ROBUST ENSEMBLE GUIDANCE FOR SCIENTIFIC INVERSE PROBLEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Ensemble guidance combines pretrained diffusion priors with black-box forward models to solve inverse problems without differentiating through the physical simulator. However, observation coordinates with large predictive spread or extreme residuals can dominate the ensemble correction, degrading reconstruction accuracy. We show that two simple modifications, weighting and clipping, substantially improve this correction. Our method, Robust Ensemble Guidance (REG), uses ensemble predictive spread to balance observation scales and adaptively clips standardized residuals to limit the influence of extreme discrepancies. Both operations reuse existing particles and forward predictions, requiring no additional denoiser or forward-model evaluations. Under a local linear Gaussian model, we derive conditions for reduced one-step estimation risk, bound the influence of individual observation coordinates, and characterize when these benefits persist with finite ensembles. Experiments on Navier–Stokes inversion, black-hole imaging, and acoustic full-waveform inversion demonstrate improved reconstruction over the underlying ensemble solver. In particular, REG **increases black-hole reconstruction PSNR by 6.2–8.2 dB** across three observation regimes and **reduces Navier–Stokes reconstruction error by 26.4%** in a matched-budget comparison. These findings highlight the importance of observation heterogeneity and residual influence in designing reliable generative solvers for scientific inverse problems.

1 INTRODUCTION

Inverse problems seek an unknown state $x^\star$ from an indirect, incomplete, and noisy observation y (Tarantola, 2005). Since many states may explain the same measurement, recovery requires a strong prior (Stuart, 2010; Bora et al., 2017). Pretrained diffusion models (Ho et al., 2020; Song et al., 2020; Karras et al., 2022) provide reusable priors for restoration, sensing, and scientific inference (Song et al., 2021; Jalal et al., 2021). The prior keeps reconstructions near the data distribution; observation correction promotes measurement consistency (Kadkhodaie & Simoncelli, 2021).

Most observation corrections use a data-fidelity gradient. They propagate the measurement error through a denoised prediction and the forward model (Chung et al., 2022a;b). Other methods use pseudoinverses (Song et al., 2023a; Wang et al., 2022) or known operator structure (Kawar et al., 2022; Choi et al., 2021). These forms of access are natural for many imaging operators, but they are not always available. A PDE adjoint may be expensive to derive and store, and a discrete or long-running simulator may provide unstable sensitivities. Such cases require guidance from forward evaluations alone.

Ensemble guidance provides this capability by passing a set of generative particles through the forward model (Zheng et al., 2024). Empirical correlations between particle states and predicted observations then transport the measurement error back to the state space (Evensen, 2003; Iglesias et al., 2013; Schillings & Stuart, 2017). This approach removes the need for a forward derivative, but it introduces a different problem. The correction itself is estimated from finitely many particles, and its behavior depends on how observation coordinates and residual magnitudes enter the update.

Two effects can make this correction unstable. First, predicted observations can vary on very different scales across coordinates. Without standardization, a large-scale coordinate can dominate the interaction between a residual and the ensemble predictions, amplifying any noise or estimation error in that coordinate. A large spread need not imply weak information: the issue is how unequal scales affect the combined update. Second, a quadratic data-fidelity loss gives increasing influence

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.

to extreme residuals (Huber & Ronchetti, 2011; Maronna et al., 2019). A few corrupted or poorly predicted measurements can therefore distort an otherwise useful direction. A smaller step limits this error's size but not its direction. These effects motivate balancing coordinate scales and limiting extreme residuals before particle correction. We introduce Robust Ensemble Guidance (REG), a derivative-free correction designed for this finite-ensemble setting. REG uses the predictive spread of the ensemble to standardize both the residuals and the predicted-observation variations. This produces a diagonal observation geometry that reduces the influence of high-spread coordinates. REG then clips the standardized residuals at a shared, quantile-adaptive threshold, limiting the influence of extreme discrepancies on the particle interaction. Both operations reuse the particles and forward predictions already needed for the correction. They add no denoiser call and no forward-model evaluation. Uniform coordinate weights without clipping recover the standard quadratic ensemble correction.

Our analysis explains when these modifications help. Under a local linear Gaussian model, balancing predictive scales can improve conditioning and reduce the error component constrained by the observations. We derive conditions for lower one-step estimation risk and characterize when this advantage persists under finite-ensemble estimation of the weights. Our clipping analysis further establishes bounds on residual influence and identifies conditions under which suppressing extreme residuals reduces estimation error. Controlled experiments validate these mechanisms, while experiments on Navier–Stokes inversion, black-hole imaging, and acoustic full-waveform inversion demonstrate improved reconstruction over other methods. These results establish REG's effectiveness as a derivative-free solver across diverse scientific inverse problems and observation regimes.

2 Related Work

Diffusion priors for inverse problems. Diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020; Song et al., 2020; Karras et al., 2022) provide expressive priors that can be reused across observations. Guidance methods combine this prior with an observation likelihood, building on denoiser-based regularization (Romano et al., 2017) and plug-and-play reconstruction (Venkatakrishnan et al., 2013). A common strategy forms a clean prediction at each diffusion step and differentiates its measurement mismatch through both the denoiser and a known forward operator (Chung et al., 2022a;b). Other formulations use variational optimization (Mardani et al., 2024) or decouple prior updates from data consistency (Zhu et al., 2023; Li et al., 2024). Known operator structure (Kawar et al., 2022) and pseudoinverses (Song et al., 2023a; Wang et al., 2022) also enable observation conditioning. REG addresses the complementary derivative-free setting, using generative particles and forward values.

Ensemble Kalman inversion. Ensemble Kalman methods infer state–observation relations from particles (Iglesias et al., 2013; Kovachki & Stuart, 2019). Their theory addresses convergence (Schillings & Stuart, 2017; Blömker et al., 2019) and mean-field limits (Ding & Li, 2021). Iterative regularization (Iglesias, 2016) and repeated data assimilation (Emerick & Reynolds, 2013) extend this framework. General-likelihood extensions likewise modify the score transported through the empirical cross-covariance (Duffield & Singh, 2022). EnKG brings this statistical transport to diffusion guidance using an identity observation geometry and a quadratic residual (Zheng et al., 2024). Sample covariance concentration quantifies finite-sample error (Koltchinskii & Lounici, 2017). We study finite-ensemble observation scaling and residual influence at fixed coordinates.

Robust estimation. Huber losses are quadratic near zero (Huber, 1992); their bounded scores limit residual influence (Hampel, 1974). Related robust mean (Catoni, 2012) and regression estimators (Sun et al., 2020) control heavy-tailed fluctuations. Unlike oracle noise weights, inverse spread is observable but fallible. Our analysis identifies when the geometry and bounded score do or do not improve the conventional identity-quadratic update.

3 Robust Ensemble Guidance

Robust Ensemble Guidance (REG) combines diffusion-prior prediction with an ensemble correction. Paired state and observation variations transport data discrepancy to the unknown state. Predictive-spread weighting balances local observation scales, while quantile-adaptive clipping limits extreme residuals. We derive these components and conditions for lower one-step risk.

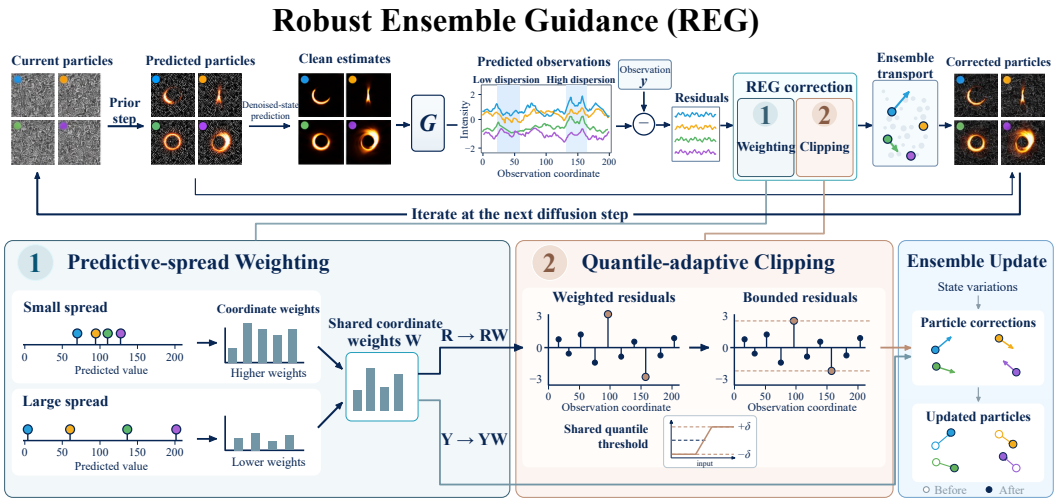


Figure 1: Overview of Robust Ensemble Guidance. Forward predictions from a denoised particle ensemble determine coordinate weights and a shared clipping threshold. The resulting bounded residuals are transported back to the particle space to form the next correction.

3.1 Ensemble guidance from forward predictions

Let $G : \mathbb{R}^n \to \mathbb{R}^m$ denote the forward model, where n and m are the state and observation dimensions. The unknown state $x^\star \in \mathbb{R}^n$ produces the observation $y \in \mathbb{R}^m$ according to:

$$y = G(x^\star) + \xi, \qquad \Gamma = \mathrm{Cov}(\xi), \tag{1}$$

where ξ is zero-mean observation noise and $\Gamma \in \mathbb{R}^{m\times m}$ is its covariance. The decreasing diffusion-time schedule $\tau_T > \cdots > \tau_0 = 0$ comprises T reverse steps. At each step $t \in \{T, \dots, 1\}$, REG maintains an ensemble of J particles $\{x_t^{(j)}\}_{j=1}^J$ indexed by j. For example, one Euler step of the probability-flow dynamics gives the diffusion prediction:

$$x_{t-1}^{\mathrm{pred},(j)} = x_t^{(j)} - \dot\sigma(\tau_t)\sigma(\tau_t)s_\theta(x_t^{(j)}, \tau_t)(\tau_{t-1} - \tau_t), \tag{2}$$

Here $\sigma(\tau)$ is the diffusion noise schedule with time derivative $\dot\sigma$. The score model s_θ has pretrained parameters θ. The superscript pred distinguishes the prediction from the corrected particle $x_{t-1}^{(j)}$ below. The correction does not depend on the numerical solver.

The denoising map ϕ_t operates at the predicted particle's noise level. Correction quantities are indexed by reverse step t. The clean-state estimate $\widehat{x}_t^{(j)}$, predicted observation $\widehat{y}_t^{(j)}$, and residual $r_t^{(j)}$ are:

$$\widehat{x}_t^{(j)} = \phi_t(x_{t-1}^{\mathrm{pred},(j)}), \qquad \widehat{y}_t^{(j)} = G(\widehat{x}_t^{(j)}), \qquad r_t^{(j)} = y - \widehat{y}_t^{(j)}. \tag{3}$$

$\bar{x}_{t-1}^{\mathrm{pred}}$ and $\bar{y}_t$ denote the predicted-particle and observation means. The state-anomaly matrix $E_t \in \mathbb{R}^{J\times n}$, observation-anomaly matrix $Y_t \in \mathbb{R}^{J\times m}$, and residual matrix $R_t \in \mathbb{R}^{J\times m}$ have rows:

$$(E_t)_{j,:} = (x_{t-1}^{\mathrm{pred},(j)} - \bar{x}_{t-1}^{\mathrm{pred}})^\top, \qquad (Y_t)_{j,:} = (\widehat{y}_t^{(j)} - \bar{y}_t)^\top, \qquad (R_t)_{j,:} = (r_t^{(j)})^\top. \tag{4}$$

The matrix $X_{t-1}^{\mathrm{pred}} \in \mathbb{R}^{J\times n}$ stacks the predicted particles by row. The basic coupling matrix $\mathcal{A}_t^{(0)} \in \mathbb{R}^{J\times J}$ and the corrected particle matrix $X_{t-1} \in \mathbb{R}^{J\times n}$ are:

$$\mathcal{A}_t^{(0)} = \frac{1}{J} R_t Y_t^\top, \qquad X_{t-1} = X_{t-1}^{\mathrm{pred}} + \gamma_t \mathcal{A}_t^{(0)} E_t, \tag{5}$$

where $\gamma_t > 0$ is the correction step size. The entry $(\mathcal{A}_t^{(0)})_{j\ell}$ measures the alignment between particle j's residual and particle ℓ's observation anomaly. The corresponding state anomaly of particle ℓ transfers this information to the state space. The empirical state covariance C_t^{xx} and state–observation cross-covariance C_t^{xy} are computed from the ensemble as:

$$C_t^{xx} = \frac{1}{J} E_t^\top E_t, \qquad C_t^{xy} = \frac{1}{J} E_t^\top Y_t. \tag{6}$$

For $G \circ \phi_t$ locally linear over the ensemble, its linearization $H_t = \nabla(G \circ \phi_t)(\bar{x}_{t-1}^{\mathrm{pred}}) \in \mathbb{R}^{m \times n}$ gives:

$$C_t^{xy} \approx C_t^{xx} H_t^\top, \tag{7}$$

which supplies a covariance-preconditioned local direction without forming H_t. Weighting and clipping control how observation coordinates and residual magnitudes enter this direction.

The analysis considers a single local correction with fixed linearization and covariance. A fixed step size $\alpha > 0$ serves as the theoretical counterpart of γ_t. The local reference state x_t^{ref} is in the particle space and satisfies $G(\phi_t(x_t^{\mathrm{ref}})) = G(x^\star)$. This idealization isolates the observation correction from denoising and linearization errors. Appendix A gives assumptions, proofs, and additional analysis.

3.2 Reliability-weighted observation geometry

Identity observation weighting assigns the same coefficient to every coordinate, even when their predictive scales differ substantially. The resulting update can amplify a large-scale coordinate together with its perturbations. A two-coordinate example makes the scale instability explicit. Consider a scalar local state with unit ensemble variance and linearized observation vector $h = (1, L)^\top$, where $L > 0$ is the scale ratio between the two coordinates. The two predictive standard deviations are then proportional to 1 and L. Here e denotes the scalar mean-state error and ζ represents a perturbation in the second residual. The identity-weighted correction g_I is:

$$g_I = (1 + L^2)e + L\zeta. \tag{8}$$

The corrected error is $e^+ = e - \alpha g_I$. The iteration is stable only when $0 < \alpha < 2/(1 + L^2)$. The perturbation enters with gain L. Thus a large spread ratio simultaneously narrows the stable step-size range and amplifies finite-ensemble or measurement perturbations.

For observation coordinate $k \in \{1, \ldots, m\}$, empirical standard deviation $v_{t,k}$ defines raw weight $\widetilde{w}_{t,k}$ and unit-mean normalized weight $w_{t,k}$:

$$v_{t,k} = \mathrm{Std}_j[\widehat{y}_{t,k}^{(j)}], \qquad \widetilde{w}_{t,k} = \frac{1}{v_{t,k} + \varepsilon_w}, \qquad w_{t,k} = \frac{\widetilde{w}_{t,k}}{m^{-1} \sum_{k'=1}^{m} \widetilde{w}_{t,k'}}. \tag{9}$$

Here Std_j denotes the standard deviation across particles and k' indexes the sum. The positive floor ε_w prevents unbounded raw weights. Unit-mean normalization separates relative coordinate importance from the overall step size. The diagonal weight matrix $W_t \in \mathbb{R}^{m \times m}$ and the weighted residual and observation-anomaly matrices $\widetilde{R}_t, \widetilde{Y}_t \in \mathbb{R}^{J \times m}$ are defined by:

$$W_t = \mathrm{diag}(w_{t,1}, \ldots, w_{t,m}), \qquad \widetilde{R}_t = R_t W_t, \qquad \widetilde{Y}_t = Y_t W_t. \tag{10}$$

The weighted coupling $\widetilde{R}_t \widetilde{Y}_t^\top / J$ uses identical weights for residuals and observation anomalies.

Why weighting helps. At a fixed reverse step, $C = C_t^{xx}$ and $H = H_t$ denote the fixed state covariance and linearization. The prediction covariance is $\Sigma_y = HCH^\top$. Applying the weight rule to the population spreads $v_k = \sqrt{(\Sigma_y)_{kk}}$ yields the population geometry $Q \in \mathbb{R}^{m \times m}$. Its empirical counterpart is $\widehat{Q} = W_t^2$. The diagonal variance matrix is $D = \mathrm{diag}(v_1^2, \ldots, v_m^2)$. For positive spreads in the zero-floor limit, the normalization factor $c = (m^{-1} \sum_k v_k^{-1})^{-1}$ gives:

$$Q = c^2 D^{-1}. \tag{11}$$

Thus weighting diagonally whitens observation space and removes arbitrary coordinate scales in this limit. With a nonzero floor, $Q_{kk} \propto (v_k + \varepsilon_w)^{-2}$. For correlated observations, the magnitude of the standardized off-diagonal correlations is measured by $\rho = \|D^{-1/2}(\Sigma_y - D)D^{-1/2}\|_{\mathrm{op}} < 1$. The condition number κ of the standardized covariance then satisfies:

$$\kappa\Big(D^{-1/2} \Sigma_y D^{-1/2}\Big) \le \frac{1 + \rho}{1 - \rho}. \tag{12}$$

Hence weighting improves conditioning when marginal scale imbalance dominates the standardized correlations. In the noiseless diagonal case, the update contracts covariance-whitened observable error for $0 < \alpha c^2 < 2$. The unobservable component remains unchanged.

Predictive spread is a local reliability proxy, so its benefit under noise is conditional. Let $e = \bar{x}_{t-1}^{\mathrm{pred}} - x_t^{\mathrm{ref}}$ denote the local mean-state error and e^+ its value after correction. With C, H, Q, α fixed, the one-step risk $\mathcal{R}(Q)$ averages squared error over observation noise:

$$\mathcal{R}(Q) := \mathbb{E}\|e^+\|_2^2 = \|(I_n - \alpha CH^\top QH)e\|_2^2 + \alpha^2 \,\mathrm{tr}\left(CH^\top Q \Gamma Q H C\right), \tag{13}$$

The two terms are the remaining state error and transferred observation noise. Define the local direction $d_Q = CH^\top QHe$ and transferred noise level $\nu_Q = \mathrm{tr}(CH^\top Q\Gamma QHC)$. Setting $Q = I_m$ gives the identity-weighted counterparts d_I and ν_I. Risk expansion gives the exact criterion:

$$2e^\top(d_Q - d_I) > \alpha\left(\|d_Q\|_2^2 - \|d_I\|_2^2 + \nu_Q - \nu_I\right). \tag{14}$$

The left side is the additional alignment with the state error. The right side is the added magnitude and noise cost. A positive alignment margin therefore guarantees improvement for a sufficiently small step. For the empirical geometry $\widehat{Q} = W_t^2$, define the population risk advantage $\Delta_{\mathcal{R}} = \mathcal{R}(I_m) - \mathcal{R}(Q) > 0$. This advantage persists whenever $\|\widehat{Q} - Q\|_{\mathrm{op}} < \Delta_{\mathcal{R}}/L_{\mathcal{R}}$, where $L_{\mathcal{R}}$ is the local sensitivity of risk to Q. Under sub-Gaussian predictions and nonvanishing spreads, the error is $O(\sqrt{\log(m/\eta)/J})$ with probability at least $1-\eta$, where $\eta \in (0,1)$ is the failure probability.

3.3 Quantile-adaptive bounded residuals

Weighting does not bound residual influence. For a fixed quantile level $q \in (0,1)$, REG computes the threshold δ_t from all weighted residuals. The elementwise clipping function ψ_{δ_t} for $a \in \mathbb{R}$ is:

$$\delta_t = \mathrm{Quantile}_q\left(\{|(\widetilde{R}_t)_{jk}|\}_{j,k}\right), \qquad \psi_{\delta_t}(a) = \mathrm{sign}(a)\min(|a|, \delta_t). \tag{15}$$

Clipping preserves signs and small entries, caps magnitudes at δ_t, and yields:

$$\mathcal{A}_t = \frac{1}{J}\psi_{\delta_t}(\widetilde{R}_t)\widetilde{Y}_t^\top, \tag{16}$$

where $\mathcal{A}_t \in \mathbb{R}^{J\times J}$ couples residuals to particle anomalies. A quantile adapts the threshold to each step's residual scale. Upper-tail contamination below $1-q$ cannot make it unbounded.

Why clipping helps. For a fixed threshold $\delta \geq 0$ and scalar residual a, clipping satisfies:

$$a\,\psi_\delta(a) \geq \psi_\delta(a)^2 \geq 0. \tag{17}$$

Clipping preserves residual signs while shrinking tails. With $W_t, \widetilde{Y}_t, \delta_t$ fixed, changing the weighted residual entry (j,k) induces a coupling change $\Delta\mathcal{A}_t$. Its Frobenius norm obeys:

$$\|\Delta\mathcal{A}_t\|_F \leq \frac{2\delta_t}{J}\|(\widetilde{Y}_t)_{:,k}\|_2, \tag{18}$$

Unlike step normalization, this bound limits an entry's influence, not just update length.

Clipping trades estimation variance for tail bias. Consider a random weighted residual $u \in \mathbb{R}^m$ with clean distribution P. A fixed local map $M \in \mathbb{R}^{n\times m}$ transfers observation scores to the state space. For J independent samples with fixed M, δ, clipping reduces sample-mean MSE if and only if:

$$\frac{1}{J}\,\mathrm{tr}\big(M\left[\mathrm{Cov}_P(u) - \mathrm{Cov}_P(\psi_\delta(u))\right]M^\top\big) > \|M\mathbb{E}_P[u - \psi_\delta(u)]\|_2^2. \tag{19}$$

The left side is removed variance and the right side is squared clean-tail bias. Clipping is therefore favored when rare large residuals add more variance than useful direction.

The trade-off remains bounded under arbitrary contamination. The contaminated residual fraction is denoted by $\pi \in [0,1]$. The clean clipping bias is $b_\delta = \|\mathbb{E}_P[u - \psi_\delta(u)]\|_2$. For J residual samples $\{u_j\}_{j=1}^J$, the clipped estimate $\widehat{g}_\delta = MJ^{-1}\sum_j \psi_\delta(u_j)$ and the clean target $g_P = M\mathbb{E}_P u$ satisfy:

$$\mathbb{E}\|\widehat{g}_\delta - g_P\|_2^2 \leq \|M\|_{\mathrm{op}}^2\left[\left(b_\delta + 2\pi\delta\sqrt{m}\right)^2 + \frac{m\delta^2}{J}\right]. \tag{20}$$

The terms capture clean-tail bias, bounded contamination, and finite-ensemble variance. Unlike the error for an unbounded linear residual, this bound is independent of corrupted-residual magnitudes.

3.4 Complete REG update

Combining transport, weighting, and clipping yields:

$$\mathcal{A}_t = \frac{1}{J}\psi_{\delta_t}(R_tW_t)(Y_tW_t)^\top, \qquad \gamma_t = \frac{\lambda_t}{\|\mathcal{A}_t\|_F + \varepsilon_\gamma}, \qquad X_{t-1} = X_{t-1}^{\mathrm{pred}} + \gamma_t\mathcal{A}_tE_t, \tag{21}$$

where $\lambda_t > 0$ is the prescribed coupling scale and $\varepsilon_\gamma > 0$ is the numerical floor used in step normalization. The ensemble relation determines the state-space direction. The weights W_t balance

Algorithm 1 Robust Ensemble Guidance

Require: Forward model G, observation y, particles $\{x_T^{(j)}\}_{j=1}^J$, denoisers $\{\phi_t\}_{t=1}^T$, score model s_θ, noise schedule σ, times $\{\tau_t\}_{t=0}^T$, floors $\varepsilon_w, \varepsilon_\gamma$, quantile q, coupling scales $\{\lambda_t\}_{t=1}^T$

for $t = T, T-1, \ldots, 1$ **do**
  Predict $x_{t-1}^{\mathrm{pred},(j)}$ by Equation (2)
  Evaluate $\widehat{x}_t^{(j)} = \phi_t(x_{t-1}^{\mathrm{pred},(j)})$ and $\widehat{y}_t^{(j)} = G(\widehat{x}_t^{(j)})$
  Form the state anomalies E_t, observation anomalies Y_t, and residuals R_t
  Compute W_t by Equation (9); set $\widetilde{R}_t \leftarrow R_t W_t$ and $\widetilde{Y}_t \leftarrow Y_t W_t$
  Compute δ_t by Equation (15) and $\mathcal{A}_t \leftarrow \psi_{\delta_t}(\widetilde{R}_t)\widetilde{Y}_t^\top / J$
  $\gamma_t \leftarrow \lambda_t / (\|\mathcal{A}_t\|_F + \varepsilon_\gamma)$
  $X_{t-1} \leftarrow X_{t-1}^{\mathrm{pred}} + \gamma_t \mathcal{A}_t E_t$
end for
return $\{x_0^{(j)}\}_{j=1}^J$

observation scales while clipping bounds the contribution of extreme residuals. The parameter λ_t controls the normalized coupling; the state-space displacement also depends on E_t.

4 Experiments

We evaluate REG through controlled experiments and three scientific inverse problems: Navier–Stokes inversion, black-hole imaging, and acoustic full-waveform inversion. The controlled study isolates the effects of weighting and clipping, while the scientific tasks assess reconstruction quality across observation conditions. Particle-count comparisons and component ablations further examine the contribution of each modification.

Baselines. We follow the comparison set in InverseBench (Zheng et al., 2025). For **Navier–Stokes**, the baselines are EKI (Iglesias et al., 2013), DPS-fGSG and DPS-cGSG (Zheng et al., 2024; Berahas et al., 2022), DPG (Tang et al., 2023), SCG (Huang et al., 2024), and EnKG (Zheng et al., 2024). **Black-hole imaging** baselines are SMILI (Akiyama et al., 2019), EHT-Imaging (Chael et al., 2019), DPS (Chung et al., 2022a), LGD (Song et al., 2023b), RED-diff (Mardani et al., 2024), PnP-DM (Wu et al., 2024), DAPS (Zhang et al., 2025), DiffPIR (Zhu et al., 2023), and EnKG (Zheng et al., 2024). For **acoustic FWI**, we include Adam, DPS (Chung et al., 2022a), LGD (Song et al., 2023b), DiffPIR (Zhu et al., 2023), DAPS (Zhang et al., 2025), PnP-DM (Wu et al., 2024), RED-diff (Mardani et al., 2024), and EnKG (Zheng et al., 2024).

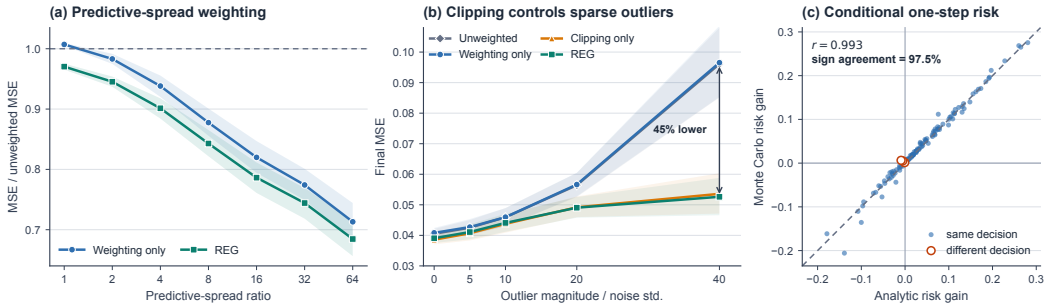


Figure 2: Controlled tests of weighting, clipping, and the conditional one-step risk criterion.

4.1 Controlled validation of the REG mechanisms

Finite-ensemble corrections can fail in two different ways. Unequal predictive scales can cause some observation coordinates to dominate the update, while a small number of extreme residuals can dominate an otherwise informative update. The controlled study separates these effects and tests whether the two parts of REG address the corresponding failure modes.

We use local linear Gaussian problems with $J = 256$ particles in Figure 2(a,b), with state dimension 16 and observation dimension 64. Within each comparison, the methods share the forward map, observation, initial ensemble, correction schedule, and random seed. Panel (a) varies predictive-spread

Table 1: Navier–Stokes relative ℓ_2 error (↓). Cells show mean above (s.d.); bold indicates the lowest mean in each column.

Subsampling	×2			×4			×8		
Method / σ_y	0	1	2	0	1	2	0	1	2
EKI	0.577 (0.138)	0.609 (0.119)	0.673 (0.107)	0.579 (0.145)	0.669 (0.131)	0.805 (0.112)	0.852 (0.167)	0.940 (0.115)	1.116 (0.090)
DPS-fGSG	1.687 (0.156)	1.612 (0.173)	1.454 (0.154)	1.203 (0.122)	1.209 (0.116)	1.200 (0.100)	1.246 (0.108)	1.221 (0.082)	1.260 (0.117)
DPS-cGSG	2.203 (0.314)	2.117 (0.295)	1.746 (0.191)	1.175 (0.079)	1.133 (0.095)	1.114 (0.144)	1.186 (0.117)	1.204 (0.115)	1.218 (0.113)
DPG	0.325 (0.188)	0.408 (0.173)	0.466 (0.171)	0.322 (0.200)	0.361 (0.187)	**0.454** (0.207)	0.596 (0.301)	0.591 (0.262)	0.846 (0.251)
SCG	0.908 (0.600)	0.928 (0.557)	0.966 (0.546)	0.869 (0.513)	0.926 (0.546)	0.929 (0.505)	1.260 (0.135)	1.284 (0.117)	1.347 (0.141)
EnKG	0.116 (0.039)	**0.215** (0.065)	0.341 (0.063)	0.137 (0.115)	0.343 (0.130)	0.555 (0.103)	0.317 (0.305)	**0.556** (0.162)	0.763 (0.106)
REG	**0.100** (0.031)	0.225 (0.103)	**0.334** (0.080)	**0.135** (0.114)	**0.330** (0.081)	0.506 (0.084)	**0.272** (0.261)	0.570 (0.185)	**0.757** (0.109)

Table 2: Matched ablation (×2, $\sigma_y = 0$). Relative ℓ_2 error: mean ± s.d.; bold: lowest mean per comparison. Changes are relative to EnKG at the same J.

(a) Particle-count comparison

J	EnKG	REG	Change
512	0.2752 ± 0.1088	$\mathbf{0.2690} \pm 0.0890$	-2.3%
1024	0.2029 ± 0.1250	$\mathbf{0.1493} \pm 0.0463$	-26.4%
2048	0.1156 ± 0.0385	$\mathbf{0.1000} \pm 0.0314$	-13.4%

(b) Component ablation ($J = 1024$)

Method	Relative ℓ_2 ↓	Change
EnKG	0.2029 ± 0.1250	–
Clipping only	0.1575 ± 0.0576	-22.4%
Weighting only	0.1670 ± 0.0556	-17.7%
REG	$\mathbf{0.1493} \pm 0.0463$	-26.4%

imbalance without contamination. Panel (b) fixes equal predictive spread and varies the outlier magnitude after corrupting 10% of the residual coordinates. Both panels report final ensemble-mean MSE with paired 95% bootstrap confidence intervals. Panel (c) separately compares analytic one-step risk gains with Monte Carlo estimates on 120 independently generated problems.

Figure 2 demonstrates the complementary roles of weighting and clipping. Weighting offers little benefit with equal predictive spreads, but improves reconstruction as scale imbalance grows (a). Clipping controls the deterioration caused by increasingly large residual outliers (b). These results show how coordinate scaling and residual influence control address distinct sources of reconstruction error. The agreement between analytic and Monte Carlo risk gains further supports the one-step analysis (c).

4.2 Navier–Stokes inverse problem

Problem setting. We recover initial vorticity from spatially subsampled, noisy observations of two-dimensional incompressible Navier–Stokes flow. Nonlinear evolution and loss of fine-scale information make recovery challenging, testing guidance without differentiating through the PDE solver. We report relative ℓ_2 error as mean and standard deviation across cases.

Results. Table 1 compares relative ℓ_2 error across nine combinations of subsampling and noise. REG achieves the lowest mean error in six settings and improves over EnKG in seven, including all subsampling factors at $\sigma_y = 2$. These gains cover both noiseless and noisy observations, demonstrating improved recovery across varied observation conditions. Also, figure 3 complements the quantitative comparison: REG preserves the main vortex structures while reducing reconstruction errors relative to other methods.

Particle-count comparison and component ablation. We next examine whether the complementary effects observed in the controlled study carry over to nonlinear flow inversion. We compare REG with EnKG (Zheng et al., 2024) at ×2 subsampling and $\sigma_y = 0$. Both methods are evaluated on the same cases with $J \in \{512, 1024, 2048\}$ particles. At $J = 1024$, *Clipping only* and *Weighting only* isolate residual clipping and predictive-spread weighting, respectively.

Table 2 shows lower mean error and lower variability for REG than EnKG at every tested particle budget. Both weighting and clipping improve the correction individually, and their combination

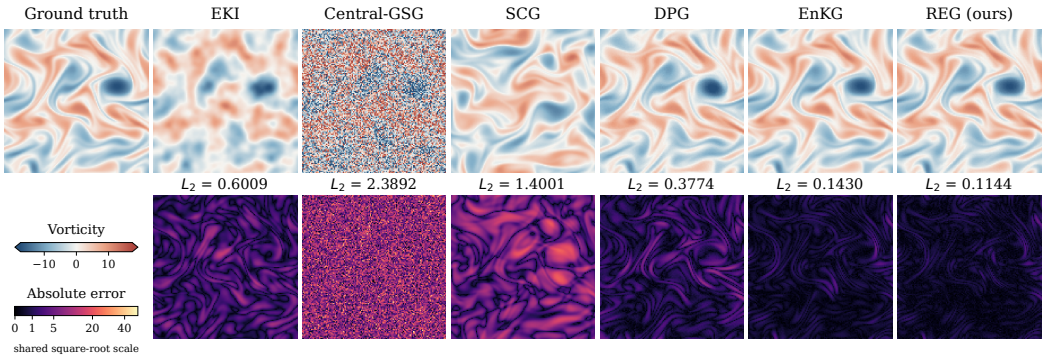


Figure 3: Navier–Stokes reconstructions (top) and absolute errors (bottom) for the same test case, with shared color scales across methods.

Table 3: Black-hole imaging over 100 images. Cells show mean above (s.d.); bold and underline mark the best and second-best PSNR and blurred PSNR (dB).

Obs. time	3%				10%				100%			
Method	PSNR↑	Blur↑	$\widetilde{\chi}^2_{cp}$	$\widetilde{\chi}^2_{\log ca}$	PSNR↑	Blur↑	$\widetilde{\chi}^2_{cp}$	$\widetilde{\chi}^2_{\log ca}$	PSNR↑	Blur↑	$\widetilde{\chi}^2_{cp}$	$\widetilde{\chi}^2_{\log ca}$
SMILI	18.51 (1.39)	23.08 (2.12)	1.478 (0.428)	4.348 (3.827)	20.85 (2.90)	25.24 (3.86)	1.209 (0.169)	21.788 (12.491)	22.67 (3.13)	27.79 (4.02)	1.878 (0.952)	17.612 (10.299)
EHT-Imaging	21.72 (3.39)	25.66 (5.04)	1.507 (0.485)	1.695 (0.539)	22.67 (3.46)	26.66 (3.93)	1.166 (0.156)	1.240 (0.205)	24.28 (3.63)	28.57 (4.52)	1.251 (0.250)	1.259 (0.316)
DPS	24.20 (3.72)	<u>30.83</u> (5.58)	8.024 (24.336)	5.007 (5.750)	24.36 (3.72)	30.79 (5.75)	13.052 (43.087)	6.614 (26.789)	25.86 (3.90)	<u>32.94</u> (6.19)	8.759 (37.784)	5.456 (24.185)
LGD	22.51 (3.76)	28.50 (5.49)	15.825 (16.838)	12.862 (12.663)	22.08 (3.75)	27.48 (5.09)	10.775 (21.684)	13.375 (56.397)	21.22 (3.64)	26.06 (4.98)	13.239 (17.231)	13.233 (39.107)
RED-diff	20.74 (2.62)	26.10 (3.35)	6.713 (6.925)	9.128 (19.052)	22.53 (3.02)	27.67 (4.53)	2.488 (2.925)	4.916 (13.221)	23.77 (4.13)	29.13 (6.22)	1.853 (0.938)	2.050 (2.361)
PnP-DM	<u>24.25</u> (3.45)	30.49 (4.93)	2.201 (1.352)	1.668 (0.551)	<u>24.57</u> (3.47)	<u>30.80</u> (5.22)	1.433 (0.417)	1.336 (0.478)	<u>26.07</u> (3.70)	32.88 (6.02)	1.311 (0.195)	1.199 (0.221)
DAPS	23.54 (3.28)	29.48 (4.88)	3.647 (3.287)	2.329 (1.354)	23.99 (3.56)	30.10 (5.13)	1.545 (0.705)	2.253 (9.903)	25.60 (3.64)	32.78 (5.68)	1.300 (0.324)	1.229 (0.532)
DiffPIR	24.12 (3.25)	30.45 (4.88)	14.085 (14.105)	10.545 (8.860)	23.84 (3.59)	30.04 (5.03)	5.374 (3.733)	5.205 (5.556)	25.01 (4.64)	31.86 (6.56)	3.271 (1.623)	2.970 (1.202)
EnKG	20.77 (5.95)	25.58 (7.54)	14.019 (22.133)	23.559 (132.848)	18.00 (9.18)	22.54 (11.88)	26.675 (41.330)	204.890 (675.821)	21.79 (7.88)	27.48 (10.36)	25.173 (32.107)	29.759 (71.019)
REG	**26.97** (3.79)	**34.08** (5.56)	4.278 (7.335)	1.740 (1.786)	**26.19** (4.54)	**33.31** (6.52)	7.198 (11.923)	2.908 (3.691)	**28.16** (4.58)	**36.84** (6.96)	5.828 (9.397)	2.852 (2.726)

performs best, supporting their complementary contribution beyond increasing ensemble size. More particles improve both methods, but the relative advantage of REG is not monotone in particle count.

4.3 Black-hole imaging

Problem setting. We reconstruct black-hole images from sparse interferometric measurements with atmospheric and thermal noise. Closure phases and log closure amplitudes reduce atmospheric sensitivity but leave image details weakly constrained. We follow InverseBench on 100 images at observation-time ratios of 3%, 10%, and 100%, with $J = 256$ particles for paired EnKG and REG.

Evaluation metrics. We report PSNR and its telescope-resolution blurred counterpart. Normalized chi-squared values near one indicate agreement at the expected noise level.

Results. Table 3 compares reconstruction quality and measurement consistency at three observation-time ratios. REG achieves the highest PSNR and blurred PSNR in all three settings and improves measurement consistency over other methods. These results show improved image recovery across observation durations, including the shortest observation time. The gains in both PSNR measures indicate improved fidelity at the original image resolution and at the telescope-resolved scale. Figure 4 further shows improved preservation of the ring structure and brightness distribution.

4.4 Acoustic full-waveform inversion

Problem setting. We recover subsurface velocity maps from seismic measurements using the InverseBench acoustic wave model. Sparse recordings and initialization sensitivity make recovery challenging. Paired EnKG and REG use $J = 64$ particles on ten official test cases.

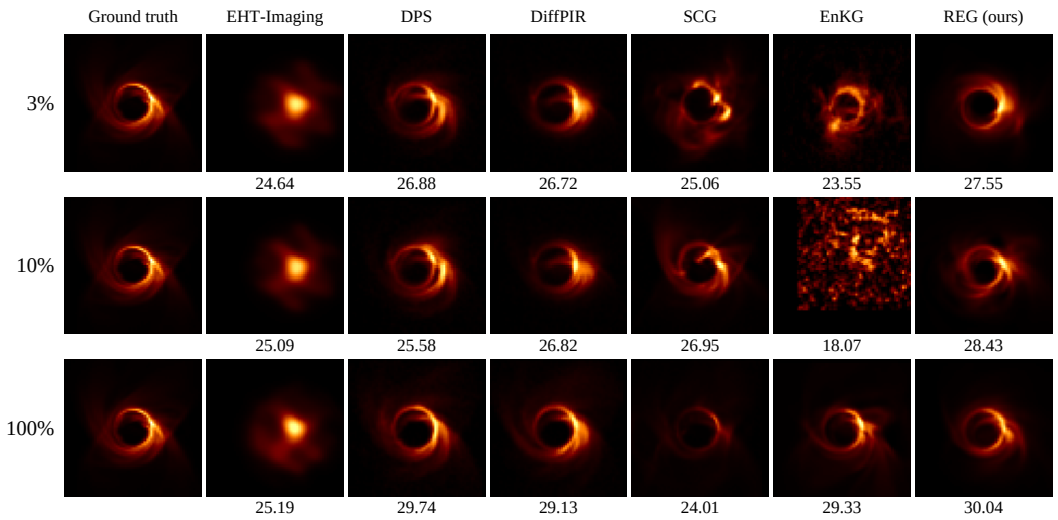


Figure 4: Selected black-hole reconstructions of one case at 3%, 10%, and 100% observation time, with PSNR (dB). EnKG and REG share observations; other baselines retain native realizations.

Table 4: Acoustic FWI results. Mean (s.d.). Bold marks the best mean in each column.

Method	Relative $\ell_2 \downarrow$	PSNR (dB) $\uparrow$	SSIM $\uparrow$	Data misfit $\downarrow$
Adam	0.333 (0.086)	9.968 (2.083)	0.305 (0.120)	115.14 (52.10)
DPS	0.250 (0.154)	14.111 (6.820)	0.491 (0.161)	155.08 (92.17)
LGD	0.244 (0.024)	12.288 (0.889)	0.341 (0.047)	258.47 (26.40)
DiffPIR	0.204 (0.129)	**16.113** (6.962)	0.554 (0.191)	**88.53** (56.91)
DAPS	0.201 (0.103)	14.914 (4.184)	0.321 (0.067)	111.13 (71.33)
PnP-DM	0.259 (0.075)	11.983 (2.269)	0.431 (0.073)	308.84 (26.34)
RED-diff	0.319 (0.102)	10.372 (2.650)	0.280 (0.108)	94.67 (41.33)
EnKG	0.170 (0.034)	15.437 (1.308)	0.565 (0.065)	269.64 (51.02)
REG	**0.160** (0.027)	15.903 (1.084)	**0.574** (0.064)	268.85 (49.51)

Evaluation metrics. We report relative ℓ_2, PSNR, SSIM, and seismic data misfit. For paired EnKG and REG, image metrics are computed on each ensemble-mean reconstruction, then averaged over test cases. Published baselines retain the official benchmark evaluation.

Results. Table 4 compares velocity-map reconstruction and seismic data misfit over ten test cases. REG achieves the lowest mean relative error and highest mean SSIM among the compared methods, while improving relative error, PSNR, and SSIM over EnKG. Relative error and SSIM assess complementary aspects of recovery. Their joint improvement reflects more accurate velocity values together with better structural agreement. These results demonstrate improved velocity-map recovery with an ensemble of only 64 particles. Additional reconstructions and absolute-error maps are provided in Appendix C.3.

5 Conclusion

We presented REG for generative inverse problems. REG combines predictive-spread weighting with adaptive residual clipping to balance observation scales and limit the influence of extreme discrepancies. Both reuse the ensemble and its forward predictions, requiring neither forward derivatives nor additional model evaluations. Under a local linear Gaussian model, our analysis identifies conditions for reduced one-step estimation risk, bounds coordinate influence, and establishes when improvement persists with a finite ensemble. Experiments support its effectiveness across varied forward operators, observation regimes, and problem scales.

AI USE STATEMENT

Generative AI tools were used during manuscript preparation to assist with language polishing, grammatical correction, and LaTeX formatting.

ETHICS STATEMENT

This work studies scientific inverse problems using controlled simulations and existing benchmarks for fluid flow, black-hole imaging, and seismic inversion. The reported experiments do not involve recruiting human participants or collecting personal data.

REPRODUCIBILITY STATEMENT

Section 3 defines the particle prediction, weighting, clipping, and correction steps, as summarized in Algorithm 1. Appendix A provides the assumptions and proofs for the local analysis, distinguishing fixed-step risk calculations from normalized algorithmic updates. Section 4.1 describes the controlled tests and their paired evaluation. Sections 4.2–4.4 identify the scientific benchmarks, observation regimes, evaluation metrics, and comparison protocols. Section 4 cites benchmark and baseline sources and distinguishes paired evaluations from published baseline results.

REFERENCES

Kazunori Akiyama, Fumie Tazaki, Kotaro Moriyama, Ilje Cho, Shiro Ikeda, Mahito Sasada, Hiroki Okino, and Mareki Honma. Smili: Sparse modeling imaging library for interferometry, 2019.

Albert S Berahas, Liyuan Cao, Krzysztof Choromanski, and Katya Scheinberg. A theoretical and empirical comparison of gradient approximations in derivative-free optimization. *Foundations of Computational Mathematics*, 22(2):507–560, 2022.

Dirk Blömker, Claudia Schillings, Philipp Wacker, and Simon Weissmann. Well posedness and convergence analysis of the ensemble kalman inversion. *Inverse Problems*, 35(8):085007, 2019.

Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *International conference on machine learning*, pp. 537–546. PMLR, 2017.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. 48(4): 1148–1185, 2012.

Andrew Chael, Katie Bouman, Michael Johnson, Maciek Wielgus, Lindy Blackburn, Chi-Kwan Chan, Joseph Rachid Farah, Daniel Palumbo, and Dominic Pesce. eht-imaging: v1. 1.0: Imaging interferometric data with regularized maximum likelihood, 2019.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. 2021.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. 2022a.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. volume 35, pp. 25683–25696, 2022b.

Zhiyan Ding and Qin Li. Ensemble kalman inversion: mean-field limit and convergence analysis. *Statistics and computing*, 31(1):9, 2021.

Samuel Duffield and Sumeetpal S Singh. Ensemble kalman inversion for general likelihoods. *Statistics & Probability Letters*, 187:109523, 2022.

Alexandre A Emerick and Albert C Reynolds. Ensemble smoother with multiple data assimilation. *Computers & Geosciences*, 55:3–15, 2013.

Geir Evensen. The ensemble kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, 53(4):343–367, 2003.

Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. volume 33, pp. 6840–6851, 2020.

Yujia Huang, Adishree Ghatare, Yuanzhe Liu, Ziniu Hu, Qinsheng Zhang, Chandramouli S Sastry, Siddharth Gururani, Sageev Oore, and Yisong Yue. Symbolic music generation with non-differentiable rule guided diffusion. 2024.

Peter J Huber. Robust estimation of a location parameter. pp. 492–518, 1992.

Peter J Huber and Elvezio M Ronchetti. *Robust statistics*. John Wiley & Sons, 2011.

Marco A Iglesias. A regularizing iterative ensemble kalman method for pde-constrained inverse problems. *Inverse Problems*, 32(2):025002, 2016.

Marco A Iglesias, Kody JH Law, and Andrew M Stuart. Ensemble kalman methods for inverse problems. *Inverse Problems*, 29(4):045001, 2013.

Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jon Tamir. Robust compressed sensing mri with deep generative priors. volume 34, pp. 14938–14954, 2021.

Zahra Kadkhodaie and Eero Simoncelli. Stochastic solutions for linear inverse problems using the prior implicit in a denoiser. volume 34, pp. 13242–13254, 2021.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. volume 35, pp. 26565–26577, 2022.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. volume 35, pp. 23593–23606, 2022.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, pp. 110–133, 2017.

Nikola B Kovachki and Andrew M Stuart. Ensemble kalman inversion: a derivative-free technique for machine learning tasks. *Inverse Problems*, 35(9):095005, 2019.

Xiang Li, Soo Min Kwon, Shijun Liang, Ismail R Alkhouri, Saiprasad Ravishankar, and Qing Qu. Decoupled data consistency with diffusion purification for image restoration. *arXiv preprint arXiv:2403.06054*, 2024.

Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 28027–28053, 2024.

Ricardo A Maronna, R Douglas Martin, Victor J Yohai, and Matías Salibián-Barrera. *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.

Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (red). *SIAM journal on imaging sciences*, 10(4):1804–1844, 2017.

Claudia Schillings and Andrew M Stuart. Analysis of the ensemble kalman filter for inverse problems. *SIAM Journal on Numerical Analysis*, 55(3):1264–1290, 2017.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International conference on learning representations*, 2023a.

Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In *International Conference on Machine Learning*, pp. 32483–32498. PMLR, 2023b.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. volume 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. 2020.

Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. 2021.

Andrew M Stuart. Inverse problems: a bayesian perspective. *Acta numerica*, 19:451–559, 2010.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

Haoyue Tang, Tian Xie, Aosong Feng, Hanyu Wang, Chenyang Zhang, and Yang Bai. Solving noisy inverse problems via posterior sampling: A policy gradient view-point. In *The Symbiosis of Deep Learning and Differential Equations III*, 2023.

Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005.

Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE global conference on signal and information processing*, pp. 945–948. IEEE, 2013.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. 2022.

Zihui Wu, Yu Sun, Yifan Chen, Bingliang Zhang, Yisong Yue, and Katherine L Bouman. Principled probabilistic imaging using diffusion models as plug-and-play priors. *Advances in Neural Information Processing Systems*, 37:118389–118427, 2024.

Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, and Yang Song. Improving diffusion inverse problem solving with decoupled noise annealing. pp. 20895–20905, 2025.

Hongkai Zheng, Wenda Chu, Austin Wang, Nikola Kovachki, Ricardo Baptista, and Yisong Yue. Ensemble kalman diffusion guidance: A derivative-free method for inverse problems. *arXiv preprint arXiv:2409.20175*, 2024.

Hongkai Zheng, Wenda Chu, Bingliang Zhang, Zihui Wu, Austin Wang, Berthy Feng, Caifeng Zou, Yu Sun, Nikola Kovachki, Zachary Ross, et al. Inversebench: Benchmarking plug-and-play diffusion priors for inverse problems in physical sciences. In *International Conference on Learning Representations*, volume 2025, pp. 90912–90940, 2025.

Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1219–1229. IEEE, 2023.

SUPPLEMENTARY MATERIAL

This supplement uses the notation of the main paper. Appendix A provides assumptions and proofs; Appendix B describes experimental protocols and evaluation conventions; Appendix C presents additional reconstructions.

A ASSUMPTIONS AND PROOFS FOR THE LOCAL ANALYSIS

This appendix specifies the scope of the statements in Section 3. The weighting results concern a frozen local linear model. The clipping results first hold with the transport and threshold fixed. We then discuss an estimated threshold.

A.1 LOCAL REFERENCE, LINEARIZATION, AND STEP NORMALIZATION

Fix a reverse step and write $F = G \circ \phi_t$ for the denoised forward map, $C = C_t^{xx}$ for state covariance, and $H = H_t$ for its local linearization. We omit t on $E, Y, R, W, \widetilde{Y}, \delta, \mathcal{A}$ and C^{xy} when referring to the corresponding quantities at this fixed step. The analysis uses the affine model $F(x) = F(x_t^{\mathrm{ref}}) + H(x - x_t^{\mathrm{ref}})$ on the region of interest and assumes $F(x_t^{\mathrm{ref}}) = G(x^\star)$. Thus the local target is a state in the particle coordinates whose denoised prediction has the correct observation. This assumption does not identify a noisy particle with the clean state. Centered rows give $Y = EH^\top$, hence:

$$C^{xy} = J^{-1}E^\top Y = J^{-1}E^\top EH^\top = CH^\top. \tag{22}$$

For nonlinear F, remainder matrix B_{nl} gives $Y = EH^\top + B_{\mathrm{nl}}$ and $C^{xy} = CH^\top + J^{-1}E^\top B_{\mathrm{nl}}$.

Let $e = \bar{x}_{t-1}^{\mathrm{pred}} - x_t^{\mathrm{ref}}$. If the mean forward prediction differs from $G(x^\star) + He$ by a deterministic mismatch $b \in \mathbb{R}^m$, the mean residual is $-He - b + \xi$. Here α is the fixed theoretical counterpart of γ_t, Q is a deterministic observation geometry, and e^+ is the error after correction. The update gives

$$e^+ = (I_n - \alpha CH^\top QH)e - \alpha CH^\top Qb + \alpha CH^\top Q\xi. \tag{23}$$

The main-text risk sets $b = 0$ and fixes C, H, Q, α, averaging over zero-mean noise of covariance Γ. A nonzero b contributes to the deterministic error term and can change the ranking of two geometries. Dependence of these quantities on the same noise requires further analysis.

The implemented normalization in Equation (21) obeys

$$\|\gamma_t \mathcal{A}_t E_t\|_F \le \lambda_t \frac{\|\mathcal{A}_t\|_F}{\|\mathcal{A}_t\|_F + \varepsilon_\gamma} \|E_t\|_{\mathrm{op}} \le \lambda_t \|E_t\|_{\mathrm{op}}. \tag{24}$$

It bounds the displacement relative to the ensemble spread, rather than setting its length to λ_t. The risk comparison uses a common fixed α.

A.2 SCALE BALANCING AND CONDITIONING

Let $\Sigma_y = HCH^\top$ be prediction covariance, $v_k^2 = (\Sigma_y)_{kk}$, and $D = \mathrm{diag}(v_1^2, \dots, v_m^2)$, with $D_{kk} > 0$. The population geometry Q uses these population spreads; its empirical counterpart is $\widehat{Q} = W_t^2$. In the idealized zero-floor limit, define $c = (m^{-1} \sum_k D_{kk}^{-1/2})^{-1}$. Equation (9) applied to population spreads gives $Q = c^2 D^{-1}$. Thus

$$Q^{1/2} \Sigma_y Q^{1/2} = c^2 D^{-1/2} \Sigma_y D^{-1/2}. \tag{25}$$

The diagonal entries are equal, while off-diagonal correlations remain. This is marginal standardization, not a claim that diagonal weights remove all correlations. With $\varepsilon_w > 0$, the exact weights are proportional to $(\sqrt{D_{kk}} + \varepsilon_w)^{-1}$; Equation (11) is the zero-floor limit.

Write $D^{-1/2} \Sigma_y D^{-1/2} = I_m + K$, where K is symmetric and $\|K\|_{\mathrm{op}} = \rho < 1$. For unit vector z,

$$1 - \rho \le z^\top (I_m + K) z \le 1 + \rho. \tag{26}$$

The smallest and largest eigenvalues therefore lie in this interval, proving Equation (12). A sufficient condition for a strict improvement over the unweighted covariance is $\kappa(\Sigma_y) > (1 + \rho)/(1 - \rho)$.

For positive coordinate rescaling $S = \mathrm{diag}(s_1, \ldots, s_m)$, let $R' = RS$, $Y' = YS$. In the zero-floor limit $W' = aS^{-1}W$ for some common $a > 0$, so $R'W' = aRW$, $Y'W' = aYW$, and $\delta' = a\delta$. Homogeneity of clipping gives $\mathcal{A}' = a^2\mathcal{A}$, where primes denote the rescaled quantities. Thus $\mathcal{A}/\|\mathcal{A}\|_F$ is invariant when $\mathcal{A} \neq 0$ and the normalization floor is zero.

A.3 Contraction of observable error

Assume C is positive definite and $\Sigma_y = HCH^\top = D$ is diagonal and positive definite, with $Q = c^2D^{-1}$ as above. Define $B = HC^{1/2}$ and the covariance-whitened error $z = C^{-1/2}e$. Since $BB^\top = D$, the orthogonal projector onto the row space of B is:

$$\Pi = B^\top(BB^\top)^{-1}B \tag{27}$$

It satisfies $\Pi^\top = \Pi$ and $\Pi^2 = \Pi$. Writing $z^+ = C^{-1/2}e^+$ for the whitened error after correction, the noiseless update and its squared error satisfy:

$$z^+ = (I_n - \alpha c^2\Pi)z, \qquad \|z^+\|_2^2 = (1 - \alpha c^2)^2\|\Pi z\|_2^2 + \|(I_n - \Pi)z\|_2^2. \tag{28}$$

For $0 < \alpha c^2 < 2$, the observable component contracts strictly unless it is already zero; the null-space component is unchanged. We retain c^2 explicitly so α has the same meaning as in the risk analysis. For singular C, the same argument applies on its range, provided H has full row rank on that range and the error lies in it.

A.4 One-step risk and finite-ensemble weight estimation

Set $b = 0$ in Equation (23). The mixed term vanishes because $\mathbb{E}\xi = 0$, and the identity $\mathbb{E}\|CH^\top Q\xi\|_2^2 = \mathrm{tr}(CH^\top Q\Gamma QHC)$ gives Equation (13). Only the first two noise moments are needed; Gaussian noise is a sufficient special case. Expanding the deterministic square yields

$$\mathcal{R}(Q) = \|e\|_2^2 - 2\alpha e^\top d_Q + \alpha^2(\|d_Q\|_2^2 + \nu_Q). \tag{29}$$

Subtracting $\mathcal{R}(I_m)$ proves Equation (14). In particular, if $a = e^\top(d_Q - d_I) > 0$ and $b_1 = \|d_Q\|_2^2 - \|d_I\|_2^2 + \nu_Q - \nu_I$, improvement holds for $0 < \alpha < 2a/b_1$ when $b_1 > 0$, and for every positive α when $b_1 \leq 0$ in this frozen model.

For completeness, the local sensitivity in the main text can be made explicit. Set $U = CH^\top$ for the frozen transport, and let $B_Q > 0$ bound the geometries: $\|Q\|_{\mathrm{op}}, \|\widehat{Q}\|_{\mathrm{op}} \leq B_Q$. The difference of squared norms and the Frobenius representation of the noise term imply

$$|\mathcal{R}(\widehat{Q}) - \mathcal{R}(Q)| \leq L_{\mathcal{R}}\|\widehat{Q} - Q\|_{\mathrm{op}}, \tag{30}$$

$$L_{\mathcal{R}} = 2\alpha\|U\|_{\mathrm{op}}\|He\|_2(\|e\|_2 + \alpha\|U\|_{\mathrm{op}}B_Q\|He\|_2) + 2\alpha^2\|U\|_{\mathrm{op}}^2B_Q\,\mathrm{tr}(\Gamma). \tag{31}$$

If $L_{\mathcal{R}} > 0$, a population advantage $\Delta_{\mathcal{R}} > 0$ therefore persists when $\|\widehat{Q} - Q\|_{\mathrm{op}} < \Delta_{\mathcal{R}}/L_{\mathcal{R}}$. If $L_{\mathcal{R}} = 0$, risk is constant on the neighborhood considered.

The stated rate follows under independent, identically distributed calibration predictions with uniformly bounded sub-Gaussian norms and marginal standard deviations v_k in $[v_{\min}, v_{\max}]$, where $v_{\min} > 0$. Write $\widehat{v}_k = v_{t,k}$ for their empirical estimates. Concentration of the empirical first and second moments, followed by a union bound over coordinates, gives

$$\max_k |\widehat{v}_k^2 - v_k^2| \leq C_0\left(\sqrt{\frac{\log(2m/\eta)}{J}} + \frac{\log(2m/\eta)}{J}\right) \tag{32}$$

with probability at least $1 - \eta$. Here C_0 depends on the uniform sub-Gaussian bound. This follows by applying Bernstein's inequality to the squared centered coordinates and a sub-Gaussian bound to each sample mean. On the event that empirical variances remain above $v_{\min}^2/2$, square root, inverse scale, unit-mean normalization, and squaring are Lipschitz on this bounded interval. The diagonal operator norm is the largest coordinate error, so for J large enough relative to $\log(2m/\eta)$,

$$\|\widehat{Q} - Q\|_{\mathrm{op}} = O\left(\sqrt{\frac{\log(2m/\eta)}{J}}\right). \tag{33}$$

The constants depend on the spread bounds and the chosen floor. Population and empirical weights must use the same floor; otherwise its deterministic approximation error must also be included. This argument controls weight estimation with C, H, e fixed.

A.5 Clipping influence and adaptive thresholds

For scalar residual a with $|a| \le \delta$, $a\psi_\delta(a) = \psi_\delta(a)^2 = a^2$; otherwise $a\psi_\delta(a) = |a|\delta \ge \delta^2$. This proves Equation (17). Fix W, $\widetilde{Y}$, and δ, and change only weighted residual entry (j,k) from a to a'. If $\mathbf{e}_j$ is the jth coordinate unit vector in $\mathbb{R}^J$, then

$$\Delta\mathcal{A} = \frac{\psi_\delta(a') - \psi_\delta(a)}{J}\mathbf{e}_j(\widetilde{Y}_{:,k})^\top. \tag{34}$$

Since the scalar difference is at most 2δ in magnitude, the norm of this rank-one matrix proves Equation (18).

Recomputing the threshold also changes other clipped entries. Let $\widetilde{R}'$ differ from $\widetilde{R} = RW$ in one entry, with recomputed threshold δ' and coupling $\mathcal{A}'$. Adding and subtracting $\psi_\delta(\widetilde{R}')$ yields

$$\|\mathcal{A}' - \mathcal{A}\|_F \le \frac{2\delta}{J}\|\widetilde{Y}_{:,k}\|_2 + \sqrt{\frac{m}{J}}\,|\delta' - \delta|\|\widetilde{Y}\|_{\mathrm{op}}, \tag{35}$$

because $|\psi_{\delta'}(a) - \psi_\delta(a)| \le |\delta' - \delta|$.

At the population level, under an upper-tail contamination fraction π, a quantile level $q < 1 - \pi$ remains below a finite clean-distribution quantile, provided that quantile is finite. For $N = Jm$ finite entries and r arbitrary replacements, the analogous condition is that the order-statistic indices used by the empirical quantile lie at or below $N - r$. For example, for linear interpolation at index $1+(N-1)q$, it suffices that $\lceil 1+(N-1)q \rceil \le N-r$. This states the finite-sample rounding condition behind the contamination-fraction interpretation. A shared measurement can affect multiple entries, so the fraction concerns threshold entries.

A.6 Variance, tail bias, and contamination risk

For Equation (19), assume $u_1, \dots, u_J$ are independent samples from the clean distribution P with finite second moment, and fix M and δ independently of these samples. Write $\mu = \mathbb{E}_P u$ and $\mu_\delta = \mathbb{E}_P\psi_\delta(u)$. The unclipped mean is unbiased for $M\mu$; clipping adds bias $M(\mu_\delta - \mu)$. Therefore

$$\mathbb{E}\left\|MJ^{-1}\sum_j u_j - M\mu\right\|_2^2 = J^{-1}\operatorname{tr}(M\operatorname{Cov}_P(u)M^\top), \tag{36}$$

$$\mathbb{E}\left\|MJ^{-1}\sum_j \psi_\delta(u_j) - M\mu\right\|_2^2 = \|M(\mu_\delta - \mu)\|_2^2 + J^{-1}\operatorname{tr}(M\operatorname{Cov}_P(\psi_\delta(u))M^\top). \tag{37}$$

Subtracting proves the exact criterion in the main text, with both covariances taken under P.

For a scalar residual a, clipping is 1-Lipschitz, so an independent copy a' gives $\operatorname{Var}(\psi_\delta(a)) = \frac{1}{2}\mathbb{E}(\psi_\delta(a) - \psi_\delta(a'))^2 \le \frac{1}{2}\mathbb{E}(a - a')^2 = \operatorname{Var}(a)$. This reduces each marginal variance, but does not imply a covariance ordering after an arbitrary map M when coordinates are correlated. If coordinates are independent, the covariance matrices are diagonal and the transported variance cannot increase. If they are also symmetric about zero, the clipping bias vanishes and the risk cannot increase, with strict improvement when a direction used by M has strictly reduced variance. For general residuals, the explicit variance–bias criterion is required.

For Equation (20), assume instead independent samples from $(1-\pi)P + \pi P_{\mathrm{bad}}$, where P has finite first moment and P_{bad} is arbitrary. Clipping satisfies $\|\psi_\delta(u)\|_2 \le \delta\sqrt{m}$, ensuring all required moments exist and giving the following bias and variance bounds:

$$\|\mathbb{E}\psi_\delta(u) - \mathbb{E}_P u\|_2 \le b_\delta + 2\pi\delta\sqrt{m}, \tag{38}$$

$$\operatorname{tr}(\operatorname{Cov}(\psi_\delta(u))) \le \mathbb{E}\|\psi_\delta(u)\|_2^2 \le m\delta^2. \tag{39}$$

The bias–variance decomposition and $\|Mz\|_2 \le \|M\|_{\mathrm{op}}\|z\|_2$ for any $z \in \mathbb{R}^m$ prove the claimed bound.

These sample-mean results use a fixed threshold and transport. They also apply conditionally if these are estimated on independent calibration data. Conditioning on a quantile computed from the

same residuals does not preserve their independence. To relate the two cases, compare a random $\widehat{\delta}$ with a fixed δ_0. Lipschitz dependence on the threshold gives, for any slack parameter $\beta > 0$,

$$\mathbb{E}\|\widehat{g}_{\widehat{\delta}} - g_P\|_2^2 \le (1+\beta)\mathbb{E}\|\widehat{g}_{\delta_0} - g_P\|_2^2 + (1+\beta^{-1})\|M\|_{\mathrm{op}}^2 m\mathbb{E}|\widehat{\delta} - \delta_0|^2. \tag{40}$$

The additional term must be controlled to obtain a sample-dependent threshold guarantee. Likewise, dependent particles add cross-covariance terms to the variance of their mean.

B Experimental protocols and evaluation details

B.1 Solver settings

Navier–Stokes. The nine-setting sweep uses $J = 2048$, batch size 64, 80 steps, two updates, guidance scale 2.0, numerical floor 0.01, and quantile 0.995. It also sets `precision_shrinkage`=1.0, `precision_power`=2.0, observation-noise standard deviation σ_y, and `guidance_decay_power`=0.05.

Black-hole imaging. The paired launcher uses the official `nfft` observation backend, $J = 256$, batch size 32, 80 reverse steps, and two updates at a correction. The numerical floor and quantile are 0.01 and 0.75. The observation seed is 0 and the inference seed is $100000 +$ case ID.

Acoustic FWI. The paired ensemble launcher uses $J = 64$, batch size 4, four reverse steps, one update per step, numerical floor 0.01, quantile 0.75, and correction stride 1. It loads method-specific guidance scales from validation selection. The observation seed is 0 and the inference seed is $100000 +$ case ID.

B.2 Controlled experiments

Panels (a) and (b) of Figure 2 use state dimension 16, observation dimension 64, and $J = 256$. Panel (a) uses spread ratios $1, 2, 4, 8, 16, 32, 64$ without contamination. Panel (b) fixes the spread ratio to one and the contaminated fraction to 10%, with outlier magnitudes $0, 5, 10, 20, 40$ times the noise standard deviation.

Panel (c) uses a separate linear experiment with state dimension 12 and observation dimension 48. It compares analytic and empirical risk gains on 120 generated problems. Population and empirical weighted directions are rescaled to match the corresponding unweighted linear-update Frobenius norm. Empirical predictions in this experiment use standard deviations with denominator J.

B.3 Metrics and visualization conventions

For a displayed reconstruction $\widehat{x}$, relative error is $\|\widehat{x} - x^\star\|_2/\|x^\star\|_2$. Absolute-error images display $|\widehat{x} - x^\star|$. NS qualitative scores use the displayed particle mean.

Paired FWI reconstruction metrics are evaluated on the ensemble mean. The saved seismic misfit is $\sqrt{\frac{1}{2}\|G(\widehat{x}) - y\|_2^2}$, with case metrics then aggregated. The NS and paired FWI records use population standard deviations.

Black-hole qualitative panels use the existing translation-alignment grid and the saved particle with the highest aligned PSNR against the reference image. Both ensemble methods use the same rule. Their blurred PSNR column corresponds to the saved evaluator field with blur factor $f = 20$. EnKG and REG share observations; other qualitative baselines retain native observation realizations for the same target and observation-time setting.

In Figure 3, NS reconstructions share a symmetric vorticity scale. Absolute-error colors use a common square-root mapping with ticks in the original error units. Black-hole panels share the range $[0, 1]$.

C ADDITIONAL QUALITATIVE RESULTS

C.1 NAVIER–STOKES INVERSION

Figure 5 extends the multi-method comparison to another target under the original matched configuration. REG preserves the vortex structure and attains the lowest reconstruction error among the displayed methods.

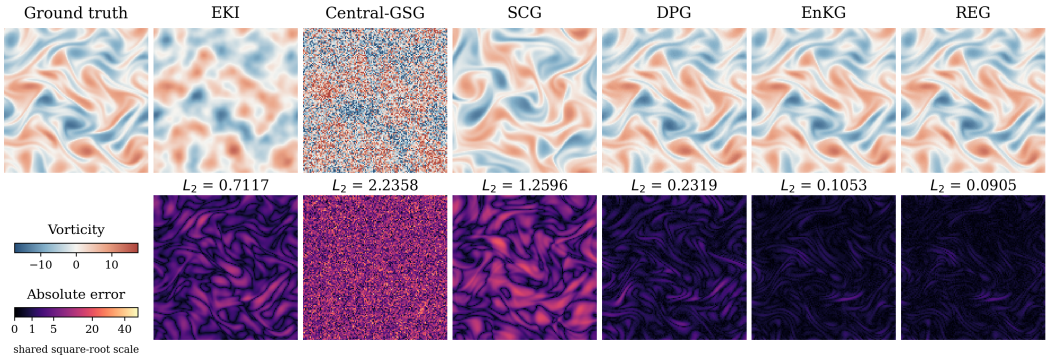


Figure 5: NS reconstructions and absolute errors at $\times 2$, $\sigma_y = 0$. EnKG and REG use $J = 2048$; numbers are relative ℓ_2 errors of the displayed reconstructions.

Figure 6 holds the target fixed across the nine settings using the noise-aware configuration in Appendix B.1. Dense, noiseless observations give the most accurate recovery, while stronger noise and coarser sampling progressively obscure finer structures.

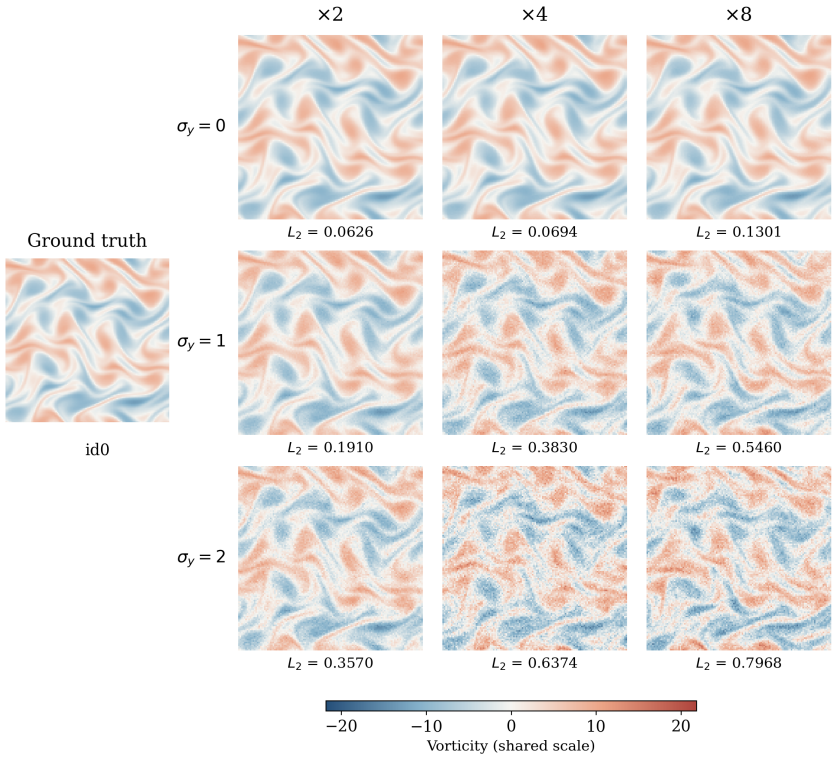


Figure 6: REG reconstructions of NS id0 across downsampling factors $\times 2/4/8$ and noise levels $\sigma_y = 0/1/2$, with $J = 2048$. All panels share a vorticity scale.

C.2 BLACK-HOLE IMAGING

Figure 7 compares three additional targets. DPS loses part of the ring or distorts its shape, while EnKG shows structural distortion or strong granular artifacts. REG better preserves the ring and brightness distribution in these examples.

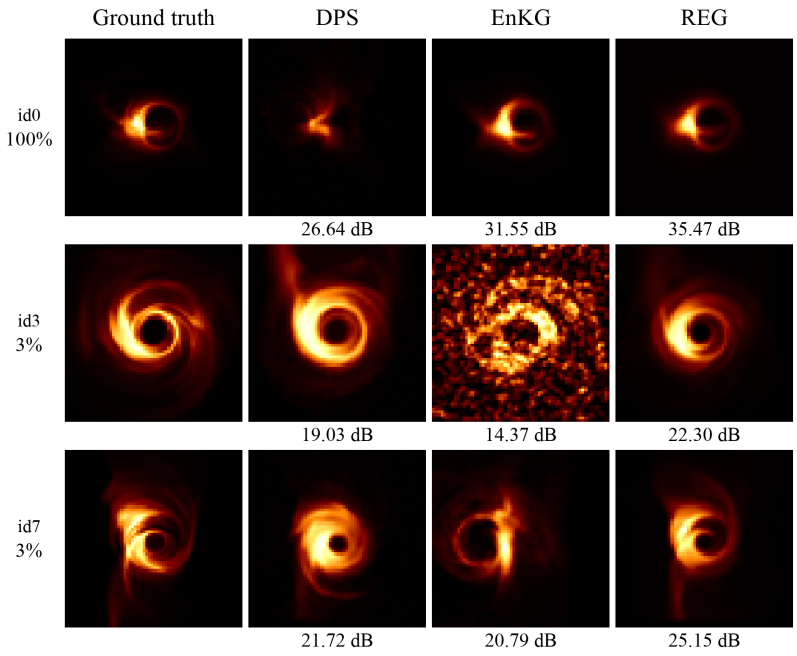


Figure 7: Selected black-hole reconstructions: id0 at 100%, id3 at 3%, and id7 at 3% observation time. Numbers are aligned PSNR (dB); evaluation follows Appendix B.3.

Figure 8 presents six further targets at two observation times. REG recovers the main ring and bright regions across different morphologies, while the intricate outer spiral patterns are less fully resolved.

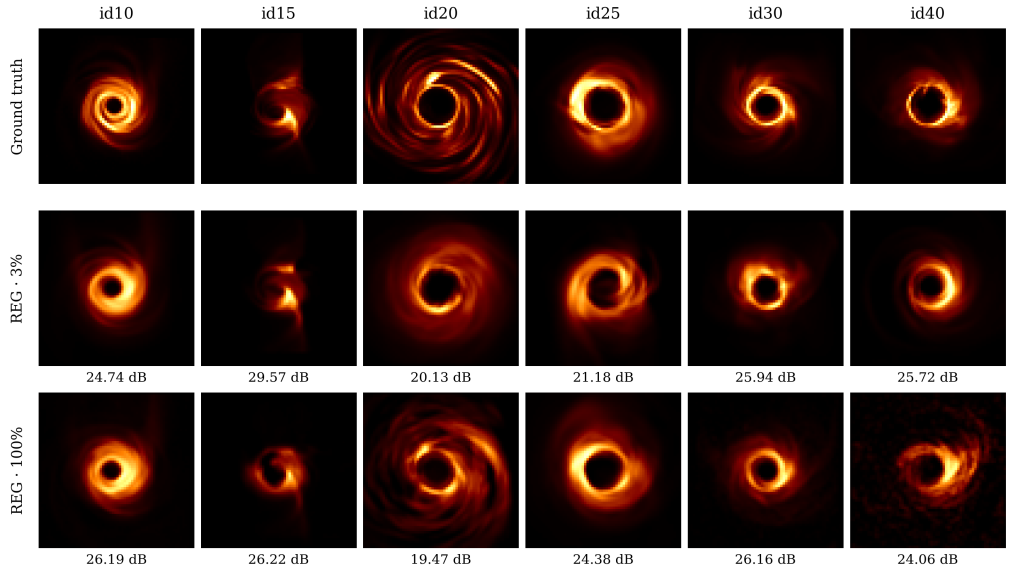


Figure 8: Additional black-hole targets and REG reconstructions at 3% and 100% observation time, with $J = 256$. Numbers are aligned PSNR (dB).

C.3 ACOUSTIC FULL-WAVEFORM INVERSION

Figure 9 compares EnKG and REG on three velocity fields. REG reduces reconstruction error in each case and better matches the broad velocity variations. Fine interfaces remain smoothed in both methods.

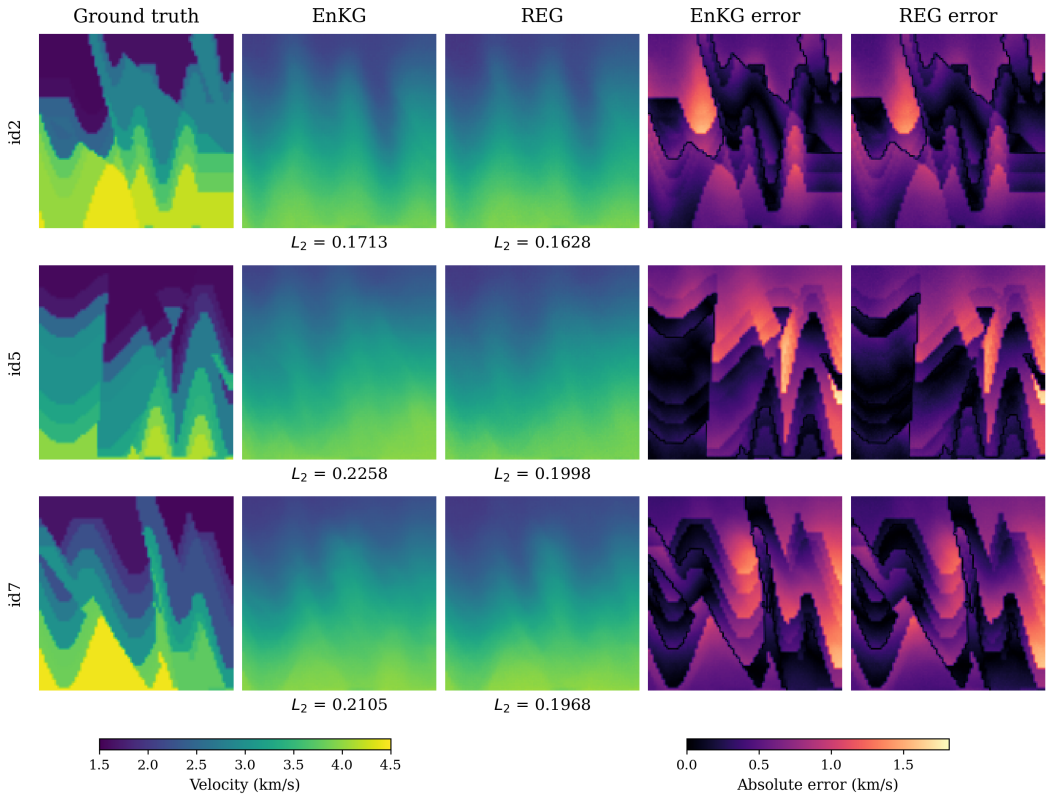


Figure 9: FWI reconstructions and absolute errors for id2, id5, and id7, with $J = 64$. Velocity and error panels use shared color scales; numbers are relative ℓ_2 errors.