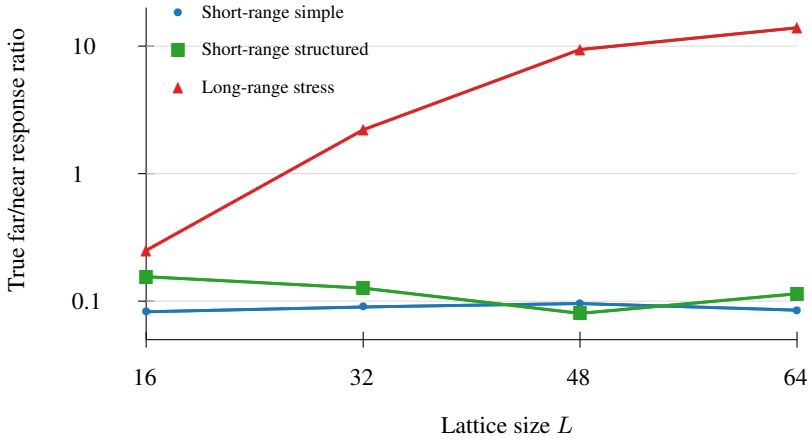


Figure 7: Teacher-only true far/near response ratios in the 2D fixed-architecture size sweep. The plotted values are computed from exact teacher score responses to finite Gaussian perturbations. Near denotes the central 5×5 patch, and far denotes its complement. The ratio uses batch-averaged root-mean-square score changes, as defined in Appendix B.3. This is a separate teacher-only check with the same teacher parameters as Figure 1.