FROM LOCAL CONTRACTIVITY TO GLOBAL GEOMETRY: POISSON POTENTIALS FOR REPRESENTATION ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Contractive autoencoders characterize representation geometry through local derivatives of an encoder, but a pointwise Jacobian statistic does not describe how these local responses are spatially organized across or away from the data distribution. Our aim is to endow learned representations with mathematical structure beyond observed samples, including regions between samples and regions of low data density. To this end, we introduce a Poisson-based framework that uses an aggregate encoder-Jacobian measure as the source of a domain-wide scalar potential. Green's identity relates the expected source to a boundary-flux term and an interior coupling between the potential gradient and the data score. Because the boundary of the effective data support is unavailable from finite samples, a corruption operator transports observations toward an off-data anchor region that serves as an operational boundary surrogate; the resulting displacements define corruption-induced measurement directions. For Gaussian corruption, the normalized denoising displacement estimates the score of the smoothed data density and aligns the bulk coupling with the denoising task. We examine whether the potential exposes non-local structure beyond its pointwise source, whether its geometry depends on the representation-learning objective, and how different corruption families condition its directional response. Two-dimensional field visualizations show that the potential is smoother and more spatially integrated than the Jacobian source, while matched autoencoder objectives produce distinguishable induced fields. Comparisons across corruption regimes suggest a contrast between diffuse responses and axis-aligned, coordinate-selective responses, while Gaussian and diffusion-style corruptions show visually similar multiscale behavior. The resulting framework provides a mathematical mechanism for organizing representation geometry across sampled, inter-sample, and low-density regions of the domain.

1 INTRODUCTION

Representation-learning objectives determine not only which information an encoder retains, but also how its output changes under perturbations of the input. Contractive autoencoders (CAEs) make this local behavior explicit by penalizing the Frobenius norm of the encoder Jacobian (Rifai et al., 2011b). For sufficiently small perturbations, the Jacobian provides the first-order change of the representation and therefore defines a local notion of encoder sensitivity. The CAE penalty encourages nearby inputs to have nearby representations, but a pointwise norm alone does not define how local Jacobian information should be organized across a region, between separated modes, or along trajectories leaving the observed data.

Corruption-based learning offers a complementary perspective. Denoising autoencoders reconstruct clean samples from corrupted observations (Vincent et al., 2008; 2010), while regularized reconstruction functions are connected to the score of a smoothed data distribution (Alain & Bengio, 2014). Diffusion models extend related score information across a schedule of noise levels (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021). These results suggest that perturbations can be used not only as augmentation, but also as controlled geometric measurement operators.

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.

Our broader aim is to endow learned representations with mathematical structure beyond observed samples, including regions between samples and regions of low data density. In this work, we pursue that aim through a Poisson-induced potential. Given an encoder f_ϕ, a nonnegative aggregate of its Jacobian acts as a source field. Solving a Poisson problem transforms this local source into a non-local scalar potential v_ϕ. We therefore propose the potential as a field induced by the spatial distribution of the local-sensitivity source, the differential operator, and the boundary realization. Green's identity then decomposes the expected source into a boundary contribution and an interior score–potential coupling.

The effective boundary of an empirical distribution is unknown. We therefore interpret a corruption operator as a stochastic transport from high-density data regions toward an off-data anchor region. When the transported samples concentrate near a thin surface surrounding the data, that region acts as an operational boundary surrogate on which the potential is softly anchored near zero. The resulting scalar is a corruption-conditioned directional response of the potential rather than an intrinsic sensitivity of the encoder.

The experiments address three questions:

RQ1 How does the Poisson construction extend sample-local encoder information across inter-sample and low-density regions?

RQ2 Do plain, denoising, contractive, and Poisson-regularized autoencoders yield distinguishable domain-wide fields?

RQ3 How do isotropic Gaussian, diffusion-style, and coordinate-deletion perturbation regimes condition the directional response of a fixed potential?

RQ1 and RQ3 form the main empirical argument. RQ2 is a supporting control that checks whether the observed field depends on how the representation is learned rather than only on the dataset or visualization procedure. The evaluation therefore centers on the induced structure beyond empirical data locations, while reconstruction and downstream performance remain secondary evaluation criteria.

Our contributions are:

- A precise separation between local encoder sensitivity, its scalar Jacobian source, the induced Poisson potential, and the potential's corruption-conditioned directional response;
- A Poisson formulation, $\Delta v = s$, that extends a local Jacobian source into a domain-wide potential and separates boundary flux from score–potential coupling;
- A Gaussian-denoising realization of the score coupling through the normalized reconstruction displacement.

2 Background and related work

Contractive and denoising autoencoders. For an encoder $f_\phi : \mathbb{R}^d \to \mathbb{R}^m$ and decoder g_θ, the classical CAE augments reconstruction with $\mathbb{E}_{p_D} \left\| J_{f_\phi}(x) \right\|_F^2$ (Rifai et al., 2011b). The original interpretation is geometric: robustness of the features corresponds to contraction of input-space neighborhoods after projection into feature space. The principal singular directions of a representation Jacobian also describe locally important manifold directions (Rifai et al., 2011a). Denoising autoencoders replace the explicit derivative penalty with reconstruction from corrupted inputs (Vincent et al., 2008; 2010).

Jacobians, sensitivity, and smoothness. The use of sensitivity in this paper follows the differential interpretation of a Jacobian: it characterizes the first-order response of a network output to nearby input perturbations. Jacobian penalties have accordingly been connected to robustness under small input changes (Srinivas & Fleuret, 2018). Sensitivity is not a unique scalar, however, because directional, average, and worst-case responses correspond to different functions of the Jacobian. More broadly, smoothness constraints depend on the task, data geometry, and selected norms (Rosca et al., 2020).

Score-based generative modeling. Score matching estimates $\nabla_x \log p(x)$ without requiring a normalized likelihood (Hyvärinen, 2005). Denoising score matching makes this estimation tractable after corruption (Vincent, 2011). Alain and Bengio show that a regularized reconstruction function estimates the data score in the small-noise limit (Alain & Bengio, 2014). Their central result concerns the reconstruction function rather than the encoder Jacobian alone, a distinction preserved in our use of the score–potential coupling.

Neural variational PDE solvers. The potential is approximated by a neural variational solver related to the Deep Ritz method (E & Yu, 2018). Because the encoder determines the source and the optimized potential enters the outer objective, the construction is bilevel and uses implicit differentiation (Lorraine et al., 2020).

3 From local encoder response to a non-local potential

3.1 Local encoder sensitivity and the Jacobian source

Let $f_\phi : \Omega \subset \mathbb{R}^d \to \mathbb{R}^m$ be differentiable at x. Although the source is derived from the latent representation $f_\phi(x)$, the potential $v_\phi : \Omega \to \mathbb{R}$ is defined over the ambient input domain. The construction therefore organizes representation-induced geometry as a function of input location and does not directly define a potential or metric on latent coordinates. For a small perturbation δx,

$$f_\phi(x + \delta x) - f_\phi(x) = J_{f_\phi}(x)\delta x + O(\|\delta x\|_2^2). \tag{1}$$

We define the *directional local encoder sensitivity* along a unit vector u as $\left\|J_{f_\phi}(x)u\right\|_2$. The spectral norm $\left\|J_{f_\phi}(x)\right\|_2$ is the worst-case directional sensitivity. The Frobenius norm $\left\|J_{f_\phi}(x)\right\|_F$ is an aggregate local sensitivity magnitude rather than a worst-case response. For $\epsilon \sim \mathcal{N}(0, I)$,

$$\mathbb{E}_\epsilon \left\|J_{f_\phi}(x)\epsilon\right\|_2^2 = \left\|J_{f_\phi}(x)\right\|_F^2, \tag{2}$$

so the squared Frobenius norm equals the expected squared first-order representation change under isotropic unit-variance perturbations.

The classical CAE source is the squared Jacobian magnitude

$$s_\phi^{\mathrm{CAE}}(x) = \left\|J_{f_\phi}(x)\right\|_F^2 = \sum_{j=1}^{m} \|\nabla_x f_{\phi,j}(x)\|_2^2. \tag{3}$$

The current experiments instead use the smoothed Jacobian magnitude

$$s_\phi^{\mathrm{mag}}(x) = \sqrt{\left\|J_{f_\phi}(x)\right\|_F^2 + \varepsilon}, \tag{4}$$

which weights large Jacobians less strongly than the squared CAE source. The reported Poisson fields use s_ϕ^{mag} as their source, while s_ϕ^{CAE} is displayed for reference in the field visualization. Accordingly, the experiments study the Poisson extension of a Jacobian-derived magnitude rather than an exact reformulation of the classical CAE penalty.

3.2 Poisson identity

Let $p : \overline{\Omega} \to (0, \infty)$ be a continuously differentiable population density on a bounded domain with sufficiently regular boundary. We assume that $p\nabla v_\phi$ is integrable and has a well-defined normal trace on $\partial\Omega$. The empirical measure $\widehat{p}_N = N^{-1}\sum_i \delta_{x_i}$ is used for Monte Carlo estimation and is not assumed to possess a classical score.

We introduce v_ϕ satisfying

$$\Delta v_\phi(x) = s_\phi(x), \qquad x \in \Omega, \tag{5}$$

with homogeneous Dirichlet data on the idealized boundary. Applying Green's first identity with the population density p gives

$$\begin{aligned}\mathbb{E}_p[s_\phi(x)] = &\int_{\partial\Omega} p(s)\, \partial_n v_\phi(s)\, \mathrm{d}S \\ &- \mathbb{E}_p\left[\nabla \log p(x)^\top \nabla v_\phi(x)\right].\end{aligned} \tag{6}$$

The left side is the expected scalar Jacobian source. The first term is a density-weighted normal flux, while the second couples the potential gradient to the density score. Equation (6) sets the potential as a non-local relation induced by the source and expressed through two geometric terms. A complete derivation and discussion of the boundary conditions are provided in Appendix B.

3.3 Variational realization and soft anchors

The target Dirichlet energy for the sign convention in Eq. (5) is

$$\mathcal{E}_\phi(v) = \int_\Omega \left[\frac{1}{2} \|\nabla v(x)\|_2^2 + s_\phi(x) v(x) \right] \mathrm{d}x. \tag{7}$$

Its Euler–Lagrange equation is

$$\Delta v_\phi = s_\phi. \tag{8}$$

The implementation uses $\mu = 0$ and approximates the domain integrals with collocation samples. A nonuniform collocation rule requires appropriate quadrature weights; otherwise, the sampled Ritz functional corresponds to a weighted elliptic operator rather than Eq. (8).

Let $q_{\psi,\alpha} = (\Pi_{\psi,\alpha})_\# p$ denote the pushforward distribution induced by the corruption. The numerical objective adds the soft anchor

$$\mathcal{B}_D(v) = \lambda_D \mathbb{E}_{\tilde{x} \sim q_{\psi,\alpha}}[v(\tilde{x})^2], \tag{9}$$

which encourages low potential values at projected off-data locations. We assume that these locations concentrate in a sufficiently thin tubular neighborhood of a surface $\Gamma_{\psi,\alpha}$ surrounding the high-density data region and that their displacement is predominantly outward. Under this assumption, Eq. (9) acts as a weighted soft approximation to homogeneous Dirichlet data on $\Gamma_{\psi,\alpha}$. For a general corruption, the pushforward may instead form a thick or disconnected region, so the anchor distribution should be interpreted as an operational boundary surrogate.

3.4 Corruptions as stochastic measurement operators

Let $\Pi_{\psi,\alpha}$ denote a corruption family ψ at strength α and define

$$\tilde{x} = \Pi_{\psi,\alpha}(x), \qquad n_{\psi,\alpha}(x) = \frac{\tilde{x} - x}{\|\tilde{x} - x\|_2 + \epsilon}. \tag{10}$$

The empirical directional response is

$$F_{\psi,\alpha}(x) = \langle \nabla v_\phi(\Pi_{\psi,\alpha}(x)), n_{\psi,\alpha}(x) \rangle . \tag{11}$$

This quantity belongs jointly to the potential and the distribution of corruption-induced directions. It is not an intrinsic encoder sensitivity, and corrupted points are not assumed to provide an unbiased quadrature rule for a literal boundary.

3.5 Gaussian denoising and the score–potential coupling

Gaussian corruption is special because the corrupted marginal p_σ is smooth. Let $r_{\phi,\theta}(\tilde{x}) = g_\theta(f_\phi(\tilde{x}))$ denote a denoising reconstruction function trained with squared error under $\tilde{x} = x + \sigma\epsilon$ and $\epsilon \sim \mathcal{N}(0, I)$. At the population optimum, $r^*_\sigma(\tilde{x}) = \mathbb{E}[x \mid \tilde{x}]$, and Tweedie's identity gives

$$\nabla_{\tilde{x}} \log p_\sigma(\tilde{x}) = \frac{\mathbb{E}[x \mid \tilde{x}] - \tilde{x}}{\sigma^2}. \tag{12}$$

We therefore use the normalized denoising displacement

$$\widehat{\nabla \log p_\sigma}_\sigma(\tilde{x}) = \frac{r_{\phi,\theta}(\tilde{x}) - \tilde{x}}{\sigma^2} \tag{13}$$

as the score proxy, rather than the gradient of a generic task loss. The corresponding bulk estimate is

$$\widehat{\mathcal{B}}_\sigma = \mathbb{E}_{\tilde{x} \sim p_\sigma} \left[\widehat{\nabla \log p}_\sigma(\tilde{x})^\top \nabla v_\phi(\tilde{x}) \right]. \tag{14}$$

This alignment assumes that the learned denoiser adequately approximates the posterior mean, so the bulk estimate contains denoising-estimation error in addition to Monte Carlo and PDE-solver error. Deriving corruption-specific bulk terms for non-Gaussian corruptions is left to later work.

4 EXPERIMENTAL DESIGN

The field-visualization experiments use a two-dimensional three-component Gaussian mixture so that the source, potential, gradient, and directional response can be inspected in the ambient coordinates. Unless otherwise stated, the potential is a three-layer SiLU multilayer perceptron of width 128, trained for $K = 100$ inner steps with $\mu = 0$, $\lambda_D = 1$, and implicit bilevel differentiation.

RQ1 uses s_ϕ^{mag} as the Poisson source and compares it with s_ϕ^{CAE}, v_ϕ, ∇v_ϕ, and directional-response maps. RQ2 applies the same field construction to plain, denoising, contractive, and Poisson-regularized autoencoders. Because raw latent coordinates are non-identifiable under rotations and rescalings, RQ2 emphasizes field topology and treats magnitude differences as descriptive. The Frobenius source is invariant to orthogonal latent rotations but not to arbitrary latent rescaling. RQ3 holds the learned representation and potential fixed while varying only the test-time corruption family and strength. This isolates the geometry of the measurement operator from the representation-training objective.

4.1 PERTURBATION REGIMES

For RQ3, the strength parameters have different meanings across corruption families, so we use family-specific symbols rather than a common parameter α. Gaussian corruption is defined by

$$\widetilde{x} = x + \sigma\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{15}$$

with

$$\sigma \in \{0.01, 0.03, 0.05, 0.10, 0.20, 0.50\}.$$

The diffusion-style corruption is

$$\widetilde{x}_t = \sqrt{\bar{\alpha}_t}\, x + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{16}$$

where $\bar{\alpha}_t = \prod_{k=0}^{t}(1 - \beta_k)$. The implementation uses $T = 200$, a linear schedule from $\beta_0 = 10^{-4}$ to $\beta_{T-1} = 2 \times 10^{-2}$, and evaluates

$$t \in \{1, 3, 5, 10, 25, 50, 100\}.$$

Coordinate deletion is defined by

$$\widetilde{x} = m \odot x, \qquad m_j \sim \mathrm{Bernoulli}(1 - \rho), \tag{17}$$

with

$$\rho \in \{0.10, 0.20, 0.30, 0.50, 0.70\}.$$

The implementation labels this last operator as dropout and applies no inverse-probability rescaling. The separately configured masking operator uses the same transformation; it is therefore omitted from the comparison rather than treated as a distinct corruption family. The parameters σ, t, and ρ define the evaluated perturbation regimes but are not directly comparable across families.

Most corruption panels use sampled corruption realizations and independent color scales. They support claims about spatial organization more strongly than monotonic magnitude claims.

5 RESULTS

5.1 RQ1: THE POTENTIAL EXPOSES NON-LOCAL ORGANIZATION

Figure 1 compares the aggregate local sensitivity source, its square, the induced potential, the potential-gradient field, and two corruption-conditioned directional responses under Gaussian corruption. The source varies locally around and between the observed modes, whereas the potential forms a smooth domain-wide structure. The potential therefore does not reproduce the source pointwise; it reflects a Green-kernel-weighted transformation of the complete source distribution together with the boundary realization. The gradient adds directional information that neither the scalar source nor the scalar potential provides. The response panels are less smooth because they also contain sampled corruption-induced directions.

The shift–scale field visualization in Appendix C produces a sharper diagonal Jacobian source and a correspondingly different non-local potential. Because its color scale differs from Fig. 1, the comparison establishes spatial reorganization rather than a calibrated magnitude ranking. Together, the maps provide a positive qualitative answer to RQ1: the potential contains non-local organization not visible from the pointwise source alone.

5.2 RQ2: THE INDUCED FIELD DEPENDS ON THE LEARNING OBJECTIVE

Figure 3 compares the four autoencoder objectives on the same data geometry. The classical CAE yields the smallest Jacobian and potential-gradient scales, consistent with direct suppression of first-order representation changes under input perturbations. The plain and denoising models produce stronger smooth variations, while the Poisson-regularized model exhibits a more heterogeneous field with corridor-like directional organization. The field is therefore not solely a visualization of the dataset locations.

This ablation is a supporting control rather than a performance ranking. The current evidence shows distinguishable fields under a common diagnostic, but invariant comparison requires latent whitening, alignment, or a normalized pullback quantity.

5.3 RQ3: QUALITATIVE RESPONSE ACROSS CORRUPTION REGIMES

Figures 4, 5, and 6 show the directional response obtained under Gaussian, diffusion-style, and coordinate-deletion corruptions at different corruption strengths. These results are interpreted qualitatively because the panels correspond to a single experimental seed and use independently normalized color scales.

The Gaussian panels exhibit spatially diffuse responses whose organization changes gradually across corruption strengths. The diffusion-style panels display a visually similar progression, particularly at low and intermediate corruption levels. In both cases, the response is distributed throughout the domain rather than separated by sharp coordinate-aligned boundaries. The present visualizations indicate a qualitative resemblance between these two corruption families, but do not establish equivalence between their responses.

Coordinate deletion produces a different spatial pattern. Its response maps contain pronounced horizontal and vertical transitions that divide the domain into approximately axis-aligned regions. These partitions become especially visible at intermediate and high corruption strengths. The corresponding field evaluations also vary across the corruption regimes, although their magnitudes cannot be compared directly because each panel uses a separate color scale.

Table 1: Qualitative patterns observed across the evaluated corruption regimes. The descriptions summarize visual evidence and are not quantitative comparisons.

Corruption family	Observed spatial pattern	Provisional interpretation
Gaussian	Diffuse, spatially distributed	Non-coordinate-selective response
Diffusion-style	Diffuse progression	Visually similar to Gaussian regimes
Coordinate deletion	Axis-aligned partitions	Coordinate-selective response

Taken together, the visualizations suggest that the directional response in Eq. (11) depends on the corruption family: Gaussian and diffusion-style corruptions are associated with diffuse domain-wide patterns, whereas coordinate deletion is associated with coordinate-selective partitions. This observation constitutes provisional visual evidence for RQ3. Quantitative comparisons across repeated corruption draws, seeds, and matched distortion levels are required before making a stronger claim about operator-conditioned response.

6 DISCUSSION

The Jacobian $J_{f_\phi}(x)$ characterizes the local first-order response of the representation to changes in the input. Its Frobenius norm aggregates this response into a scalar source, while the squared norm

admits the isotropic perturbation interpretation given in Eq. (2). Solving the Poisson equation distributes the source over the domain according to the differential operator and the imposed boundary realization. Importantly, the resulting potential defines a distinct scalar field over the input domain rather than a measure of sensitivity itself. It provides a geometric description of how locally defined encoder responses are organized across observed samples, between samples, and in regions of low data density.

The corruption operator determines both the locations at which the potential is softly anchored and the directions along which its variation is evaluated. Changing the corruption therefore changes the pushforward distribution, the displacement geometry, and the resulting directional response. A large response is not intrinsically beneficial or harmful; it indicates that the selected corruption-induced direction crosses a region of substantial variation in the potential. The response must consequently be interpreted relative to the corruption regime used to construct it.

The pushforward of the data distribution under corruption does not necessarily coincide with a geometric boundary of the data support. The boundary interpretation is most direct when corrupted samples concentrate near a thin surface outside the high-density data region and their displacements are predominantly outward. Under these conditions, the anchor distribution approximates a codimension-one boundary on which low potential values are encouraged. When corrupted samples instead occupy a thick or disconnected region, the same term remains operationally well defined, but it should be interpreted as imposing low potential over an off-data anchor region rather than as approximating a classical boundary condition.

Because the empirical data distribution consists of a finite collection of point masses, its score is not defined as an ordinary differentiable function. Gaussian corruption smooths this distribution and produces a density with a well-defined score. Moreover, the appropriately scaled conditional denoising displacement provides an estimator of this score, giving the Gaussian bulk term a direct score-based interpretation. For non-Gaussian corruptions, the induced displacement generally represents a transport direction rather than the score of a smooth density. We therefore interpret the corresponding quantity as a directional coupling with the potential gradient.

The encoder determines the source through its Jacobian, while the Poisson operator extends the resulting scalar quantity across locations $x \in \Omega$. The construction therefore produces a representation-induced field over the ambient input domain rather than a differential equation posed directly in latent coordinates. In this sense, the proposed structure is determined by the learned representation but evaluated across the domain on which the encoder acts.

7 Future work

The next phase will quantitatively characterize the spatial organization observed in the field visualizations. This includes measuring the propagation of local source variations across the domain, averaging directional responses over repeated corruption realizations, and comparing corruption families at matched expected distortion. Gaussian and diffusion-style regimes will also be aligned through their effective noise variance to support a more controlled comparison of their multiscale behavior. Complementary numerical analyses will examine the Poisson residual, anchor error, Green-identity closure error, and stability with respect to the collocation and inner-optimization configurations.

A broader theoretical direction is the study of the properties of the non-locally coupled potential. Of particular interest are its regularity, spatial decay, stability under perturbations of the encoder, and dependence on the source, differential operator, and boundary realization. Future work will also compare the magnitude source $\|J_f\|_F$ with the squared CAE source $\|J_f\|_F^2$ and investigate corruption-specific score and transport couplings. These analyses may clarify which geometric properties arise from the learned representation and which are introduced by the Poisson construction.

Another direction concerns the behavior of the potential in regions far from the training distribution. Because the Poisson operator extends representation-induced structure into low-density and off-sample regions, its level sets and gradient field may provide information about how a generative model organizes such regions. Future studies will examine whether this geometry can characterize transitions away from the data distribution and whether changes in the potential field are informative of out-of-distribution samples or distribution shift. This interpretation will be investigated without assuming that the potential is itself a probability density or an out-of-distribution score.

8 CONCLUSION

We presented a Poisson construction, $\Delta v = s$, for extending representation-induced structure beyond observed samples. The construction uses an aggregate local encoder-sensitivity measure as a source and produces a scalar field over the ambient domain whose directional variation can be queried by controlled corruption operators. Gaussian denoising aligns the score–potential coupling with the normalized reconstruction displacement, while other corruption families define geometric measurement operators. Current field visualizations indicate that the potential organizes sample-local encoder information across inter-sample and low-density regions and that different representation-learning objectives yield distinguishable induced fields. They also provide provisional evidence that the empirical directional response varies with the corruption family. These findings support Poisson fields as a mathematical mechanism for structuring off-sample representation geometry.

REFERENCES

Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution. *Journal of Machine Learning Research*, 15(110):3743–3773, 2014.

Weinan E and Bing Yu. The deep ritz method: A deep learning-based numerical algorithm for solving variational problems. *Communications in Mathematics and Statistics*, 6(1):1–12, 2018.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

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

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 1540–1552, 2020.

Salah Rifai, Grégoire Mesnil, Pascal Vincent, Xavier Muller, Yoshua Bengio, Yann Dauphin, and Xavier Glorot. Higher order contractive auto-encoder. In *Machine Learning and Knowledge Discovery in Databases*, pp. 645–660. Springer, 2011a.

Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 833–840, 2011b.

Mihaela Rosca, Theophane Weber, Arthur Gretton, and Shakir Mohamed. A case for new neural network smoothness constraints. In *Proceedings of the NeurIPS Workshop "I Can't Believe It's Not Better!"*, pp. 21–32. PMLR, 2020.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2256–2265, 2015.

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

Suraj Srinivas and François Fleuret. Knowledge transfer with jacobian matching. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 4723–4731. PMLR, 2018.

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

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096–1103, 2008.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408, 2010.

A Reproducibility details

The principal variational runs use $K = 100$, $\mu = 0$, $\lambda_D = 1$, and outer weight $\lambda = 10^{-2}$. The encoder and potential multilayer perceptrons use SiLU activations and hidden width 128 unless otherwise specified. The complete corruption strengths and dataset-specific training lengths remain stored in the archived experiment configurations. No uncertainty claim is made from a single-seed corruption comparison.

B Complete derivation of the Green decomposition

Let p satisfy the regularity and positivity assumptions stated in Section 3.2, and let v_ϕ solve Eq. (5) under boundary conditions for which Green's first identity applies. Starting from the expected source and substituting $s_\phi = \Delta v_\phi$ gives

$$\mathbb{E}_p[s_\phi(x)] = \int_\Omega p(x) s_\phi(x)\,\mathrm{d}x = \int_\Omega p(x)\Delta v_\phi(x)\,\mathrm{d}x. \tag{18}$$

Green's first identity with $u = p$ and $v = v_\phi$ yields

$$\int_\Omega p\,\Delta v_\phi\,\mathrm{d}x = \int_{\partial\Omega} p\,\partial_n v_\phi\,\mathrm{d}S - \int_\Omega \nabla p^\top \nabla v_\phi\,\mathrm{d}x. \tag{19}$$

Because $p > 0$, the identity $\nabla p = p\nabla \log p$ gives

$$\int_\Omega \nabla p^\top \nabla v_\phi\,\mathrm{d}x = \mathbb{E}_p\left[\nabla \log p(x)^\top \nabla v_\phi(x)\right]. \tag{20}$$

Substituting Eq. (20) into Eq. (19) gives the decomposition in Eq. (6).

Boundary conditions. For homogeneous Dirichlet conditions, $v_\phi|_{\partial\Omega} = 0$, the normal derivative $\partial_n v_\phi$ is generally nonzero, so the boundary-flux term remains. For homogeneous Neumann conditions, $\partial_n v_\phi|_{\partial\Omega} = 0$, integrating Eq. (5) over Ω requires

$$\int_\Omega s_\phi(x)\,\mathrm{d}x = 0. \tag{21}$$

This compatibility condition is inconsistent with a nonnegative, nontrivial Jacobian source. It therefore motivates the homogeneous Dirichlet model and its numerical approximation through corruption-induced soft anchors.

C FIELD VISUALIZATIONS

The following figures collect the complete qualitative field visualizations referenced in the main text.

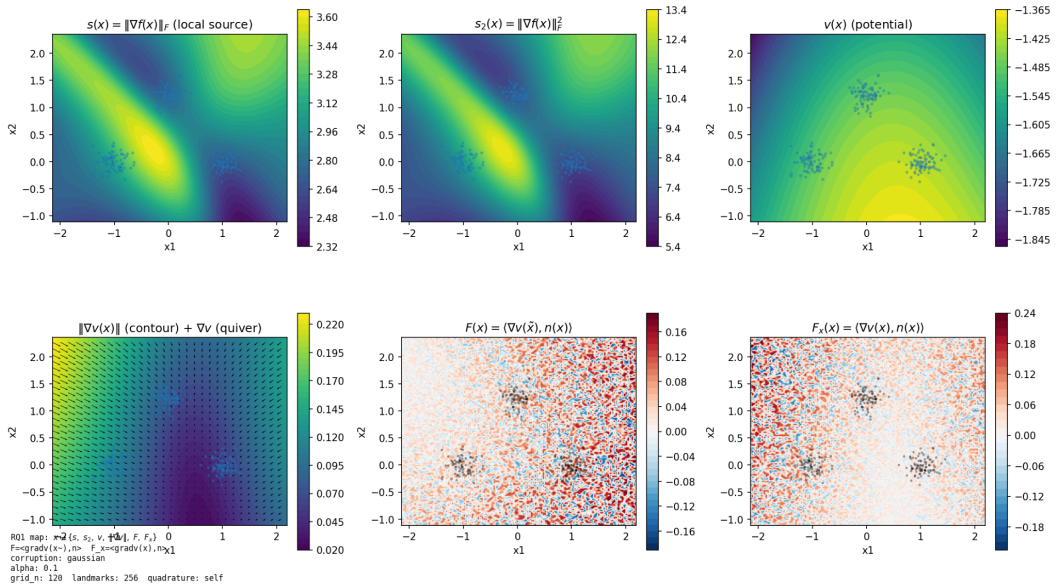


Figure 1: RQ1 field visualization under Gaussian corruption. The first row shows the aggregate local encoder-sensitivity magnitude, its square, and the induced potential. The second row shows the potential-gradient field and corruption-conditioned directional responses.

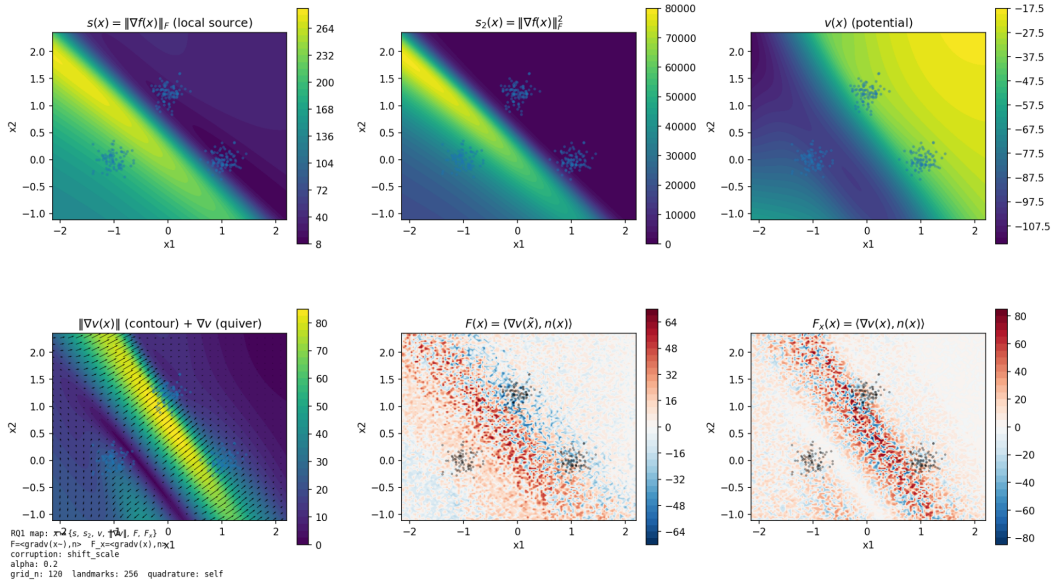
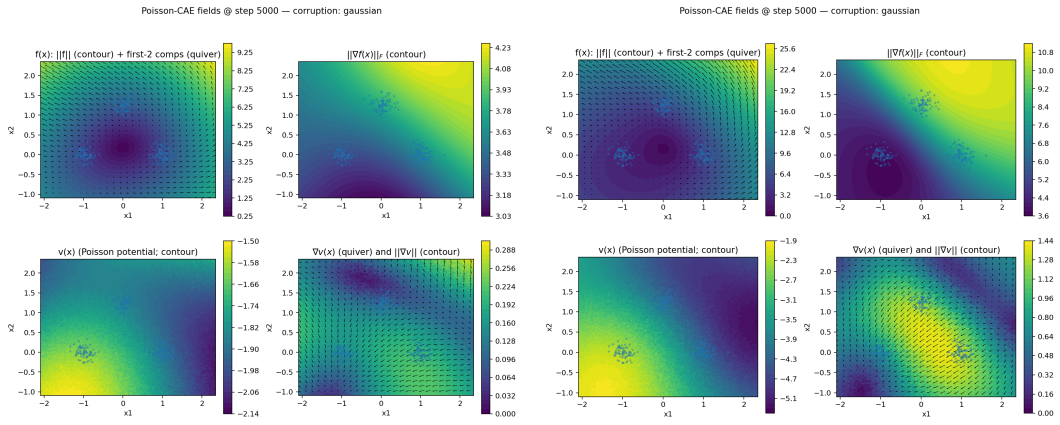
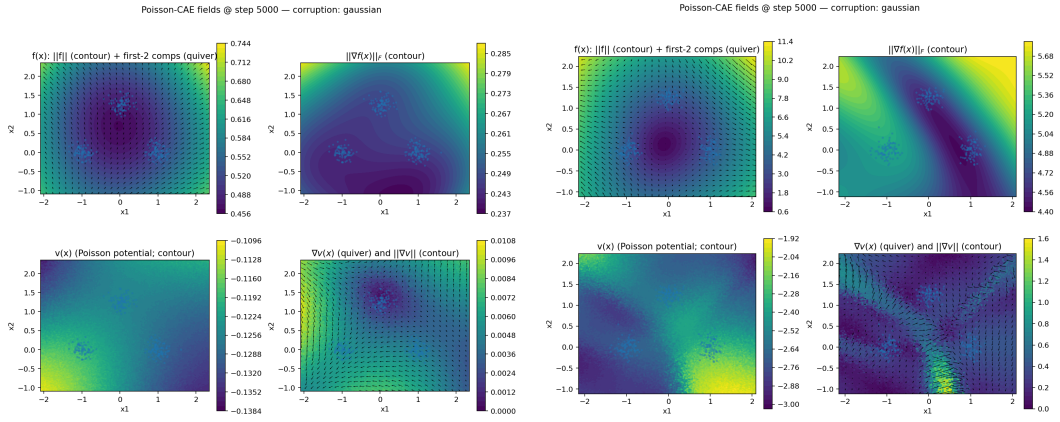


Figure 2: RQ1 field visualization under shift–scale corruption. The local Jacobian source forms a sharp diagonal band, while the induced potential is smooth and domain-wide.



(a) Plain AE

(b) Denoising AE



(c) Contractive AE

(d) Poisson-regularized AE

Figure 3: RQ2 objective ablation. The original plot headers share a legacy title; the subcaptions identify the actual model associated with each file. Color scales are model-specific, so topology is more directly comparable than magnitude.

D RQ3 RESULTS ACROSS CORRUPTION REGIMES

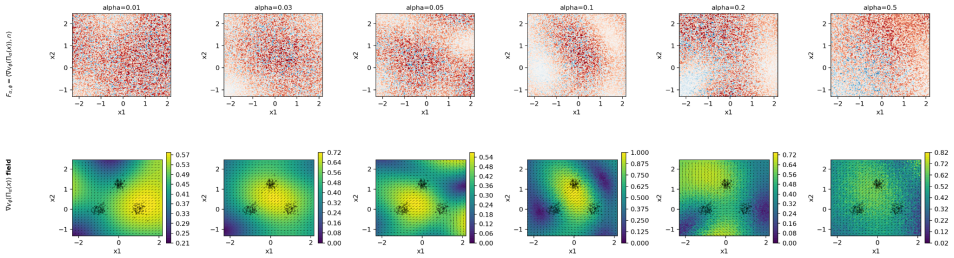


Figure 4: Gaussian corruption regimes.

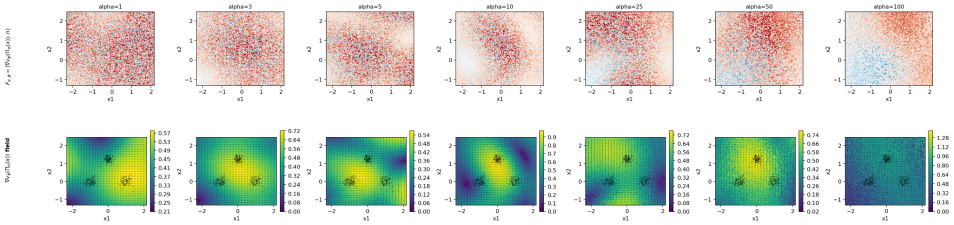


Figure 5: DDPM forward-process regimes.

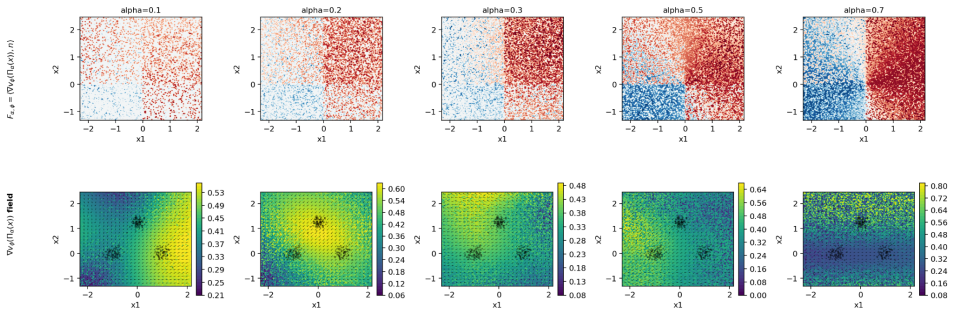


Figure 6: Coordinate-deletion regimes, implemented through the dropout configuration without inverse-probability rescaling.