ALIGNING EXACT-LIKELIHOOD NORMALIZING FLOWS WITH PHYSICS-AWARE SCORE DENOISING

Anonymous authors
Paper under double-blind review

ABSTRACT

Normalizing flows offer expressive conditional generation and exact likelihood evaluation, making them attractive for scientific field reconstruction. However, practical reconstruction often relies on score-denoised outputs, whose interpretation depends on the density learned during training. Incorporating physical constraints changes this density, creating a potential mismatch between the training objective and the denoising rule. We introduce PAIM-Flow, a framework for aligning exact-likelihood normalizing flows with physics-aware score denoising. PAIM-Flow combines maximum likelihood with physical risk evaluated on fields generated through the inverse flow, while reserving denoising for inference. At the conditional population level, we characterize the density tilt induced by physical supervision and derive an objective- dependent score correction. This analysis clarifies the assumptions required for matched denoising and distinguishes the exact likelihood of the underlying flow from the distribution of denoised outputs. Experiments on Darcy pressure reconstruction, conditional Poisson problems, and sparse turbulent field reconstruction demonstrate substantial improvements in reconstruction accuracy over the evaluated baselines. Controlled ablations disentangle the effects of conditional modeling, physical supervision, and inference-time denoising, separating overall reconstruction gains from the contribution of score correction. Our work provides a principled foundation for connecting physics-aware training objectives with score-based inference in exact-likelihood generative models.

1 INTRODUCTION

Scientific reconstruction requires a field that agrees with observations and physical structure. Given permeability in a Darcy problem, for example, a pressure reconstruction must recover spatial variation while respecting sources and boundary conditions. Data fidelity and physical residual measure distinct properties of that output. Physics-informed learning, neural operators, and generative models provide complementary approaches to this problem (Raissi et al., 2019; Karniadakis et al., 2021; Lu et al., 2021; Li et al., 2021; Bastek et al., 2025). Normalizing flows offer an invertible generative map and explicit density evaluation (Rezende & Mohamed, 2015; Papamakarios et al., 2021; Noé et al., 2019). TARFlow combines these properties with a Transformer autoregressive backbone and a practical Gaussian-noise training and score-denoising recipe(Zhai et al., 2025). These capabilities motivate our choice without assuming a speed or quality advantage over diffusion.

Adding a physical training objective changes the model used by that recipe. Maximum likelihood fits a Gaussian-corrupted data density, whereas an additional physical penalty can favor a different density. The input score remains the correct gradient of the learned model, but its interpretation as the reference Gaussian-corruption score no longer follows automatically. Reusing the original denoising formula can consequently produce a different estimator. Figure 1 isolates this training-to-inference matching question.

We focus on a setting where score denoising remains entirely outside training. The model receives no supervised reconstruction loss on a denoised estimate. Instead, physical supervision acts on model-generated raw fields, and inference accounts for the resulting score change. This differs from training a score-derived estimator by reconstruction or physical losses, as well as from simply

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.

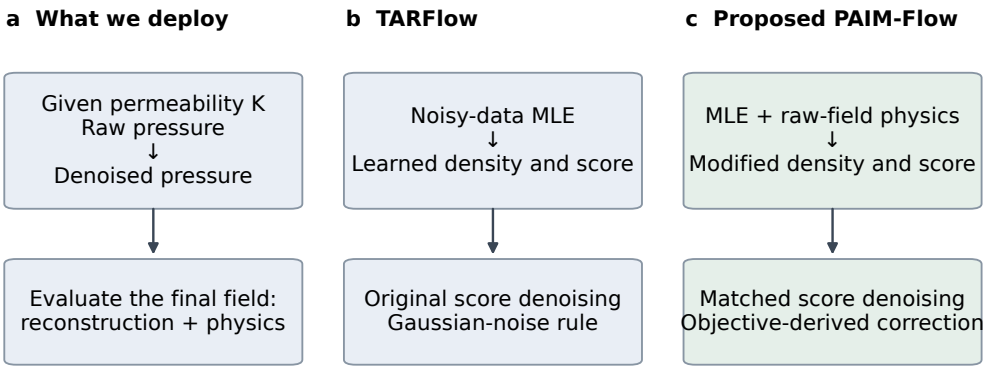


Figure 1: Research question: physical training changes the density underlying post-hoc score denoising. Both TARFlow and PAIM-Flow use score-based inference; PAIM-Flow studies a matching correction for its physical training objective. This schematic does not assert an empirical gain.

appending a PDE-descent step to a generated field. The question is not whether any physics penalty necessarily helps, but which inference correction is justified by the objective.

Two obstacles make the distinction consequential. First, noisy forward-KL maximum likelihood plus expected physical risk does not generally induce the exponential tilt often assumed in guidance arguments. A correction based on that assumption can have the wrong magnitude or sign. Second, even a correct score identity does not identify the distribution of the final denoised samples with the original flow density or guarantee improvement against paired ground truth. The density optimum, estimator, and evaluation distribution must be specified separately.

We formulate PAIM-Flow using a bounded physical risk on raw samples and derive its conditional population optimum. The optimum has a reciprocal tilt, giving an objective-dependent correction to the learned input score. A per-condition normalization quantity can be estimated from independent raw draws without ground truth. The procedure keeps TARFlow's invertible architecture and exact raw-density calculation. Its computational cost, finite-model error, and task-level reconstruction benefit require measurement; differentiable inverse generation during physical training is not free.

The contribution is threefold: an explicit physical-training and inference-matching formulation (Section 3); conditional score and denoising-risk derivations with their assumptions (Appendix A); and a compact evaluation separating task performance, matching effects, and training dynamics (Section 4). The pilot tables contain only runs of this definition. Earlier estimator-supervision experiments are preserved in the research archive and are not relabelled as evidence for the proposed method.

2 Related Work

Normalizing flows provide tractable densities through invertible transformations, including coupling, autoregressive, spline and continuous constructions (Dinh et al., 2015; 2017; Kingma et al., 2016; Papamakarios et al., 2017; Kingma & Dhariwal, 2018; Durkan et al., 2019; Grathwohl et al., 2019; Ho et al., 2019; Tabak & Vanden-Eijnden, 2010). TARFlow motivates the discrete-flow backbone(Zhai et al., 2025); flow matching is a related but distinct transport-learning framework(Lipman et al., 2023). Scientific flows have been used for physical inference and equilibrium sampling (Guo et al., 2022; Rehman et al., 2025; Noé et al., 2019).

Tweedie's identity and denoising score matching connect corrupted densities with clean estimates (Efron, 2011; Vincent, 2011); neural empirical Bayes trains energy-derived denoisers(Saremi & Hyvärinen, 2019). Diffusion uses noise-dependent score or denoising models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021; Nichol & Dhariwal, 2021). Guidance, posterior sampling

and refinement modify inference trajectories (Dhariwal & Nichol, 2021; Ho & Salimans, 2022; Chung et al., 2023; Lippe et al., 2023; Huang et al., 2024; Jacobsen et al., 2025). Our proposed matching correction is tied to a specified training objective, not assumed to be a generic clean posterior score.

Physical constraints enter neural solvers, operators and generative models (Cuomo et al., 2022; Li et al., 2024; Kovachki et al., 2023; Yang & Perdikaris, 2019; Mo et al., 2022; Schindler et al., 2026; Bastek et al., 2025; Zeng et al., 2026). REPA, REPA-P and reverse representation alignment supervise intermediate representations (Yu et al., 2025; Jia et al., 2026; Chen et al., 2026). PAIM-Flow here has no intermediate projection head and does not treat a low residual as proof of physical understanding(Geirhos et al., 2020). Weak-form flow fine-tuning and negative-Sobolev operator refinement motivate investigating residual geometry rather than only its scalar weight (Tauberschmidt et al., 2026; Cao et al., 2025). Score mismatch also arises in conditional guidance and iterative noise schedules (Chao et al., 2022; Zhong et al., 2026); these mechanisms differ from the fixed-noise conditional density deformation studied here.

3 Physical Training and Matched Score Denoising

Let x denote a clean field, c its condition, and $u = x + \sigma\epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$. The corrupted data density is $q_\sigma(u \mid c)$. For an invertible conditional flow f_θ,

$$\log p_\theta(u \mid c) = \log \phi(f_\theta(u; c)) + \log |\det J_{f_\theta}(u; c)|, \qquad s_\theta = \nabla_u \log p_\theta. \tag{1}$$

Equation 1 defines the raw density and its score. A raw sample is $u_\theta = f_\theta^{-1}(z; c)$ with $z \sim \mathcal{N}(0, I)$. Throughout, the fixed noise scale is not a diffusion timestep.

3.1 Training without a denoising branch

Let $R(u, c) \geq 0$ be the task's native physical training energy, including its declared boundary and constraint weights. Choose $\kappa > 0$ using TRAIN-only scale statistics and set $r(u, c) = R(u, c)/(\kappa + R(u, c))$. The bounded risk preserves the minimizers of R while limiting its magnitude. For $0 \leq \lambda < 1$, the proposed training objective is

$$L(\theta) = \mathbb{E}_{c,u\sim q_\sigma}[-\log p_\theta(u \mid c)] + \lambda \mathbb{E}_{c,z}[r(f_\theta^{-1}(z; c), c)]. \tag{2}$$

No denoised estimator, score-matching target or reconstruction error enters Equation 2. Physics gradients reach the backbone through inverse generation and the residual operator. Raw high-frequency noise can conflict with the residual; the bounded objective makes that distributional change explicit rather than assuming it disappears. The likelihood term is a full per-example NLL, averaged over examples. If an implementation divides NLL by the number of coordinates D, its physical coefficient must be λ/D to preserve this objective. Sparse scheduling changes the effective objective and must be compensated or recorded; the initial proposed protocol evaluates both terms every update.

Figure 2 distinguishes the two parameter-gradient training paths from the input-gradient inference path.

3.2 Objective-dependent inference matching

Under the population assumptions derived in Appendix A, the optimum is $p_\lambda = q_\sigma/(a(c) + \lambda r)$, with a positive condition-dependent normalizer $a(c)$. This motivates the matching score and estimator

$$s_{\text{match}}(u, c) = s_\theta(u, c) + \frac{\lambda \nabla_u r(u, c)}{\widehat{a}(c) + \lambda r(u, c)}, \tag{3}$$

$$D_{\text{match}}(u, c) = u + \sigma^2 s_{\text{match}}(u, c). \tag{4}$$

Here $\widehat{a}(c) = 1 - \lambda M^{-1} \sum_{m=1}^{M} r(u^{(m)}, c)$ uses an independent bank of raw draws from the frozen model. The bank size M is a validation-selected computational parameter shared across comparisons where applicable. It does not use a clean test field. Equations 3–4 use a positive physical gradient to compensate the reciprocal tilt, not the negative gradient of ordinary PDE descent. Consequently residual decrease is not guaranteed by the correction itself. For a finite trained flow, the population score identity is approximate and must be assessed empirically.

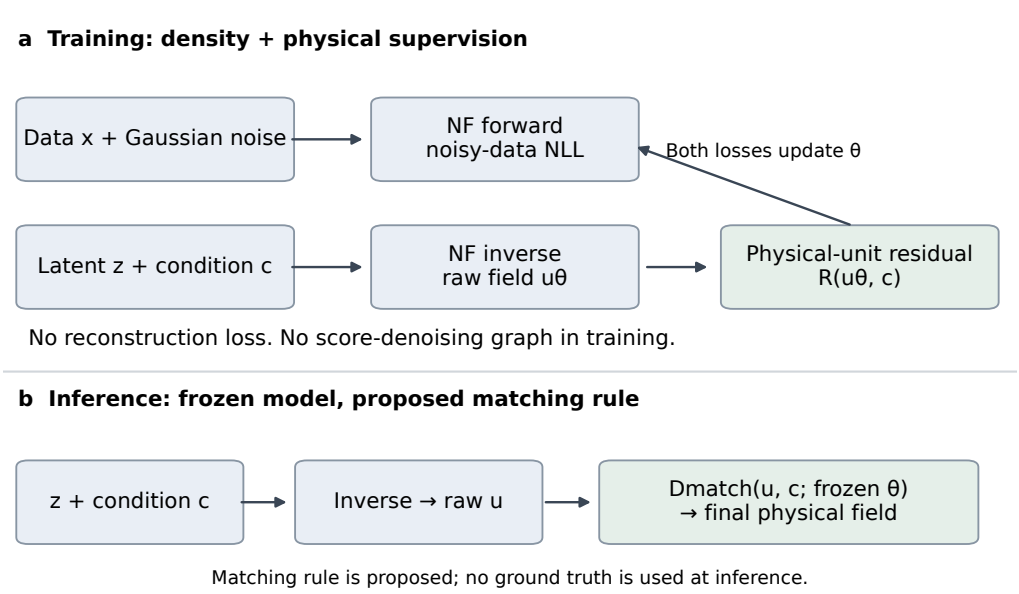


Figure 2: Proposed PAIM-Flow. Training combines noisy MLE and bounded physical risk on inverse samples. Score evaluation and matching occur only at inference with fixed model parameters. The two paths are implemented separately; no denoised estimator enters the training loss.

3.3 Baseline and inference controls

TARFlow trains only the noisy likelihood and uses $D_0(u, c) = u + \sigma^2 s_\theta(u, c)$. PAIM-Flow trains Equation 2 and uses the proposed matching estimator. An unmatched inference ablation evaluates D_0 on the very same PAIM-Flow checkpoint and raw fields. These are protocols of two model families, not extra unnamed model classes. At $\lambda = 0$, PAIM's objective and matching formula reduce to the baseline. Neither denoised output is assigned the raw flow likelihood. Preservation of invertibility does not imply unchanged NLL quality, calibrated uncertainty, or a performance guarantee.

3.4 What matching changes, and what it leaves in place

Physical training changes both the distribution of raw fields and the score evaluated on those fields. The correction compensates the second effect while retaining the first: the raw input still comes from the physics-trained flow. Consequently, recovering the corrupted-data score does not imply that the final sample law equals either the original noisy law or the clean data law. The same clean estimator can yield different output distributions under different inputs. This separates objective matching from simply undoing every effect of physical training.

The theoretical comparison is against unmatched denoising of the same ideal physics-trained density. Under the joint law specified in Appendix A, conditional-mean orthogonality yields

$$\mathcal{E}(D_{\text{ordinary}}) - \mathcal{E}(D_{\text{match}}) = \sigma^4 \lambda^2 \mathbb{E}_{p_\lambda} \frac{\|\nabla r\|^2}{(a + \lambda r)^2} \geq 0. \tag{5}$$

Equation 5 does not compare against a separately trained TARFlow checkpoint on paired benchmark targets. In particular, when the conditional clean solution is deterministic, an ideal exactly learned TARFlow already denoises to that solution. The practical question concerns approximation and finite optimization, not strict improvement over an optimal conditional mean. Related posterior-mean and covariance identities likewise require a specified corruption model (Manor & Michaeli, 2024).

3.5 Numerical implementation and physical scaling

The implementation evaluates the noisy likelihood on a minibatch and computes the physical expectation on independently generated raw fields. Microbatch accumulation preserves the physical

sample mean; it does not multiply the coefficient by the field dimension. The flow backbone is shared but the noise generators for MLE corruption, physical samples, reported raw samples and normalizer samples are separate. Training does not evaluate a score-derived reconstruction loss. Diagnostic denoising runs between checkpoints use frozen parameters.

The median raw physical energy on a fixed set of TRAIN conditions at the shared MLE prefix defines κ, which remains fixed during continuation and inference. This avoids calibrating energy units from validation or test targets. At each evaluation condition, a separate bank estimates $a(c)$; sample count, RNG seed and latent hashes are recorded. These extra inverse evaluations are included in inference timing. Likelihood and physical gradients are measured separately before judging their relative influence: small bounded risk values alone are not evidence of negligible parameter updates.

Physical loss geometry can matter as much as its scalar coefficient. Operator preconditioning changes the weighting of error directions (De Ryck et al., 2024); multiobjective generative training can also exhibit distribution–physics conflicts (Baldan et al., 2026). The present experiment keeps a fixed scalar objective and a fixed energy in both stages. Gradient surgery, a changed residual norm at inference, or an unrecorded coefficient schedule would require a new matching derivation and are not silently incorporated into this protocol.

4 EXPERIMENTS

A staged evaluation with explicit controls. We separate the ability to reconstruct a physical field from the benefit of the matching correction. A first campaign used reduced backbones and three training seeds. A recovery campaign restored stronger conditional fits before comparing physical training and denoising. Every comparison pairs TARFlow and PAIM-Flow within an architecture and initialization. The two campaigns have different budgets and are not pooled into a single multi-seed estimate. All endpoint protocols and unsuccessful architectures remain in the appendix.

Darcy reconstructs pressure from full permeability on a 64×64 grid. Poisson reconstructs potential from two independently signed point charges on the same grid with known zero Dirichlet boundaries. Turbulent reconstruction predicts a scalar 128×48 velocity-fluctuation slice from 30% observations and their mask. Its wall/smoothness criterion is a proxy, not a closed Navier–Stokes residual. The TRAIN/VALID counts are 10000/1000, 200000/2048 and 9000/1000, respectively. No evaluation target enters generation or numerical correction.

4.1 RECOVERING CONDITIONAL RECONSTRUCTION QUALITY

The first pilot used width 256, four flow blocks and coarse patches. Its weak absolute reconstruction scores cannot establish the capacity limit of TARFlow or explain all matching failures. The recovery campaign uses finer 4×4 patches and an eight-block width 384 backbone for Darcy, with 12000 shared MLE updates followed by 1000 matched updates per model. Turbulence restores a trained 25000-update TARFlow checkpoint with 8×4 patches, eight blocks and width 256, then adds 1000 matched updates. Optimizer states are freshly initialized at a weights-only warm start and shared between the two subsequent branches.

For Poisson we retain three controlled recovery configurations: width 384 with cross-attention, width 384 with concatenated conditions, and the earlier width 768 concatenated backbone. The medium variants use 16000 shared updates and 1000 continuation updates; the full-capacity run planned 5000 and 300. Architecture selection uses the shared-MLE checkpoint's validation RMSE before either continuation, never the difference between PAIM-Flow and TARFlow. The two inferior architectures were stopped adaptively before completed physical endpoints. Their checkpoint diagnostics and resource failures remain reported; absent results are not filled.

Table 1 reports the selected recovery configurations at seed 42. The full native validation sets are evaluated, with common latent seed 1042 within each pair. Bold denotes the best measured value in that pair, not significance or state of the art. The appendix includes PSNR, raw fields, ordinary PAIM-Flow denoising and all Poisson architectures. The new table must not be compared causally with historical checkpoints that used different architecture, data exposure or training objectives.

The restored turbulent baseline gave RMSE 0.02224 and SSIM 0.91990 before further training, compared with RMSE 0.21297 in the reduced pilot. Thus the previous poor visual quality was

Table 1: Recovery validation results, training seed42. Each task compares identical model architecture and starting weights; PAIM-Flow uses matched denoising and TARFlow ordinary denoising. These are single-seed validation results, not SOTA claims. Bold marks the smaller/larger measured value within the pair only.

Task	Model	RMSE ↓	SSIM ↑	RMAE ↓
Darcy	TARFlow	0.00649624	0.9272932	**16.02491**
	PAIM-Flow	**0.006495043**	**0.9273098**	16.03952
Poisson	TARFlow	0.00413499	**0.9923106**	9.224214
	PAIM-Flow	**0.004134638**	0.9922939	**9.223116**
Turbulent	TARFlow	**0.02218148**	**0.9205203**	0.002673981
	PAIM-Flow	0.02218671	0.9204745	**0.002673904**

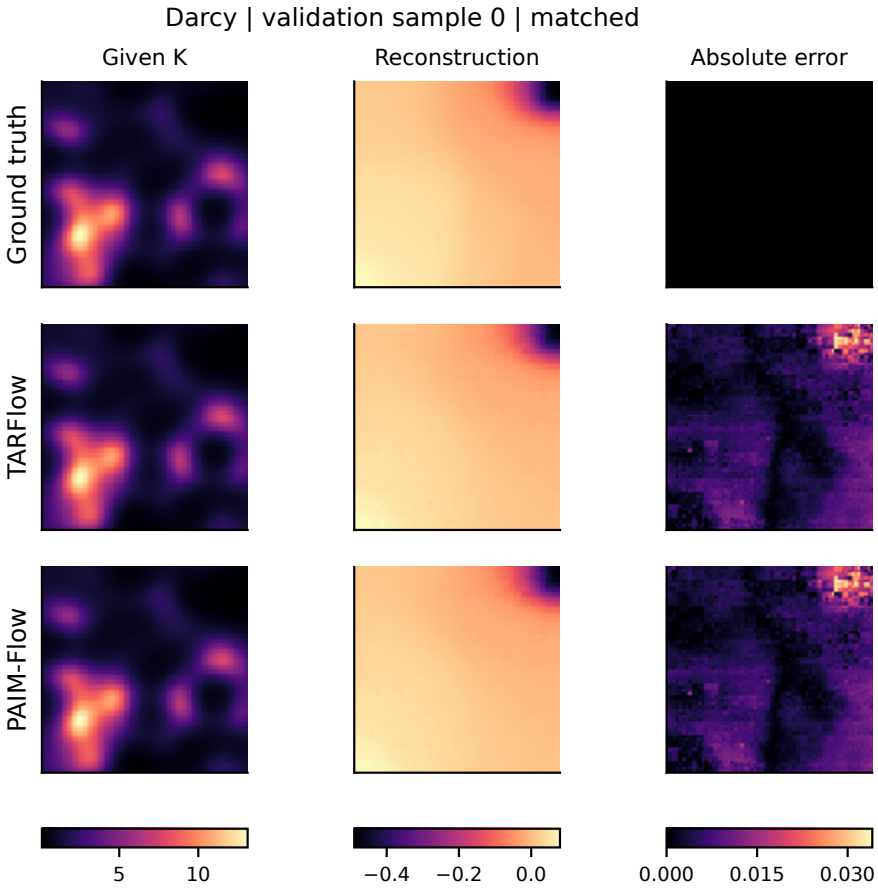


Figure 3: Recovered Darcy reconstruction, fixed validation sample 0. Rows show ground truth, TARFlow ordinary denoising and PAIM-Flow matched denoising. All field and error scales are shared across rows. Strong absolute reconstruction does not imply a visible or statistically meaningful advantage between the two methods.

Table 2: Recovered Darcy, 1000 validation examples, seed 42. PAIM-Flow rows share weights and raw draws. Matching has a minute RMSE effect and slightly worsens physical RMAE.

Model	Output	RMSE ↓	RMAE ↓
TARFlow	Ordinary denoising	0.006496240	**16.02491**
PAIM-Flow	Ordinary denoising	0.006495049	16.02540
PAIM-Flow	Matched denoising	**0.006495043**	16.03952

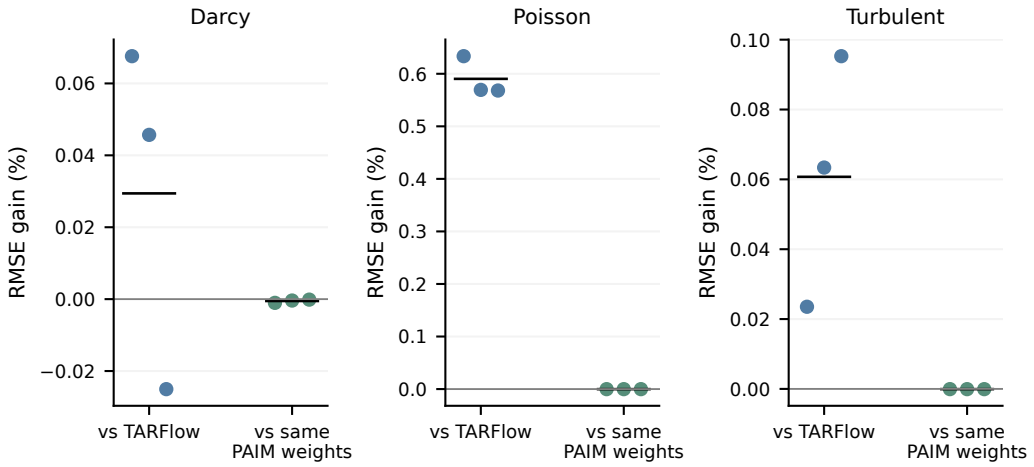


Figure 4: Earlier reduced-backbone pilots, retained as a separate mechanism check. Each dot is a training seed; positive values favor matched PAIM-Flow. The first contrast uses TARFlow and the second uses ordinary denoising of the same PAIM-Flow weights. These are not replications of the larger recovery models in Table 1.

not evidence that the task or data had changed. A data audit independently matched the Poisson validation mean, standard deviation and an archived ground-truth array exactly. Conditioning also matters: at update 8000, medium Poisson concat gave diagnostic RMSE 0.01284, whereas the cross-attention variant gave 0.08489 on the same 32 conditions. These diagnostic values are not substituted for full-validation endpoints.

4.2 Matching on identical checkpoint outputs

For a PAIM-Flow checkpoint, ordinary and matched denoising operate on identical saved raw fields. This comparison isolates the additional inference formula from the training objective. Raw-field evaluation separates both from inverse generation. Table 2 shows the recovered Darcy case without assigning different model names to these protocols.

On turbulent recovery, ordinary TARFlow reached RMSE 0.02218148, compared with 0.02218661 for ordinary PAIM-Flow and 0.02218671 for matched PAIM-Flow. The difference is small, but recovering a strong conditional model did not reverse the ordering. The earlier three-seed reduced campaign showed slightly worse matched than ordinary PAIM-Flow RMSE in all nine endpoints. Figure 4 retains those replications separately, preventing single-seed recovery results from being mistaken for new multi-seed confirmation.

The current evidence therefore distinguishes a functioning physical-gradient path from a demonstrated matching benefit. Exact population score recovery does not establish finite-checkpoint accuracy. Nor does a small mean improvement against another trained model identify which component caused it. Training-seed uncertainty, samplewise pairing and the same-checkpoint control answer different questions and are reported separately.

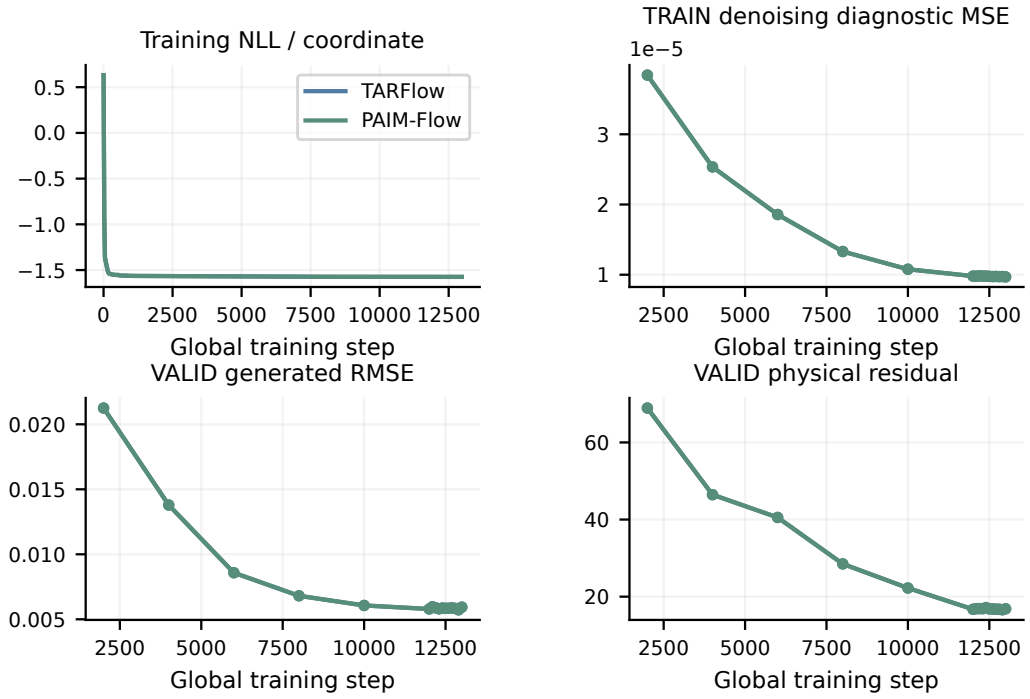


Figure 5: Actual recovered Darcy training diagnostics, seed 42. NLL is displayed as 25-update means; other points are measured checkpoint evaluations. Generated panels use 32 fixed validation conditions. Endpoint tables use all 1000 examples.

4.3 Training diagnostics and explicit consistency controls

Checkpoint diagnostics record noisy TRAIN NLL, frozen-parameter denoising error on TRAIN corruptions, generated VALID RMSE and native physical residual. The diagnostic denoising error is not a training loss. For Poisson cross-attention at update 4000, generated RMSE was 0.09090 while denoising a corrupted true field gave 0.00351 on the same 32 conditions. Shuffling conditions increased generated error to 0.14963. The model used its condition, but accurate denoising near real data did not imply accurate conditional generation.

An oracle direction audit asks whether rescaling the saved matching displacement could help. Writing $e = D_{\text{ordinary}} - x$ and $d = D_{\text{match}} - D_{\text{ordinary}}$, the best unrestricted coefficient is $\alpha^* = -\langle e, d\rangle / \|d\|^2$. Even selecting a separate coefficient with ground truth for every example reduces mean RMSE by at most 0.0049% for Darcy, 0.0373% for Poisson and 0.0120% for turbulence. These are diagnostic upper bounds, not deployable model results. Darcy has 94.6% of error energy in the first 8×8 orthonormal DCT modes, but only 0.000115% of correction energy there. For these checkpoints, coefficient tuning cannot repair the direction mismatch.

We additionally test a separately labelled quadratic consistency operator on the final denoised field. It combines an anchor to the model output with the declared physical risk and, for turbulence, observed-value consistency. The same operator and validation-selected strength are applied to both model families. On 1000 turbulent validation examples it reduced TARFlow RMSE to 0.01690066 and PAIM-Flow RMSE to 0.01690317, with SSIM about 0.9541. This is useful reconstruction improvement without an independent PAIM training advantage. The strength was selected on 64 validation examples; suffix results excluding those examples are retained and are not called a fresh test.

For Poisson, an exact spectral version enforces known zero Dirichlet boundaries. As its strength grows, it approaches the numerical PDE solution and discards the NF anchor. Such near-ground-truth output cannot validate a superior learned score. Appendix L derives these

controls and their limits. The Darcy conjugate-gradient control is iteration-limited; its residual and stationarity diagnostics are reported without claiming an exact solve.

These controls separate three outcomes: stronger conditional fitting, useful numerical postprocessing and an objective-matched denoiser that adds value beyond either. Only the first two currently have clear practical support. The matching hypothesis remains an empirical question despite the population identity. A substantial advantage requires additional evidence against tuned baselines and a fresh test protocol.

5 Discussion and Conclusion

Physical training and denoising are distinct operations. Our proposed objective makes their density relationship explicit and motivates a correction at inference rather than a denoising loss during training. The conditional population analysis supports algebraic matching and a same-distribution risk statement, not unconditional superiority over TARFlow. Finite model capacity, optimization, normalization estimation and a shifted clean marginal remain material limitations. Differentiable raw inverse training and the inference calibration bank add costs that must be measured. Turbulent wall/smoothness priors offer weaker evidence than a closed PDE. The pilot and recovery evaluations separate conditional reconstruction ability from the additional inference correction. Recovery substantially improves absolute quality, but does not establish a substantial PAIM-Flow advantage. The results identify a gap between an exact population identity and its influence in a finite conditional flow. A larger benchmark claim requires new evidence, not a stronger interpretation of these comparisons.

Two rival explanations remain open. The bounded penalty may be too weak relative to full-field likelihood to change the relevant low-frequency errors. Finite physical continuations may also remain far from the population optimum. Restored baseline quality alone does not resolve these issues. The experiments distinguish them from a missing-gradient bug without identifying a universally successful objective. A larger coefficient cannot be inserted into the denoiser alone: it would cease to match the training objective. A different residual norm must likewise be used consistently in training and inference.

The exploratory mean and residual-law objectives preserve the intended Gaussian corruption at a feasible population optimum, but their short Poisson probes also fail to improve reconstruction. Their gradient variance and solver-derived supervision deserve further analysis before expansion. A physical criterion appropriate to the actual turbulent coordinates is also needed. These requirements make the remaining empirical question explicit rather than assuming that exact likelihood, physical training and a matching identity together imply a reconstruction improvement.

Reproducibility statement

The appendix gives the population objective, inference assumptions, discrete tasks and full metric reductions. The executed pilots retain resolved configurations, data hashes, checkpoints, raw and denoised fields, random seeds and cloud-rendered figures. Historical estimator-supervision results are archived separately and do not populate the new method's tables.

AI use statement

AI tools assisted implementation planning, mathematical exposition, figure assembly and writing. Authors remain responsible for verifying all proofs, citations and experimental claims.

REFERENCES

Giacomo Baldan, Qiang Liu, Alberto Guardone, and Nils Thuerey. Physics vs distributions: Pareto optimal flow matching with physics constraints. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a57483b394a3654f4317051e4ce3b2b8-Abstract-Conference.html.

Jan-Hendrik Bastek, WaiChing Sun, and Dennis M. Kochmann. Physics-informed diffusion models. In *International Conference on Learning Representations*, 2025. URL https://iclr.cc/virtual/2025/poster/28020.

Shuhao Cao, Francesco Brarda, Rui Peng Li, and Yuanzhe Xi. Spectral-refiner: Accurate fine-tuning of spatiotemporal fourier neural operator for turbulent flows. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/972eea97af7fe9a7588f5af136edef0c-Abstract-Conference.html.

Chen-Hao Chao, Wei-Fang Sun, Bo-Wun Cheng, Yi-Chen Lo, Chia-Che Chang, Yu-Lun Liu, Yu-Lin Chang, Chia-Ping Chen, and Chun-Yi Lee. Denoising likelihood score matching for conditional score-based data generation. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2203.14206.

Yang Chen, Xiaowei Xu, Shuai Wang, Chenhui Zhu, Ruxue Wen, Xubin Li, Tiezheng Ge, and Limin Wang. Flowing backwards: Improving normalizing flows via reverse representation alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3074–3082, 2026. doi: 10.1609/aaai.v40i4.37300.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.14687.

Salvatore Cuomo, Vincenzo Schiano Di Cola, Fabio Giampaolo, Gianluigi Rozza, Maziar Raissi, and Francesco Piccialli. Scientific machine learning through physics-informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3):88, 2022.

Tim De Ryck, Florent Bonnet, Siddhartha Mishra, and Emmanuel de Bézenac. An operator preconditioning perspective on training in physics-informed machine learning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f122a6778999d1d2b25cdc6927930f1a-Abstract-Conference.html.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear independent components estimation. In *International Conference on Learning Representations Workshop*, 2015.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations*, 2017. URL https://iclr.cc/archive/www/doku.php%3Fid=iclr2017:conference_posters.html.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2019. URL https://iclr.cc/virtual/2019/poster/778.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

Ling Guo, Hao Wu, and Tao Zhou. Normalizing field flows: Solving forward and inverse stochastic differential equations using physics-informed flow models. *Journal of Computational Physics*, 461:111202, 2022. doi: 10.1016/j.jcp.2022.111202.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2722–2730, 2019.

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

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196.

Jiahe Huang, Guandao Yang, Zichen Wang, and Jeong Joon Park. DiffusionPDE: Generative PDE-solving under partial observation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 130291–130323, 2024. doi: 10.52202/079017-4140. URL https://papers.nips.cc/paper_files/paper/2024/hash/eb3878c1dcbfff9ee95d5d033e5f5942-Abstract-Conference.html.

Samuel Hurault, Arthur Leclaire, and Nicolas Papadakis. Gradient step denoiser for convergent plug-and-play. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2110.03220.

Christian Jacobsen, Yilin Zhuang, and Karthik Duraisamy. CoCoGen: Physically consistent and conditioned score-based generative models for forward and inverse problems. *SIAM Journal on Scientific Computing*, 47(2):C399–C425, 2025. doi: 10.1137/24M1636071.

Haozhe Jia, Pengyu Yin, Wenshuo Chen, Shaofeng Liang, Lei Wang, Bowen Tian, Xiucheng Wang, Nanqian Jia, and Yutao Yue. Learning to think in physics: Breaking shortcut learning in scientific diffusion via representation alignment. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/66236.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440, 2021.

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/d139db6a236200b21cc7f752979132d0-Abstract.html.

Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://papers.nips.cc/paper_files/paper/2016/hash/ddeebdeefdb7e7e7a697e1c3e3d8ef54-Abstract.html.

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

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL https://iclr.cc/virtual/2021/poster/3281.

Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM / IMS Journal of Data Science*, 1(3):1–27, 2024. doi: 10.1145/3648506.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://iclr.cc/virtual/2023/poster/11309.

Phillip Lippe, Bas Veeling, Paris Perdikaris, Richard Turner, and Johannes Brandstetter. PDE-Refiner: Achieving accurate long rollouts with neural PDE solvers. In *Advances in Neural Information Processing Systems*, volume 36, pp. 67398–67433, 2023. doi: 10.52202/075280-2946. URL https://papers.nips.cc/paper_files/paper/2023/hash/d529b943af3dba734f8a7d49efcb6d09-Abstract-Conference.html.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3:218–229, 2021. doi: 10.1038/s42256-021-00302-5.

Hila Manor and Tomer Michaeli. On the posterior distribution in denoising: Application to uncertainty quantification. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/d6adf82fa531dcb8bfa53c224ce5e466-Abstract-Conference.html.

Zhaobin Mo, Yongjie Fu, Daran Xu, and Xuan Di. TrafficFlowGAN: Physics-informed flow based generative adversarial network for uncertainty quantification. In *Machine Learning and Knowledge Discovery in Databases*, pp. 323–339. Springer, 2022. doi: 10.1007/978-3-031-26409-2_20.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171, 2021.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019. doi: 10.1126/science.aaw1147.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

Danyal Rehman, Oscar Davis, Jiarui Lu, Jian Tang, Michael Bronstein, Yoshua Bengio, Alexander Tong, and Avishek Joey Bose. Efficient regression-based training of normalizing flows for boltzmann generators. *arXiv preprint arXiv:2506.01158*, 2025.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538, 2015.

Saeed Saremi and Aapo Hyvärinen. Neural empirical bayes. *Journal of Machine Learning Research*, 20(181):1–23, 2019. URL https://jmlr.org/papers/v20/19-216.html.

Benjamin Schindler, Melle Mendikowski, and Thomas Schmid. Bridging physical constraints and deep generative models via physics-aware normalizing flows. *Neurocomputing*, 661:131880, 2026. doi: 10.1016/j.neucom.2025.131880.

Rajesh Shrestha and Xiao Fu. Taming score-based denoisers in ADMM: A convergent plug-and-play framework. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/16e18fa3b3add076c30f2a2598f03031-Paper-Conference.pdf.

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*, volume 37 of *Proceedings of Machine Learning Research*, 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. URL https://iclr.cc/virtual/2021/poster/3177.

Esteban G. Tabak and Eric Vanden-Eijnden. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010.

Jan Tauberschmidt, Sophie Fellenz, Sebastian J. Vollmer, and Andrew B. Duncan. Physics-constrained fine-tuning of flow-matching models for generation and inverse problems. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e8b782d03c623e3e587fdbcc521dd581-Abstract-Conference.html.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

Yibo Yang and Paris Perdikaris. Adversarial uncertainty quantification in physics-informed neural networks. *Journal of Computational Physics*, 394:136–152, 2019. doi: 10.1016/j.jcp.2019.05.027.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

Tianyi Zeng, Tianyi Wang, Jiaru Zhang, Zimo Zeng, Feiyang Zhang, Yiming Xu, Sikai Chen, Junfeng Jiao, Christian Claudel, and Xinbo Chen. PILD: Physics-informed learning via diffusion. *arXiv preprint arXiv:2601.21284*, 2026.

Shuangfei Zhai, Ruixiang Zhang, Preetum Nakkiran, David Berthelot, Jiatao Gu, Huangjie Zheng, Tianrong Chen, Miguel Ángel Bautista, Navdeep Jaitly, and Joshua M. Susskind. Normalizing flows are capable generative models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74348–74369, 2025. URL https://proceedings.mlr.press/v267/zhai25d.html.

Hongkai Zheng, Wenda Chu, Bingliang Zhang, Zihui Wu, Austin Wang, Berthy T. Feng, Caifeng Zou, Yu Sun, Nikola Kovachki, Zachary E. Ross, Katherine L. Bouman, and Yisong Yue. Inversebench: Benchmarking plug-and-play diffusion priors for inverse problems in physical sciences. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e297068f5ec7f483a017948715b591fb-Abstract-Conference.html.

Jincheng Zhong, Boyuan Jiang, Xin Tao, Pengfei Wan, Kun Gai, and Mingsheng Long. Mitigating the noise shift for denoising generative models via noise awareness guidance. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/39b6b296c5d19b9bbf67a5320ff6117f-Abstract-Conference.html.

A A CONDITIONAL THEORY OF PHYSICS-MATCHED POSTPROCESSING

This appendix specifies a concrete candidate for physics training followed by inference-only score correction. It does not reinterpret experiments with denoiser-supervised training as results for this candidate. The identities below hold at an unrestricted population optimum; their finite-model approximation and benchmark benefits require new experiments.

A.1 VARIABLES, UNITS, AND TRAINING OBJECTIVE

Fix a condition c. Let $q_0(x \mid c)$ denote the clean field distribution and $q_\sigma(u \mid c) = \int q_0(x \mid c)\phi_\sigma(u - x)\, dx$ its Gaussian convolution. Here $u, x \in \mathbb{R}^D$ are in the same model coordinates and ϕ_σ has covariance $\sigma^2 I$. The physical energy $R(u, c) \geq 0$ is evaluated after decoding to physical units, including the appropriate boundary and observation terms when available. All derivatives with respect to u include the chain rule through this decoding. Define the dimensionless bounded energy

$$r(u, c) = \frac{R(u, c)}{\kappa + R(u, c)}, \qquad \kappa > 0, \qquad 0 \leq r < 1. \tag{6}$$

The energy scale κ is fixed using training data only. Boundedness is a sufficient condition for the normalization result below; it also limits the influence of extreme residuals. Saturation can weaken gradients at extreme states and must therefore be checked empirically.

For $0 \leq \lambda < 1$, consider

$$J[p] = \mathbb{E}_{q_\sigma}[-\log p(u \mid c)] + \lambda \mathbb{E}_p[r(u, c)], \qquad \int p(u \mid c)\, du = 1. \tag{7}$$

For an NF, the second expectation is computed on raw generated fields $u_\theta = F_\theta^{-1}(z; c)$ with $z \sim \mathcal{N}(0, I)$, and its gradient must pass through u_θ. There is no reconstruction loss and no score denoising inside this training graph. This objective can be expensive for an autoregressive inverse; a detached generated sample would not implement its pathwise physics gradient.

Equation equation 7 uses a full-field NLL in nats and a scalar dimensionless energy per example, followed by a batch mean. If the implementation averages NLL over D coordinates, the equivalent coefficient of r is λ/D. Auxiliary-update frequency must also be accounted for in the expected objective. In particular, the scalar λ here is not an arbitrary legacy coefficient multiplying an unnormalized PDE loss.

A.2 THE OPTIMUM IS A RATIONAL TILT, NOT AN EXPONENTIAL TILT

Proposition 1 (population density). Assume $q_\sigma > 0$, and optimize over normalized positive densities with finite objective. Then the unique minimizer is

$$p_\lambda(u \mid c) = \frac{q_\sigma(u \mid c)}{a(c) + \lambda r(u, c)}, \qquad \int \frac{q_\sigma(u \mid c)}{a(c) + \lambda r(u, c)}\, du = 1, \tag{8}$$

where $1 - \lambda \leq a(c) \leq 1$.

Proof. The first variation of the constrained objective is $-q_\sigma(u)/p(u) + \lambda r(u) + a = 0$, which gives the stated density. The normalization function is continuous and strictly decreasing in a. At $a = 1 - \lambda$ its integrand denominator is at most one, so its value is at least one. At $a = 1$ it is at most one. Thus a unique root exists in the given interval. Finally, $-\int q_\sigma \log p$ is strictly convex in p on the support of q_σ, while the energy term is linear, so this stationary point is the unique global minimizer. The case $\lambda = 0$ gives $a = 1$ and $p_0 = q_\sigma$. □

An exponential tilt follows from a different objective, such as an appropriate reverse KL plus expected energy. It does not follow from the forward-KL MLE objective in Equation equation 7.

A.3 MATCHED SCORE AND A COMPUTABLE NORMALIZER

Assume r is differentiable. Taking the input gradient of Equation equation 8 gives

$$s_\lambda(u,c) = \nabla_u \log q_\sigma(u \mid c) - \frac{\lambda \nabla_u r(u,c)}{a(c) + \lambda r(u,c)}, \tag{9}$$

$$s_{\text{match}}(u,c) := s_\lambda(u,c) + \frac{\lambda \nabla_u r(u,c)}{a(c) + \lambda r(u,c)} = \nabla_u \log q_\sigma(u \mid c). \tag{10}$$

Thus the matching term has a positive sign: it compensates for the density deformation induced by training. It is not an additional downhill physics step, and it need not lower the residual of the denoised field. Integrating $q_\sigma = p_\lambda(a + \lambda r)$ yields

$$a(c) = 1 - \lambda \mathbb{E}_{p_\lambda(\cdot|c)} r(U,c), \qquad \widehat{a}_\theta(c) = 1 - \frac{\lambda}{M} \sum_{m=1}^{M} r(U_m, c), \quad U_m \sim p_\theta(\cdot \mid c). \tag{11}$$

The estimate uses independent calibration samples for each condition and requires no evaluation ground truth. Conditions cannot be pooled unless a separate approximation is declared. A fixed condition-indexed calibration bank would avoid dependence on evaluation batch size. The executed pilot instead fixes batch size and draw order and records latent hashes; changing batching changes the bank assignment. At the population optimum, boundedness gives $\Pr(|\widehat{a} - a| \geq \delta) \leq 2\exp(-2M\delta^2/\lambda^2)$ for $\lambda > 0$. Additional inverse samples are a real inference cost.

The implementable candidate is

$$D_{\text{match},\theta}(u,c) = u + \sigma^2 \left[\nabla_u \log p_\theta(u \mid c) + \frac{\lambda \nabla_u r(u,c)}{\widehat{a}_\theta(c) + \lambda r(u,c)} \right]. \tag{12}$$

The NF term must be a full log-density gradient per example, not a batch-mean or coordinate-mean gradient unless the missing scale is restored. The variance σ^2 is in model coordinates. Nonisotropic Gaussian corruption requires its covariance matrix instead.

A.4 WHICH POSTERIOR DOES THE MATCHED ESTIMATOR ESTIMATE?

Tweedie's identity follows by differentiating the Gaussian convolution:

$$\nabla_u \log q_\sigma(u \mid c) = \frac{\mathbb{E}_q[X \mid U = u, c] - u}{\sigma^2}. \tag{13}$$

Consequently, exact score correction recovers the original conditional denoising mean. However, raw samples now come from p_λ, not q_σ. To state the resulting inference problem, define the normalized joint law

$$\pi_\lambda(x, u \mid c) = \frac{q_0(x \mid c)\phi_\sigma(u - x)}{a(c) + \lambda r(u,c)}. \tag{14}$$

Its u marginal is p_λ, while $\pi_\lambda(x \mid u, c) = q(x \mid u, c)$ because the tilt depends only on u. Therefore $D_{\text{match}}(u,c)$ is the posterior mean for this joint law. Its clean marginal is

$$\pi_\lambda(x \mid c) = q_0(x \mid c)\mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)} \left[\frac{1}{a(c) + \lambda r(x + \sigma\epsilon, c)} \right]. \tag{15}$$

This is not generally $q_0(x \mid c)\exp[-\beta R(x,c)]$ and is not generally the original clean distribution. Matching the score does not undo the change in the sampling marginal. Nor is a posterior mean a sample from that clean marginal.

Proposition 2 (comparison with unmatched postprocessing). Assume finite second moments. Under the joint law in Equation equation 14, let $D_{\text{naive}}(u,c) = u + \sigma^2 s_\lambda(u,c)$. Then

$$\begin{aligned} &\mathbb{E}_{\pi_\lambda} \|X - D_{\text{naive}}(U,c)\|^2 - \mathbb{E}_{\pi_\lambda} \|X - D_{\text{match}}(U,c)\|^2 \\ &\quad = \sigma^4 \lambda^2 \mathbb{E}_{p_\lambda} \frac{\|\nabla_u r(U,c)\|^2}{[a(c) + \lambda r(U,c)]^2} \geq 0. \end{aligned} \tag{16}$$

Proof. For any estimator $g(U)$, conditional-mean orthogonality gives $\mathbb{E}\|X - g(U)\|^2 = \mathbb{E}\|X - \mathbb{E}[X \mid U]\|^2 + \mathbb{E}\|g(U) - \mathbb{E}[X \mid U]\|^2$. Substitute Equation equation 9. □

The proposition compares two postprocessors of the same ideal physics-trained density. It does not establish dominance over TARFlow, which has a different raw marginal, or over an externally paired benchmark ground truth. In a deterministic conditional problem with an exactly learned Gaussian-corrupted target, TARFlow's exact conditional denoiser already returns the true solution; strict improvement over that ideal estimator is impossible.

A.5 Physical bias and finite-model limitations

At the population optimum, lower-energy raw states receive greater relative weight. In particular,

$$\mathbb{E}_{p_\lambda} r - \mathbb{E}_{q_\sigma} r = \mathrm{Cov}_{q_\sigma}\left(r, \frac{1}{a + \lambda r}\right) \leq 0, \tag{17}$$

since the second argument is a decreasing function of the first. This only guarantees a decrease in the bounded training energy on raw states, not in unbounded residuals, clean-field RMSE, or energy after denoising.

Let $e_\theta = \nabla \log p_\theta - \nabla \log p_\lambda$ and use the same r, λ in training and inference. Pointwise,

$$\|D_{\mathrm{match},\theta} - D_q\| \leq \sigma^2 \left(\|e_\theta\| + \frac{\lambda \|\nabla r\| \, |\widehat{a}_\theta - a|}{(1-\lambda)^2} \right). \tag{18}$$

Finite capacity, incomplete optimization, and estimating a under an approximate p_θ all contribute error. Training loss convergence alone does not bound e_θ. The denominator bound deteriorates as λ approaches one. Changing energy, units, or effective loss coefficient between training and inference invalidates the matching identity. Finally, inverse sampling must remain differentiable during physics training; a cheap detached physics diagnostic is not an equivalent training objective.

Falsifiable empirical claims. The proposed advantages are retention of the exact NF parameterization, an explicit correction derived from a specified training objective, and an idealized reduction of denoising risk relative to unmatched postprocessing. Improved benchmark reconstruction, lower clean physical residual, favorable runtime, and superiority to TARFlow remain empirical hypotheses. They require matched checkpoints, identical raw samples, multiple training seeds, and separate reporting of calibration cost and raw versus denoised metrics.

B Benchmark Definitions and Complete Metrics

B.1 Data, units and physical operators

Darcy uses positive K and pressure on a boundary-inclusive unit-square grid, $h = 1/63$, with opposite 8×8 source/sink blocks of $+10$ and -10. The reference solves the stacked PDE, zero normal-derivative boundary, and trapezoidal pressure-gauge equations in least squares. Its residual floor can be nonzero. The expanded discrete PDE is

$$r_\Omega = -K\Delta_h p - (D_x K)(D_x p) - (D_y K)(D_y p) - f. \tag{19}$$

The solver-consistent risk preserves relative row weights: $R = \mathrm{mean}(r_\Omega^2) + (256/4096)\,\mathrm{mean}(r_\partial^2) + g(p)^2/4096$. Physical training must record this exact convention; do not compare reweighted objectives as if they were identical reference problems.

Poisson uses two independently signed charges with amplitudes in $[0.5, 1.5]$; coincident charges accumulate. Stored charge amplitudes ρ correspond to source ρ/h^2. The interior residual is $\Delta_h U + \rho/h^2$, with $U = 0$ on the boundary. Neither this task nor Darcy normalizes state coordinates. Turbulent velocity uses a TRAIN-only max-absolute scale; its single-component data do not support a closed PDE residual. The wall proxy penalizes the first wall-normal index; the opposite edge follows the reference's centerline interpretation. The stored first column is not exactly zero, so the proxy is not asserted to be an exact feasible boundary condition of these arrays.

B.2 FIELD ERRORS

For N examples, let $x_i, \widehat{x}_i \in \mathbb{R}^D$ be target and prediction in the declared evaluation coordinates. Define $e_i = \widehat{x}_i - x_i$. We report

$$\mathrm{MSE} = \frac{1}{N}\sum_i \frac{\|e_i\|_2^2}{D}, \qquad \mathrm{RMSE} = \frac{1}{N}\sum_i \sqrt{\frac{\|e_i\|_2^2}{D}}, \tag{20}$$

$$\mathrm{NMSE} = \frac{1}{N}\sum_i \frac{\|e_i\|_2^2}{\max(\|x_i\|_2^2, \varepsilon)}. \tag{21}$$

Mean per-example RMSE is not the square root of dataset MSE. With target range $L_i = \max(\max_j x_{ij} - \min_j x_{ij}, \varepsilon)$ and $m_i = \|e_i\|^2/D$,

$$\mathrm{PSNR} = \frac{1}{N}\sum_i \left[20\log_{10} L_i - 10\log_{10}\max(m_i, \varepsilon)\right]. \tag{22}$$

The implementation uses the target tensor dtype epsilon (float32 in native saved bundles). Ranges are sample-specific, never inferred from predictions. PSNR is in decibels.

B.3 STRUCTURAL SIMILARITY

For a uniform 7×7 window W of $n = 49$ pixels, compute

$$\mu_x = \frac{1}{n}\sum_{j\in W} x_j, \qquad \mu_y = \frac{1}{n}\sum_{j\in W} \widehat{x}_j, \tag{23}$$

$$v_x = \frac{1}{n-1}\sum_{j\in W}(x_j - \mu_x)^2, \qquad v_y = \frac{1}{n-1}\sum_{j\in W}(\widehat{x}_j - \mu_y)^2, \tag{24}$$

$$v_{xy} = \frac{1}{n-1}\sum_{j\in W}(x_j - \mu_x)(\widehat{x}_j - \mu_y). \tag{25}$$

For $C_1 = (0.01L_i)^2$ and $C_2 = (0.03L_i)^2$,

$$\mathrm{SSIM}_{i,W} = \frac{(2\mu_x\mu_y + C_1)(2v_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(v_x + v_y + C_2)}, \qquad \mathrm{SSIM} = \frac{1}{N}\sum_i \frac{1}{|\mathcal{W}_i|}\sum_{W\in\mathcal{W}_i} \mathrm{SSIM}_{i,W}. \tag{26}$$

This matches uniform-window scikit-image SSIM with sample covariance, valid window-center averaging, no Gaussian weighting and arithmetic channel averaging where applicable. For small grids use the largest odd window at most seven, requiring at least three pixels per axis. Native constant-target handling assigns one to an exactly identical field and otherwise marks SSIM undefined; report undefined counts rather than hiding excluded cases. The saved-field sidecar's epsilon-clamped constant-range policy is distinct and must not silently replace it.

B.4 PHYSICAL RESIDUALS

Let $r_{\Omega,i}$ and $r_{\partial,i}$ contain native equation and boundary residuals. The project's RMAE is *residual mean absolute error*, not relative percentage error:

$$\mathrm{RMAE} = \frac{1}{N}\sum_i \frac{\|r_{\Omega,i}\|_1 + \|r_{\partial,i}\|_1}{n_\Omega + n_\partial}. \tag{27}$$

Gauge/constraint penalties are recorded separately. Additional diagnostics are

$$\mathrm{PDE\ MAE} = \frac{1}{N}\sum_i \frac{\|r_{\Omega,i}\|_1}{n_\Omega}, \quad \mathrm{PDE\ MSE} = \frac{1}{N}\sum_i \frac{\|r_{\Omega,i}\|_2^2}{n_\Omega}, \quad \mathrm{BC\ MSE} = \frac{1}{N}\sum_i \frac{\|r_{\partial,i}\|_2^2}{n_\partial}. \tag{28}$$

For turbulent scalar slices, replace RMAE by the declared proxy

$$\mathrm{RMAE}_{\mathrm{proxy}} = \frac{1}{N}\sum_i \frac{\|\sqrt{w_b}\, u_i'|_{y=0}\|_1 + \|\sqrt{w_s}\, \Delta_h u_i'|_{\mathrm{interior}}\|_1}{n_b + n_s}. \tag{29}$$

The existing convention is $w_b = 0.1$, $w_s = 0.01$, unit index spacing for smoothness and TRAIN-normalized velocity. No Navier–Stokes interpretation is assigned to this number.

B.5 Sparse errors, likelihood and uncertainty

For binary observation mask M_i, unobserved error is

$$\mathrm{RMSE}_{\mathrm{unobs}} = \frac{1}{N}\sum_i \sqrt{\frac{\|(1-M_i)\odot e_i\|_2^2}{\sum_j (1-M_{ij})}}. \tag{30}$$

Observed RMSE replaces $1-M_i$ with M_i. The legacy masked-L2 metric divides the Euclidean norm by the count, not its square root, and must not be labelled RMSE. Raw density metrics are

$$\mathrm{NLL} = -\frac{1}{N}\sum_i \log p_\theta(x_i \mid c_i), \qquad \mathrm{BPD} = \mathrm{NLL}/(D\log 2). \tag{31}$$

These continuous density values may be negative. Clean-target and noisy-target NLL use different evaluation inputs and must be labelled; neither is the density of the denoised output. Report mean and sample standard deviation across independently trained seeds. Paired example bootstrap intervals resample identical indices for both methods; they do not replace seed-level uncertainty. PSNR/RMAE aggregate per example before across-seed aggregation.

C Exact Conditional Density and the Computational Graph

C.1 An autoregressive density with a conditional transformer

The flow maps a field to independent Gaussian coordinates through invertible blocks. Patchification first rearranges the D field coordinates into T tokens of width d, where $Td = D$. This operation does not compress the field. Its inverse is a permutation, and the absolute Jacobian determinant is one. Conditioning on a permeability or charge field does not include that condition in the generated state or its density dimension.

A scalar-coordinate notation makes the autoregressive dependency explicit. Within a block, consider the affine transformation

$$z_j = \{u_j - \mu_j(u_{<j}, c)\}\exp\{-\alpha_j(u_{<j}, c)\}. \tag{32}$$

The conditioning transformer can be arbitrarily nonlinear in its permitted context. Invertibility follows from strict autoregressive dependence and finite scales, rather than from invertibility of that transformer. The diagonal Jacobian entry is $\partial z_j/\partial u_j = \exp(-\alpha_j)$, giving

$$\log|\det J| = -\sum_j \alpha_j. \tag{33}$$

In a patch-autoregressive implementation, each patch has elementwise scales and shifts conditioned on preceding patches. The same argument applies with block-triangular Jacobians. A shifted input sequence or an equivalent strict causal mask is essential. Allowing the scale or shift to depend on the current state token would invalidate Equation equation 33 unless the additional derivative terms were included.

Composing invertible blocks adds their log determinants. Token permutations between blocks preserve the absolute determinant and let subsequent blocks use different contexts. The conditional density is therefore

$$\log p_\theta(u \mid c) = -\tfrac{1}{2}\|f_\theta(u;c)\|_2^2 - \tfrac{D}{2}\log(2\pi) + \sum_{\ell=1}^{L}\log|\det J_{f_{\theta,\ell}}|. \tag{34}$$

These are the density principles underlying autoregressive normalizing flows and the TARFlow backbone (Papamakarios et al., 2017; Zhai et al., 2025). The condition is held fixed when evaluating the Jacobian and input score. A condition encoder may be shared across blocks without adding a state-density determinant, provided it depends only on the supplied condition.

C.2 Why likelihood and generated-field training have different costs

The likelihood direction observes the full input field, so all permitted transformer contexts are available in parallel. The inverse instead solves

$$u_j = \mu_j(u_{<j}, c) + \exp\{\alpha_j(u_{<j}, c)\} z_j \tag{35}$$

in autoregressive order. A differentiable physical loss on the complete generated field must retain the dependencies through this sequence. Two GPUs do not remove the sequential dependence within a sample. Data parallelism distributes samples or microbatches, while memory-saving recomputation trades additional evaluations for fewer saved activations.

The physical branch in the proposed formulation differentiates

$$\nabla_\theta r(u_\theta, c) = \left(\frac{\partial u_\theta}{\partial \theta}\right)^\top \nabla_u r(u_\theta, c), \qquad u_\theta = f_\theta^{-1}(z; c). \tag{36}$$

The sampled z is independent of the trainable parameters. The model dependence comes from its inverse map. A detached raw sample produces a valid residual diagnostic but eliminates this pathwise gradient. Likewise, evaluating a physical loss on the training target alone does not constrain the network parameters. Evaluating it on an exact round trip $f_\theta^{-1}(f_\theta(x; c); c) = x$ also removes the intended dependence: the derivative of this identity with respect to θ, at fixed x, c, is zero.

At inference, the parameters are frozen but the raw input must support differentiation. The score requires $\nabla_u \log p_\theta(u \mid c)$, not a parameter update. Constructing this input gradient inside a context that disables all differentiation would therefore be incorrect. No higher-order parameter derivative of the score is needed by the inference-only matching formulation. This distinguishes it from the archived approach that differentiated through a denoised estimator during training.

C.3 Tensor shapes and gradient ownership

For a batch size B, generated state tensors have shape $[B, 1, H, W]$ in all three reported tasks. Darcy and Poisson conditions have shape $[B, 1, H, W]$; the sparse velocity condition has shape $[B, 2, H, W]$. The latter concatenates observed values and their Boolean mask, represented in the model's input dtype. The state dimension is $D = HW$, not the sum of state and condition coordinates. Tokens have shape $[B, T, d]$, while residuals are flattened separately for each example before scalar reduction.

During noisy MLE, the corrupted training state is data, and parameters receive gradients through the forward flow. During physical training, the generated state retains its inverse graph. During inference matching, the raw field is a leaf with respect to input differentiation; no optimizer is applied. An independent calibration bank estimates a scalar normalizer for each condition and is held fixed when taking the query-state gradient. Differentiating the bank as a function of the query state would add a term absent from the conditional population identity.

D Gaussian Denoising Identities and Their Scope

D.1 A complete conditional Tweedie derivation

Fix a condition c and assume that the clean law has finite second moments. For Gaussian corruption with covariance $\Sigma \succ 0$, define

$$q_\Sigma(u \mid c) = \int \phi_\Sigma(u - x)\, q_0(dx \mid c), \qquad U = X + \eta, \quad \eta \sim \mathcal{N}(0, \Sigma). \tag{37}$$

The clean law may be singular, including a deterministic solution for a fixed condition. Its Gaussian convolution is a smooth positive density. Assuming the required integrability for exchanging differentiation and integration,

$$\nabla_u q_\Sigma(u \mid c) = \int \Sigma^{-1}(x - u)\phi_\Sigma(u - x) q_0(dx \mid c), \tag{38}$$

$$\nabla_u \log q_\Sigma(u \mid c) = \Sigma^{-1}\{\mathbb{E}[X \mid U = u, c] - u\}. \tag{39}$$

Consequently,

$$m(u,c) := \mathbb{E}[X \mid U = u, c] = u + \Sigma \nabla_u \log q_\Sigma(u \mid c). \tag{40}$$

The scalar update used by TARFlow corresponds to $\Sigma = \sigma^2 I$. The variance is the variance of the corruption in the same coordinates as the score. It is neither an arbitrary step size nor a diffusion-time index. A separately tuned multiplier changes the estimator and must be recorded as a distinct protocol. The empirical-Bayes interpretation of this identity is classical (Efron, 2011).

To establish its squared-error optimality, let $g(U, c)$ be any square-integrable estimator. Expanding around $m(U, c)$ gives

$$\mathbb{E}\|X - g(U,c)\|_2^2 = \mathbb{E}\|X - m(U,c)\|_2^2 + \mathbb{E}\|m(U,c) - g(U,c)\|_2^2, \tag{41}$$

because the conditional expectation of $X - m(U, c)$ is zero. This proof specifies the relevant joint distribution; it does not identify all other reconstruction metrics with squared error. A posterior mean need not maximize SSIM, minimize an absolute residual, or produce a sample distributed according to the clean law.

For a deterministic conditional target $q_0(dx \mid c) = \delta_{x_c}(dx)$, $q_\sigma = \mathcal{N}(x_c, \sigma^2 I)$ and the exact score is $(x_c - u)/\sigma^2$. Equation equation 40 then returns x_c for every u. This limiting case shows why strict superiority over an ideal, exactly learned TARFlow cannot be a general theorem. Practical improvements must concern finite data, finite capacity, optimization, or a specified inference budget.

D.2 Mean, uncertainty, and generated-field diversity

The posterior covariance also has a derivative representation. Differentiating the conditional mean under isotropic Gaussian corruption yields

$$\mathrm{Cov}(X \mid U = u, c) = \sigma^2 J_u m(u,c) = \sigma^2 I + \sigma^4 \nabla_u^2 \log q_\sigma(u \mid c). \tag{42}$$

For example, differentiation of the normalized conditional density gives $\partial m_i / \partial u_j = \mathrm{Cov}(X_i, X_j \mid u, c)/\sigma^2$. This connects local denoiser derivatives with uncertainty, as developed further by Manor & Michaeli (2024). For an exact Gaussian-corruption model, the right-hand side is positive semidefinite. A negative eigenvalue computed from an approximate learned score is evidence that this exact posterior-covariance interpretation does not hold there; it is not a negative physical variance.

Different latent draws followed by deterministic denoising produce a distribution of outputs. That distribution is not automatically the posterior distribution of the clean field. An ensemble mean, a single draw, and the covariance above answer different questions. We therefore distinguish reconstruction quality, raw density quality, and uncertainty calibration rather than assigning the tractable raw likelihood to the final denoised field.

D.3 Why density fit alone does not bound score error

The original appendix's density-versus-score distinction remains relevant even when no score loss is used in training. Let $q_\sigma = \mathcal{N}(0, v)$ in one dimension and define, for integers $n \geq 2$,

$$p_n(u) = q_\sigma(u)\{1 + n^{-1}\sin(nu)\}. \tag{43}$$

This density is positive and normalized because the Gaussian is symmetric and the sine term is odd. It has an exact one-dimensional flow representation: $f_n = \Phi^{-1} \circ F_n$, where F_n and Φ are its CDF and the standard Gaussian CDF. A uniform Taylor bound for $-\log(1 + t)$, $|t| \leq 1/2$, gives

$$\mathrm{KL}(q_\sigma \| p_n) = O(n^{-2}). \tag{44}$$

However, its score discrepancy is

$$\partial_u \log p_n - \partial_u \log q_\sigma = \frac{\cos(nu)}{1 + n^{-1}\sin(nu)}. \tag{45}$$

The denominator converges uniformly to one, while $\mathbb{E}_{q_\sigma} \cos^2(nU) = (1 + \exp(-2n^2 v))/2$. Thus the squared $L^2(q_\sigma)$ score discrepancy converges to $1/2$, despite vanishing forward KL. The corresponding Tweedie estimators have limiting mean squared disagreement $\sigma^4/2$.

This construction does not show that MLE is inconsistent. At the correctly specified population optimum it recovers the target density and score. It shows that a small finite density-fit error, by itself, is insufficient to certify small input-gradient error. The same limitation applies to claiming that a nearly converged physics-regularized training loss certifies exact matching.

E What the Matching Correction Preserves and Changes

E.1 A local expansion of the population tilt

The rational-tilt result in Appendix A provides a useful small-λ diagnostic. Let $\mu_r = \mathbb{E}_{q_\sigma} r$ for a fixed condition. Differentiating the normalization equation at $\lambda = 0$ gives $a'(0) = -\mu_r$. Therefore,

$$a(\lambda) = 1 - \lambda\mu_r + O(\lambda^2), \tag{46}$$

$$\frac{p_\lambda(u)}{q_\sigma(u)} = 1 - \lambda\{r(u) - \mu_r\} + O(\lambda^2), \tag{47}$$

$$s_\lambda(u) - s_q(u) = -\lambda\nabla r(u) + O(\lambda^2). \tag{48}$$

The expansion is conditional on a fixed smooth energy and suitable derivative bounds. It explains why a positive correction is required to undo the local score deformation. A negative residual-gradient step would reinforce it.

The exact density ratio also satisfies

$$\frac{1}{1+\lambda} \le \frac{p_\lambda(u \mid c)}{q_\sigma(u \mid c)} \le \frac{1}{1-\lambda}, \qquad 0 \le \lambda < 1. \tag{49}$$

This follows from $1 - \lambda \le a \le 1$ and $0 \le r < 1$. For small coefficients, the allowed population redistribution is modest. Near one, stronger redistribution is possible but normalizer sensitivity also increases. Neither regime supplies an unconditional reconstruction guarantee.

E.2 Undoing a score deformation does not undo sample selection

Exact matching restores s_q at the queried raw input. The queried inputs, however, are drawn from the physics-trained p_λ rather than q_σ. Consequently the distributions of $D_q(U)$ generally differ under these two raw marginals. Physics can still affect the output distribution through the locations where the common denoiser is evaluated. The correction is therefore not equivalent to replacing the entire physics-trained pipeline by TARFlow.

For a measurable statistic h with integrable expectations,

$$\mathbb{E}_{p_\lambda} h(D_q(U)) = \mathbb{E}_{q_\sigma} \frac{h(D_q(U))}{a + \lambda r(U)}. \tag{50}$$

This expression also identifies the remaining uncertainty. Lower bounded energy of U does not imply a lower error or residual of $D_q(U)$. Such a conclusion would require a relation between the raw training energy and the chosen statistic after denoising. The final benchmarks measure that relation rather than assuming it.

E.3 Finite-model mismatch has two distinct contributions

Let the learned density be p_θ and the ideal density be p_λ. Write $e_s = \nabla \log p_\theta - \nabla \log p_\lambda$. With an estimated normalizer $\widehat{a}$, the matched score error is exactly

$$s_{\mathrm{match},\theta} - s_q = e_s + \lambda\nabla r \left\{ \frac{1}{\widehat{a} + \lambda r} - \frac{1}{a + \lambda r} \right\}. \tag{51}$$

The first term measures the model's score approximation error. The second measures normalizer error. An accurate Monte Carlo bank only reduces the sampling component of the second term; it cannot correct a poorly trained score.

For a bank drawn from p_θ, its normalizer error further decomposes as

$$\widehat{a} - a = -\lambda \left\{ \frac{1}{M} \sum_{m=1}^{M} r(U_m) - \mathbb{E}_{p_\theta} r \right\} - \lambda\{\mathbb{E}_{p_\theta} r - \mathbb{E}_{p_\lambda} r\}. \tag{52}$$

The first bracket has Monte Carlo standard deviation at most $1/(2\sqrt{M})$. The second persists as $M \to \infty$ unless the bounded-energy expectation is correct. Reporting bank-size sensitivity without checking model mismatch can therefore overstate convergence of the matching procedure.

A direct scalar diagnostic is available when independent Gaussian-corrupted training or validation fields can be sampled for the same condition. Define

$$\widehat{H}_c(\widehat{a}) = \frac{1}{J}\sum_{j=1}^{J} \frac{1}{\widehat{a} + \lambda r(V_j, c)}, \qquad V_j \sim q_\sigma(\cdot \mid c). \tag{53}$$

At the exact population normalizer its expectation equals one. This is an offline diagnostic requiring clean paired data; it must not be used to obtain test-time normalizers from hidden test targets. It complements, but does not replace, the target-free calibration bank in deployed inference.

E.4 Why matching is different from posterior guidance

Ordinary physics guidance often decreases an energy containing a prior, observation mismatch, and physical residual. If

$$E(v) = -\log p(v \mid c) + \gamma L_{\text{obs}}(v, y) + \beta R(v, c), \tag{54}$$

then its negative gradient is

$$-\nabla_v E = \nabla_v \log p(v \mid c) - \gamma \nabla_v L_{\text{obs}} - \beta \nabla_v R. \tag{55}$$

The positive physical term in PAIM's matching identity has a different role: it compensates a specified training-induced density deformation. It is not an extra physical likelihood and need not decrease R in one step. Adding Equation equation 55 after matching would define another inference protocol with a separate target and compute cost.

Related inverse-problem work evaluates both observation fit and source accuracy because neither alone certifies a useful reconstruction (Zheng et al., 2025). The same principle applies here. For sparse conditions, the observation mask and measured values are legitimate inference inputs; the remaining ground truth is not. If a learned conditional prior already uses these observations, an additional observation likelihood must be motivated to avoid counting the same information twice under an unjustified posterior interpretation.

F Physical Coordinates, Discretization, and Conditioning

F.1 Coordinate transforms and score units

Let a physical state be $x_{\text{phys}} = b + Sv$, where S is an invertible diagonal scaling matrix fixed from training statistics. The density and score transform as

$$\log p_v(v \mid c) = \log p_{\text{phys}}(b + Sv \mid c) + \log|\det S|, \tag{56}$$

$$s_v(v, c) = S^\top s_{\text{phys}}(b + Sv, c), \tag{57}$$

$$\nabla_v R(b + Sv, c) = S^\top \nabla_{x_{\text{phys}}} R(x_{\text{phys}}, c). \tag{58}$$

Thus an isotropic model-space noise variance σ^2 corresponds to physical covariance $\sigma^2 SS^\top$. Applying a physical-space residual gradient directly to a normalized state, without this chain rule, changes both units and direction. The physics energy and its scale κ must use the same decoding at training and inference.

Darcy and Poisson use the native stored state coordinates in the established benchmark. Turbulent velocity uses a training-only maximum-absolute scale. For Darcy, an optional standardization of $\log K$ changes only the condition features supplied to the network. The physical operator still receives the original positive permeability K. Exponentiating that already positive array would define a different PDE. Conversely, a genuine predicted log permeability would require its declared inverse transform before evaluation.

F.2 Darcy derivatives, boundary rows, and gauge

On a uniform one-dimensional grid with spacing h, interior derivative stencils are

$$(D_h v)_i = \frac{v_{i+1} - v_{i-1}}{2h}, \qquad (D_h^2 v)_i = \frac{v_{i+1} - 2v_i + v_{i-1}}{h^2}. \tag{59}$$

The left endpoint uses

$$(D_h v)_0 = \frac{-3v_0 + 4v_1 - v_2}{2h}, \qquad (D_h^2 v)_0 = \frac{2v_0 - 5v_1 + 4v_2 - v_3}{h^2}, \tag{60}$$

with reflected right-endpoint stencils. Applying these separately to both axes produces the expanded Darcy operator used by the reference generator. The implementation evaluates PDE rows on all 64^2 grid nodes, including boundary nodes. This is not interchangeable with a face-flux discretization on cell centers, even though both approximate a continuum divergence form.

The zero-normal-derivative boundary concatenates four sides, retaining both normal components at corners. Its row count is $4 \times 64 = 256$. The pressure gauge is a normalized tensor-product trapezoidal mean,

$$g(p) = \frac{\sum_{i,j} w_i w_j p_{ij}}{\sum_{i,j} w_i w_j}, \qquad w_0 = w_{63} = \tfrac{1}{2}, \quad w_i = 1 \text{ otherwise.} \tag{61}$$

The gauge eliminates the pressure-offset ambiguity of a Neumann problem. For a reference computed by stacking equation, boundary, and gauge rows, the matching least-squares energy is proportional to the sum of all row squares. Separate means require row-count factors to preserve this target.

An overdetermined discrete system can have nonzero residual at its reference solution. The meaningful operator check is consistency with that reference solver and its residual floor, rather than assuming the stored field exactly annuls every row. A manufactured smooth solution with a corresponding source provides an additional convergence test. It tests differential consistency without conflating the numerical solver's row weighting with a new model.

F.3 Poisson source scaling and spectral conditioning

Stored point charges are amplitudes assigned to grid nodes. Their physical right-hand side is divided by the grid-cell area $h_x h_y$. The interior Poisson equation is evaluated on 62×62 nodes; each Dirichlet boundary value is included once, for 252 distinct boundary coordinates. This differs from Darcy's two normal-derivative rows at a corner and must not share its boundary count by convenience.

For homogeneous Dirichlet boundaries and equal spacing h, the positive operator $L = -\Delta_h$ has sine-basis eigenvalues

$$\nu_{kl} = \frac{4}{h^2}\left[\sin^2\left(\frac{k\pi}{2(n+1)}\right) + \sin^2\left(\frac{l\pi}{2(n+1)}\right)\right], \quad 1 \le k, l \le n, \quad n = 62. \tag{62}$$

For a homogeneous-boundary prediction error e, a squared residual weights its orthonormal sine coefficients by ν_{kl}^2:

$$\|Le\|_2^2 = \sum_{k,l} \nu_{kl}^2 |\widehat{e}_{kl}|^2. \tag{63}$$

Fine-scale errors can consequently dominate the training residual while large-scale errors dominate field MSE. This is a discretization-level explanation, not proof of any particular trained model's failure mechanism. The importance of the operator's Hermitian square and its interaction with the parameterization is analyzed by De Ryck et al. (2024).

A positive-definite residual preconditioner W changes the energy to $R_W = \|W^{1/2}(Lv - f)\|^2$. With $W = L^{-2}$ and homogeneous boundaries, this equals field error when the discrete solution is unique. Implementing such a choice may require a global solve and can remove some of the claimed computational advantage. For nonzero predicted boundary errors, the interior forcing also contains boundary contributions; an interior spectral formula cannot simply ignore them. Any selected preconditioner must be fixed in the training objective and reused inside the matching correction.

F.4 Sparse velocity observations and the scope of its physics prior

The velocity tensor is ordered as $[B, 1, X, Y]$ with $X = 128$, $Y = 48$. The reference wall proxy penalizes $Y = 0$; the opposite edge follows its half-channel interpretation. The stored first-column

targets are not exactly zero, so this is not asserted to be an exactly feasible boundary constraint. The condition is $(M \odot v, M)$, where M marks approximately 30% of spatial sites. Displaying this rectangle as near-square changes only the rendering aspect. The native samples, masks, pixel counts, and metrics remain unchanged.

The available single velocity component does not determine incompressibility, momentum balance, pressure gradients, or time derivatives. The implemented physical prior therefore penalizes the observed wall and nonperiodic smoothness, with optional gradient regularization. Its finite differences use the declared index spacing. The resulting scalar is a wall/smoothness proxy, not a Navier–Stokes equation residual.

Physical plausibility and preservation of turbulent structure can compete: aggressive smoothing lowers this proxy while erasing physically meaningful fluctuations. Reconstruction errors on unobserved sites and spectrum changes are therefore necessary companions to the proxy. Observed-site errors are reported separately. Replacing those pixels by measurements is a clamp protocol and cannot be silently applied to only one model.

G Objective Scaling and Numerical Verification

G.1 Reduction, accumulation, and scheduled physics updates

The population coefficient λ multiplies a bounded scalar energy per example. It is defined relative to a full-field log density in nats. Writing $\ell_i = -\log p_\theta(u_i \mid c_i)$, the minibatch objective is

$$\widehat{L} = \frac{1}{B}\sum_i \ell_i + \frac{\lambda}{B_r}\sum_j r(u_{\theta,j}, c_j). \tag{64}$$

The likelihood and physics batches may have different sizes while estimating the same expectations, provided conditions follow the same declared law. Dividing the first term by D requires dividing the physical coefficient by D too. Dividing a per-example input-gradient objective by B requires restoring that factor before a denoising update. Computing a sum of independent per-example log densities avoids a batch-dependent score step.

If physics is evaluated every k optimizer updates, applying weight $k\lambda$ on those updates matches the time-averaged scalar coefficient in a simple stationary gradient calculation. It does not make an Adam trajectory identical to physics on every update, because moment estimation, clipping, and changing parameters intervene. The schedule and active coefficient must therefore be logged. If the population formula is used for inference, the effective trained objective must be justified explicitly.

Microbatch accumulation weights each partial mean by its number of examples divided by the full effective batch size. Equal weighting of unequal microbatches biases the objective toward the final smaller batch. Distributed reduction likewise uses global example counts. A reference small-batch test should compare gradients from the unpartitioned and accumulated versions.

G.2 Energy saturation and parameter-gradient diagnostics

For $r = R/(\kappa + R)$,

$$\nabla_u r = \frac{\kappa}{(\kappa + R)^2}\nabla_u R. \tag{65}$$

Boundedness prevents arbitrarily large loss values but can suppress gradients at highly unphysical raw states. A useful diagnostic records the distribution of R/κ, r, and $\|\nabla_u r\|$ on generated training states. Calibrating κ using clean fields alone can be misleading if generated raw states have much larger residuals. A training-only raw calibration bank is an alternative, provided its procedure is fixed and shared with inference.

Let g_{NF} and g_r denote separately accumulated parameter gradients before clipping. Report

$$\rho = \frac{\lambda\|g_r\|_2}{\|g_{\mathrm{NF}}\|_2 + \varepsilon}, \qquad \chi = \frac{g_{\mathrm{NF}}^\top g_r}{\|g_{\mathrm{NF}}\|_2\|g_r\|_2 + \varepsilon}. \tag{66}$$

Neither nonzero gradients nor a favorable cosine guarantees better validation reconstruction. They distinguish ineffective scale from conflicting directions and numerical failure. Diagnostics should include early, late, and representative condition batches rather than only the most favorable update.

Conflict-aware updates have been studied for physics-constrained flow matching (Baldan et al., 2026). Importing such updates into this model changes the optimization rule and may remove the interpretation as minimization of the fixed scalar objective. The rational-tilt correction would then require a new derivation or an explicitly approximate justification. Logging conflicts is compatible with the present theory; silently replacing its objective is not.

G.3 Required scalar and field-level numerical checks

The scalar population construction can be tested independently of the transformer. Choose a one-dimensional Gaussian q_σ and a bounded smooth energy, solve the normalizer equation by numerical quadrature, and evaluate both scores. The corrected score must agree with the Gaussian score. This checks the sign and denominator without optimization noise. Replacing λ by zero must recover the ordinary score estimator.

For every field operator, compare autograd directional derivatives with central finite differences in double precision:

$$\nabla R(v)^\top d \approx \frac{R(v+\epsilon d) - R(v-\epsilon d)}{2\epsilon}. \tag{67}$$

Several ϵ values distinguish truncation from floating-point error. The test includes nonzero boundary values so that omitted boundary gradients cannot pass unnoticed. A small negative-gradient step should lower a smooth energy when the chosen step is sufficiently small. This descent check applies to $-\nabla R$, not to the positive score-matching correction.

At the backbone level, the physical gradient must reach representative early and late flow blocks and condition parameters. The inverse/forward round trip and log-determinant tests verify the exact-density claims separately. Checkpoint reload must reproduce raw generation, input scores, and matched outputs for fixed conditions and latents. Parameter freezing at inference must not suppress the input derivative.

H Experimental Controls, Selection, and Reproducibility

H.1 Three experimental questions

The experimental narrative groups measurements by scientific question. First, the three-task comparison measures whether the complete PAIM-Flow pipeline improves useful reconstruction over TARFlow. Second, the matched versus unmatched comparison holds a PAIM-Flow checkpoint and all raw fields fixed, isolating the proposed postprocessor. Third, training and calibration diagnostics test whether the assumptions behind that postprocessor are approximately attained. These are experimental groups, not extra named model families.

The first group reports reconstruction, physical or proxy residual, raw density, and compute. The second evaluates ordinary score denoising and matching on the same raw tensor; changing the latent draw would weaken the comparison. A conventional negative-gradient physical correction is useful as a separately budgeted control when claiming an advantage over generic postprocessing. The third includes loss trends, physical gradient scale, normalizer-bank sensitivity, and paired improvement by condition.

The benchmark structure follows the distinction between source accuracy, observation consistency, and computational effort emphasized in InverseBench (Zheng et al., 2025). Its diffusion priors and scientific inverse problems are not substituted for the tasks here. Its evaluation principles inform the controls; its numerical results are not copied into our tables.

H.2 Split roles and selection criteria

Training data determine normalization statistics, energy-scale calibration, and parameter updates. Validation data select the declared coefficient, checkpoint, and inference settings. Test data assess a frozen choice. Previously inspected test results cannot serve as a fresh independent confirmation of a subsequently selected method. If new independent data are generated, their seeds and generator provenance are frozen before evaluation. The turbulent validation suffix remains labelled as validation unless a genuinely independent split is obtained.

Model selection must reflect the main claim. If the claim is improved reconstruction with physical consistency, selecting solely by the smallest residual can choose an oversmoothed field. A transparent rule either fixes a primary reconstruction metric with a physical admissibility threshold, or reports the validation Pareto frontier before choosing a point. The rule and tolerances must precede inspection of the final test. Selecting a different checkpoint for every metric creates a collection of models, not a single deployable PAIM-Flow result.

For two validation candidates a, b, candidate a Pareto-dominates b when it is no worse in all declared minimized metrics and strictly better in at least one. A frontier visualizes a trade-off rather than proving that either objective is unimportant. The physics-versus-distribution perspective in Baldan et al. (2026) motivates reporting this relationship, while the mechanism and objective remain different in PAIM-Flow.

H.3 Paired uncertainty and independent training seeds

For sample-level metric m, define the paired difference

$$\Delta_i = m(\widehat{x}_i^{\mathrm{PAIM}}, x_i) - m(\widehat{x}_i^{\mathrm{TAR}}, x_i), \qquad \overline{\Delta} = \frac{1}{N}\sum_i \Delta_i. \tag{68}$$

The sign is reversed for metrics where larger is better when reporting an improvement. Paired bootstrap resampling draws the same example indices for both models. Its interval describes uncertainty over examples at the evaluated trained checkpoints. Multiple latent seeds describe sampling variability. Neither replaces independently trained seeds.

Across S independent training seeds, compute the mean and sample standard deviation of each seed's dataset metric. Preserve paired seeds and initialization where the training protocol permits it. For correlated temporal velocity slices, an ordinary example bootstrap may underestimate uncertainty; block resampling with a declared block length is a more suitable sensitivity analysis. Multiple metrics are reported together, without turning each unadjusted interval into a separate significance claim.

H.4 Runtime, memory, and calibration accounting

Inference time separates raw inverse generation, NF input-score evaluation, physical energy differentiation, and normalizer calibration. If a condition is used once, all M calibration inverse draws belong to that query's cost. If it is reused for K outputs and the bank is cached, the amortized cost is

$$T_{\mathrm{output}} = T_{\mathrm{raw}} + T_{\mathrm{score}} + T_{\mathrm{match}} + T_{\mathrm{bank}}/K. \tag{69}$$

Both $K = 1$ and the declared reuse setting are informative. Caching one bank per condition must not silently share estimates across unrelated conditions. Timing includes device synchronization, a warm-up policy, precision, batch size, and hardware. Peak allocated memory and total wall-clock time are reported separately. A smaller number of mathematical update steps alone does not establish lower runtime.

Training comparisons report both optimizer steps and wall-clock/GPU hours. The generated-field physics branch is more expensive than a likelihood pass; matched steps and matched compute therefore answer different questions. A checkpoint initialized from a common TARFlow warm start must record that pretraining cost for both models, followed by each continuation cost.

H.5 A bound on the total distributional intervention

The bounded-risk objective also limits how far its exact optimum can move from the corrupted-data law in forward KL. Since q_σ is a feasible density, $J[p_\lambda] \le J[q_\sigma]$. Subtracting the entropy term gives

$$\mathrm{KL}(q_\sigma \| p_\lambda) + \lambda \mathbb{E}_{p_\lambda} r \le \lambda \mathbb{E}_{q_\sigma} r, \qquad \mathrm{KL}(q_\sigma \| p_\lambda) \le \lambda < 1. \tag{70}$$

For $\lambda > 0$, this also proves that the population optimum cannot increase the expected bounded raw risk relative to q_σ. Neither inequality concerns the residual of a denoised field. The KL is a full-field quantity in nats, not a per-coordinate divergence. This distinction explains why a coefficient

that looks moderate numerically can represent a small global intervention in a high-dimensional likelihood objective.

This bound is not a gradient bound or a guarantee of small score error: the oscillatory example above explicitly separates likelihood and score behavior. It nevertheless constrains interpretation of a coefficient change. Replacing full-field NLL with coordinate-mean NLL while retaining the same numerical physical coefficient multiplies the effective risk weight by D. That would leave the proved $\lambda < 1$ regime for the present grids. Such a change can be a research direction only with a revised objective, existence analysis and inference rule; it cannot be reported as a harmless implementation correction.

H.6 Artifacts and figure integrity

Related plug-and-play analyses separate the denoiser from the data-consistency operator. Gradient-step denoisers can support convergence to stationary points of an explicit functional under stated regularity assumptions (Hurault et al., 2022). More recent score-based ADMM work addresses the mismatch between optimization iterates and noisy states used to train a denoiser (Shrestha & Fu, 2026). Neither result supplies a convergence guarantee for arbitrary repeated TARFlow denoising. A one-step matched estimator, a posterior optimization procedure and a numerical PDE solver must therefore remain separately labelled. Applying an added consistency operator only to PAIM-Flow would confound its benefit with the contribution of physical training. Each result links a resolved configuration, source revision, checkpoint hash, dataset manifest, split, training seed, sampling seed, and inference protocol. The normalizer bank records its own independent seed and hash. Saved bundles retain the target, condition, mask when relevant, raw field, ordinary denoised field, matched field, and per-example metrics. This permits recomputation of paired tables without repeating autoregressive generation.

Optimizer-free inference checkpoints preserve all required model tensors but are not equivalent to resumable training checkpoints. Resumption also requires optimizer, scheduler, and random-generator state. Failed runs and rejected validation candidates remain in the audit trail. Historical denoiser-supervised checkpoints are not relabelled as results of the new inference-only matching method.

Reconstruction figures use fixed example IDs and shared field scales across methods. Absolute-error panels share their own scale, distinct from the physical-field scale. Unobserved pixels are explicitly masked rather than displayed as physical zeros. Poisson figures show ground truth and the two final model outputs; raw comparisons belong to a separate diagnostic if needed. Bold table entries denote the best measured value within a declared comparison, not a claim of state of the art across untested methods.

I Executed pilot protocol and reproducibility

I.1 Dataset ownership and native-grid evaluation

The pilot campaign contains nine training runs: three physical cases and three training seeds. Each run produces one shared-prefix checkpoint and two continuation checkpoints. The shared prefix is an initialization control, not a third comparison model. All reported reconstruction conditions belong to validation, and no independent test was opened in this campaign. This prevents optimization diagnostics from becoming undeclared test-set selection. Darcy uses the repository's 10000 training and 1000 validation fields. The conditional Poisson generator produces 200000 training and 2048 validation pairs, with generation seeds 42 and 1042. Turbulent scalar slices retain the 9000/1000 split, TRAIN-derived normalization and fixed sparse-observation construction.

The validation counts are 1000, 2048 and 1000 for Darcy, Poisson and turbulence. Evaluation is performed on native 64×64, 64×64 and 128×48 arrays. Near-square turbulent panels change the display aspect only; the evaluator never resamples fields. No mask clamp, numerical PDE solver, target-derived initialization or supervised reconstruction head is introduced into the endpoint comparison. The target appears only in metric computation. The complete permeability or charge field is a legitimate condition in the first two tasks; the turbulent condition contains observed values and their binary mask.

I.2 MATCHED MODEL AND OPTIMIZER SETTINGS

Table 3: Executed pilot settings shared by TARFlow and PAIM-Flow unless stated.

Setting	Darcy	Poisson	Turbulent
Native grid	64×64	64×64	128×48
Patch size	8×8	8×8	16×8
Flow blocks / layers per block	4 / 2	4 / 2	4 / 2
Width / head dimension	256 / 64	256 / 64	256 / 64
Trainable parameters	8,761,856	6,647,296	9,139,200
Gaussian noise σ	0.05	0.02	0.05
Condition dropout	0	0	0
MLE prefix updates	3000	3000	5000
Continuation updates per model	300	300	300
MLE batch size	64	64	128
Physics samples/update (PAIM-Flow)	2	2	2
Physics coefficient (PAIM-Flow)	0.8	0.8	0.8
Normalizer draws/condition	8	8	8
Evaluation batch size	32	32	32

The shared prefix uses AdamW with momentum parameters $(0.9, 0.95)$, weight decay 10^{-4} and a cosine learning-rate schedule from 10^{-4} to 10^{-6}. Both continuations restore the same model and optimizer state, then use learning rate 10^{-5} and gradient clipping at norm one. MLE computation uses bfloat16 autocast; differentiable inverse physics uses float32. This mixed-precision policy applies to the actual code path, not to an assumed memory estimate. The physical expectation is accumulated from two independent one-example inverse samples at every continuation update.

Full per-field NLL determines the optimizer's gradient, while displayed NLL curves divide by the coordinate count for readability. The physical term is not accidentally multiplied by the MLE batch size, coordinate count, number of microbatches or number of flow blocks. The two raw physics samples estimate a conditional expectation with considerable variance; their small count is a resource choice and a limitation. Increasing that count could reduce variance without changing the objective, whereas changing the relative gradient by surgery would require a different mathematical analysis.

I.3 CALIBRATION, RANDOMNESS AND CHECKPOINT SELECTION

For each case and training seed, the shared-prefix checkpoint generates a raw field on each of the first 32 TRAIN conditions. The median native physical energy gives κ, which remains fixed through continuation and inference. The calibration file records those indices, its random seed, the measured scale and the hash of the shared checkpoint. Neither model can select its calibration using validation reconstruction quality. The fixed final continuation checkpoint supplies every endpoint protocol and metric.

Private random streams separate training indices, MLE corruption, inverse physics draws, validation latent draws and normalizer draws. The final reconstruction seed is 1042. The normalizer bank uses seed 101042 and the evaluated model's own conditional distribution. Its draw order is part of the protocol: changing evaluation batch size or bank size changes which random values are assigned to a condition. We therefore record batch size, bank size and latent hashes rather than claim batch-invariant Monte Carlo assignments. Ordinary and matched PAIM-Flow denoising use exactly the same saved raw field. TARFlow and PAIM-Flow share the evaluation latent stream, although their different weights produce different raw fields.

The three-seed experiment was a fixed diagnostic replication, not a search for a successful seed. The initial Darcy effect did not meet a 2% improvement threshold for promoting this candidate into a large benchmark campaign. Replications were retained to determine whether the small effect and weak matching contribution persisted. All nine endpoints, including unfavorable comparisons, remain in the released tables. Validation observations motivated interpretation and stopping, not selective removal of model outputs.

I.4 Endpoint audits and limits of timing

A separate frozen-endpoint audit evaluates the first 128 validation conditions with nested normalizer banks of 4, 8 and 16 draws. These are diagnostic subsets, not replacements for the full validation table. The audit saves the risk bank, each conditional normalizer, every corrected field and a hash of the loaded checkpoint. An opposite-sign raw-risk update subtracts the same correction that objective matching adds. Its step magnitude is not optimized against the validation target. This control tests the consequence of the sign at a fixed field and does not establish an optimal physics-guidance procedure.

Experiments ran on one 96 GB RTX PRO 6000 GPU. Concurrent seed queues shared this device. The measured full-evaluation durations include raw generation, ordinary denoising, matched denoising where applicable, metric computation and the independent bank. Their ratio does not isolate denoiser latency, and concurrent wall times are not fair throughput comparisons. A speed claim would require isolated, synchronized, equal-work timings and a separately specified amortization policy for repeated conditions. No such speed superiority is claimed.

I.5 Numerical verification and artifact checks

Before the cloud runs, a native-grid two-step resource check verified finite losses and parameter gradients. Focused tests check the bounded risk, conditional normalizer bounds, the $\lambda = 0$ reduction, input-gradient correction, parameter-gradient connectivity through a real tiny autoregressive flow, and preservation of inference parameter gradients. The new optional score-function risk estimator is separately checked against an analytic Gaussian expectation; it is not enabled in the reported experiment.

The score-function estimator uses

$$\nabla_\theta \mathbb{E}_{p_\theta}[r] = \mathbb{E}_{p_\theta}\big[(r - b(c))\nabla_\theta \log p_\theta(U \mid c)\big], \tag{71}$$

where the baseline is independent of each evaluated draw. It can avoid retaining the inverse sampling graph, but still requires fresh model samples. A detached empirical mean from the same sample batch is not an independent baseline. Its variance and actual resource savings require measurement; the numerical identity alone does not establish faster or better learning.

Every run stores the resolved configuration, source snapshot, data hashes, RNG seeds, calibration, training logs, diagnostic evaluations and full checkpoints. Each saved prediction bundle includes targets and conditions for verification, with a recorded latent hash. Postprocessing verifies hashes and pairing before drawing a comparison. Displayed fields are direct bundle values without cosmetic smoothing. Remote artifacts are copied locally and checked before the compute instance is shut down. These records permit reproduction of the pilot even when subsequent implementations change.

J Complete validation replication results

All nine declared runs are retained. Each table reports all five output protocols, with mean $\pm$ sample standard deviation across seeds 42, 43 and 44. The repeated validation examples are not additional independent training replications. PSNR and SSIM use target-based ranges defined above. No fresh test split was opened for these pilots.

Table 4: Darcy: all validation protocols.

Model	Output	RMSE ↓	PSNR ↑	SSIM ↑	RMAE ↓
TARFlow	raw	$\mathbf{0.05505} \pm 0.00028$	$\mathbf{18.14} \pm 0.043$	$\mathbf{0.08282} \pm 0.00086$	$\mathbf{954.5} \pm 4.6$
TARFlow	score	$\mathbf{0.02101} \pm 0.00015$	27.07 ± 0.076	0.6422 ± 0.00069	46.07 ± 1.2
PAIM-Flow	raw	0.05507 ± 0.00023	$\mathbf{18.14} \pm 0.035$	0.08272 ± 0.00083	955 ± 3.8
PAIM-Flow	score	$\mathbf{0.02101} \pm 0.00014$	$\mathbf{27.08} \pm 0.073$	$\mathbf{0.6423} \pm 0.00055$	$\mathbf{45.91} \pm 1.3$
PAIM-Flow	matched	$\mathbf{0.02101} \pm 0.00014$	$\mathbf{27.08} \pm 0.073$	$\mathbf{0.6423} \pm 0.00056$	45.98 ± 1.3

Table 5: Poisson: all validation protocols.

Model	Output	RMSE ↓	PSNR ↑	SSIM ↑	RMAE ↓
TARFlow	raw	0.1171 ± 0.0067	19.26 ± 0.53	0.08499 ± 0.0026	286 ± 3.3
TARFlow	score	0.115 ± 0.0068	19.49 ± 0.56	0.1131 ± 0.0029	66.6 ± 6.4
PAIM-Flow	raw	$\mathbf{0.1165} \pm 0.0066$	$\mathbf{19.31} \pm 0.53$	$\mathbf{0.08531} \pm 0.0026$	$\mathbf{285.8} \pm 3.1$
PAIM-Flow	score	$\mathbf{0.1143} \pm 0.0068$	$\mathbf{19.55} \pm 0.56$	$\mathbf{0.1136} \pm 0.0029$	$\mathbf{65.97} \pm 6.3$
PAIM-Flow	matched	$\mathbf{0.1143} \pm 0.0068$	$\mathbf{19.55} \pm 0.56$	$\mathbf{0.1136} \pm 0.0029$	65.98 ± 6.3

Table 6: Turbulent: all validation protocols.

Model	Output	RMSE ↓	PSNR ↑	SSIM ↑	RMAE ↓
TARFlow	raw	0.2179 ± 0.0017	$\mathbf{13.05} \pm 0.062$	$\mathbf{0.03198} \pm 0.00027$	$\mathbf{0.01941} \pm 6.6\times10^{-5}$
TARFlow	score	0.2111 ± 0.0017	$\mathbf{13.33} \pm 0.066$	$\mathbf{0.04274} \pm 0.00048$	$0.00463 \pm 6.3\times10^{-5}$
PAIM-Flow	raw	$\mathbf{0.2178} \pm 0.0017$	$\mathbf{13.05} \pm 0.064$	0.03196 ± 0.00019	$\mathbf{0.01941} \pm 7\times10^{-5}$
PAIM-Flow	score	$\mathbf{0.211} \pm 0.0018$	$\mathbf{13.33} \pm 0.069$	0.04272 ± 0.0004	$\mathbf{0.004623} \pm 6.5\times10^{-5}$
PAIM-Flow	matched	$\mathbf{0.211} \pm 0.0018$	$\mathbf{13.33} \pm 0.069$	0.04272 ± 0.0004	$0.004624 \pm 6.5\times10^{-5}$

J.1 Individual training seeds

The table below retains every endpoint, including improvements and regressions. Raw and denoised outputs are separate protocols of the same checkpoint. Values are rounded only for display; machine-readable files retain full precision.

Table 7: All training seeds and output protocols.

Model	Output	Seed	RMSE ↓	PSNR ↑	SSIM ↑	RMAE ↓
Darcy						
TARFlow	raw	42	0.0553208	18.1019	**0.0818689**	959.388
TARFlow	score	42	0.0210324	27.1001	0.64144	46.8572
PAIM-Flow	raw	42	**0.0553066**	**18.1042**	0.0818279	**959.18**
PAIM-Flow	score	42	**0.021018**	**27.1078**	**0.641663**	**46.6717**
PAIM-Flow	matched	42	0.0210182	27.1076	0.641643	46.8117
TARFlow	raw	43	**0.0550607**	**18.1409**	**0.0835364**	**953.918**
TARFlow	score	43	0.0211531	26.9887	0.642679	46.7042
PAIM-Flow	raw	43	0.0550623	18.1406	0.0834821	954.016
PAIM-Flow	score	43	**0.0211433**	**26.9935**	**0.642707**	**46.609**
PAIM-Flow	matched	43	0.0211434	**26.9935**	0.642701	46.6564
TARFlow	raw	44	**0.0547701**	**18.188**	**0.0830411**	**950.143**
TARFlow	score	44	**0.0208581**	**27.1329**	**0.642574**	44.6417
PAIM-Flow	raw	44	0.0548557	18.1744	0.0828424	951.861
PAIM-Flow	score	44	0.0208633	27.1303	0.642464	**44.4443**
PAIM-Flow	matched	44	0.0208634	27.1303	0.642464	44.457
Poisson						
TARFlow	raw	42	0.124216	18.7237	0.0826566	289.439
TARFlow	score	42	0.12219	18.9246	0.109937	71.0736
PAIM-Flow	raw	42	**0.123454**	**18.7778**	**0.0830114**	**289.047**
PAIM-Flow	score	42	**0.121415**	**18.9814**	**0.110495**	**70.399**
PAIM-Flow	matched	42	0.121416	18.9813	0.110494	70.4019
TARFlow	raw	43	0.110909	19.7851	0.0845531	**282.868**
TARFlow	score	43	0.108602	20.0531	0.113512	59.2486
PAIM-Flow	raw	43	**0.110309**	**19.8314**	**0.0848467**	282.943
PAIM-Flow	score	43	**0.107984**	**20.1031**	**0.114076**	**58.7558**
PAIM-Flow	matched	43	**0.107984**	**20.1031**	0.114074	58.7587
TARFlow	raw	44	0.116274	19.2761	0.0877632	285.826
TARFlow	score	44	0.114155	19.5002	0.115766	69.4842
PAIM-Flow	raw	44	**0.115636**	**19.3255**	**0.0880822**	**285.341**
PAIM-Flow	score	44	**0.113506**	**19.5523**	**0.116289**	**68.7645**
PAIM-Flow	matched	44	**0.113506**	**19.5523**	0.116288	68.7662

Model	Output	Seed	RMSE ↓	PSNR ↑	SSIM ↑	RMAE ↓
Turbulent						
TARFlow	raw	42	0.219693	12.9792	0.0317158	0.0193474
TARFlow	score	42	0.212973	13.2579	0.042198	0.00456468
PAIM-Flow	raw	42	**0.21964**	**12.9814**	**0.0317582**	**0.0193376**
PAIM-Flow	score	42	**0.212923**	**13.2601**	**0.0422585**	**0.00455478**
PAIM-Flow	matched	42	**0.212923**	**13.2601**	0.0422581	0.00455534
TARFlow	raw	43	0.217706	13.0543	0.0319597	0.0194039
TARFlow	score	43	0.210897	13.3389	0.042876	0.00463535
PAIM-Flow	raw	43	**0.217576**	**13.0595**	**0.0319651**	**0.0194007**
PAIM-Flow	score	43	**0.210763**	**13.3444**	**0.0428945**	**0.00463079**
PAIM-Flow	matched	43	**0.210763**	**13.3444**	0.0428941	0.00463137
TARFlow	raw	44	0.216398	13.1017	**0.032255**	0.019479
TARFlow	score	44	0.209533	13.3897	**0.0431369**	0.00469091
PAIM-Flow	raw	44	**0.216205**	**13.1091**	0.032145	**0.0194773**
PAIM-Flow	score	44	**0.209333**	**13.3976**	0.0430018	**0.00468444**
PAIM-Flow	matched	44	**0.209333**	**13.3976**	0.0430015	0.00468507

J.2 Physical-gradient and displacement magnitudes

Each branching-point gradient probe uses 32 MLE examples and one generated physics field. Ratios establish scale and connectivity, not precise population-gradient statistics. The last column is the endpoint mean matching-displacement norm divided by the NF denoising-displacement norm.

Table 8: Calibration and measured intervention magnitudes.

Task	Seed	κ	Parameter norm ratio	Displacement ratio
Darcy	42	2.9457×10^{6}	0.0017507	0.00029062
Darcy	43	2.8741×10^{6}	0.0016284	0.00029079
Darcy	44	2.8762×10^{6}	0.0016898	0.00029068
Poisson	42	0.56421	0.0042651	0.00017618
Poisson	43	0.43216	0.0066529	0.0001902
Poisson	44	0.49254	0.0042337	0.00017899
Turbulent	42	0.0008816	0.00096985	0.00015768
Turbulent	43	0.00088961	0.00113	0.00015789
Turbulent	44	0.00089658	0.0010172	0.00015728

J.3 Frozen-endpoint bank and sign audit

All rows use the first 128 validation examples at seed42, with nested independent banks. The negative control subtracts the M16 matching correction at the raw state. It is not the PAIM-Flow inference rule. No target-based coefficient tuning is performed.

K Recovery Campaign: All Architectures and Output Protocols

The primary table uses the best shared-MLE Poisson architecture. Two inferior architectures were stopped adaptively before completing their physical branches; their diagnostics, failures and checkpoints are retained, without invented PAIM endpoints. Raw, ordinary, matched and proximal protocols are distinct. No new independent test was opened.

Table 10: All recovery validation endpoints, seed42. Proximal correction is available to both models; for Poisson it includes an explicit spectral PDE solver.

Model	Output	RMSE ↓	PSNR ↑	SSIM ↑	RMAE ↓
Darcy					

Model	Output	RMSE ↓	PSNR ↑	SSIM ↑	RMAE ↓
TARFlow	raw	**0.0506412**	**18.849**	**0.107597**	950.354
TARFlow	score	0.00649624	37.63	0.927293	**16.0249**
PAIM-Flow	raw	0.0506413	**18.849**	0.107595	**950.334**
PAIM-Flow	score	0.00649505	**37.6317**	0.927309	16.0254
PAIM-Flow	matched	**0.00649504**	**37.6317**	**0.92731**	16.0395
TARFlow	prox	0.00616713	38.2505	0.937753	**0.237903**
PAIM-Flow	prox	**0.00616584**	**38.2525**	**0.937772**	0.237929
Turbulent					
TARFlow	raw	**0.0561115**	**24.7689**	**0.639046**	**0.0180288**
TARFlow	score	**0.0221815**	**32.9008**	**0.92052**	0.00267398
PAIM-Flow	raw	0.0561146	24.7684	0.639002	0.0180289
PAIM-Flow	score	0.0221866	32.8988	0.920475	**0.00267383**
PAIM-Flow	matched	0.0221867	32.8988	0.920474	0.0026739
TARFlow	prox	**0.0169007**	**35.2801**	**0.954107**	**0.00189953**
PAIM-Flow	prox	0.0169032	35.2788	0.95409	0.00189964
Poisson concat					
TARFlow	raw	**0.0206282**	**33.6213**	**0.733586**	**266.456**
TARFlow	score	0.00413499	**48.8908**	**0.992311**	9.22421
PAIM-Flow	raw	0.0206299	33.6206	0.733554	266.49
PAIM-Flow	score	**0.00413457**	48.8899	0.992294	**9.22297**
PAIM-Flow	matched	0.00413464	48.8897	0.992294	9.22312
TARFlow	prox	$\mathbf{2.69854}\times10^{-6}$	**69.1329**	**1**	4.95467×10^{-5}
PAIM-Flow	prox	2.69953×10^{-6}	**69.1329**	**1**	$\mathbf{4.95391}\times10^{-5}$

K.1 Configuration and cost boundary

Postprocessing strengths were selected symmetrically on a fixed64-example validation prefix. Suffix metrics exclude that prefix but are not independent-test results. The original-capacity Poisson run encountered an evaluation out-of-memory failure; its saved model was preserved. Inferior architecture branches were subsequently retired using shared-MLE diagnostics, not a favorable treatment margin.

L Separating Density Matching from Numerical Consistency

Objective matching and numerical consistency address different questions. The reciprocal correction estimates the corrupted-data score under the assumptions of Appendix A. A physical solver instead changes a field to satisfy an equation or observation. Improved output after a solver does not establish that the learned score, or its matching correction, improved. We therefore expose the following operation as an inference control equally available to TARFlow and PAIM-Flow.

Let d_θ be a frozen ordinary or matched denoised field. With D state coordinates, the proximal control minimizes

$$Q(x; d_\theta, c, y) = \frac{1}{2D}\|x - d_\theta\|_2^2 + \tau R(x, c) + \nu \frac{\|M \odot (x - y)\|_2^2}{\max(1, \|M\|_1)}. \tag{72}$$

Here y contains observed values only; unobserved ground truth is not an optimizer input. Darcy and Poisson have no sparse-observation term in their conditional forward tasks. Turbulent reconstruction uses its known observation mask and the declared wall/smoothness energy. The positive strengths τ and ν are inference parameters, distinct from the physical training coefficient λ.

For the fixed-coefficient linear operators in these tasks, the energy is quadratic. The anchor term makes the unconstrained objective strictly convex. Write $R(x, c) = \frac{1}{2}x^\top H_R x + b_R^\top x + k_R$, and absorb the observation term into H_{obs} and b_{obs}. The unique stationary point solves

$$[I + D\tau H_R + DH_{\text{obs}}]x = d_\theta - D\tau b_R - Db_{\text{obs}}. \tag{73}$$

Matrix-free conjugate gradients use Hessian-vector products without constructing a dense field Hessian. Each example has its own residual, step and stopping condition. The implementation

Table 9: Normalizer sensitivity and opposite-sign raw-risk control.

Task	Protocol	RMSE ↓	RMAE ↓
Darcy	Ordinary	0.02120448	47.57056
Darcy	Matched, M4	0.02120472	47.73524
Darcy	Matched, M8	0.02120472	47.73539
Darcy	Matched, M16	0.02120472	47.73565
Darcy	Negative raw-risk control	**0.02120425**	**47.43009**
Poisson	Ordinary	0.1288418	73.2813
Poisson	Matched, M4	0.1288419	73.28393
Poisson	Matched, M8	0.1288419	73.28394
Poisson	Matched, M16	0.1288419	73.28394
Poisson	Negative raw-risk control	**0.1288416**	**73.27866**
Turbulent	Ordinary	0.2077963	0.004397759
Turbulent	Matched, M4	0.2077965	0.00439831
Turbulent	Matched, M8	0.2077965	0.00439831
Turbulent	Matched, M16	0.2077965	0.00439831
Turbulent	Negative raw-risk control	**0.2077962**	**0.00439721**

Table 11: Executed recovery configurations. Prefix counts exclude any warm-start pretraining.

Run	Parameters	Prefix	Continuation	Physics interval	Bank
Darcy	29,417,728	12000	1000	10	8
Turbulent	9,157,120	0	1000	10	8
Poisson concat	38,870,272	16000	1000	10	8

checks positive curvature, records the remaining linear residual and falls back to the anchor if the objective increases. A finite iteration budget yields an approximate solve, which must be distinguished from exact constraint satisfaction.

This objective can be interpreted using an isotropic Gaussian surrogate centered at d_θ and physical/observation energy factors. That surrogate is not the learned NF density or its exact clean posterior. The operation preserves the original flow parameters, but the final corrected output is not assigned the raw flow likelihood. Plug-and-play analyses motivate explicit objectives and careful denoiser assumptions; they do not make every combination of score denoising and projection convergent (Hurault et al., 2022; Shrestha & Fu, 2026).

L.1 Exact spectral Poisson control

For conditional Poisson, a separate control enforces the known zero Dirichlet boundary. Let L be the symmetric interior discrete Laplacian, $b = \rho/(h_x h_y)$, and n the number of interior nodes. With PDE residual scale s and weight w, the interior problem is

$$\min_v \frac{\|v - d_{\theta,\mathrm{int}}\|_2^2}{2D} + \frac{\tau w}{n s^2}\|Lv + b\|_2^2. \tag{74}$$

Table 12: Poisson architecture selection using32 fixed validation conditions at the shared-MLE endpoint. These are diagnostic subset scores, not full-validation PAIM results.

Architecture	Diagnostic RMSE ↓	Status
Poisson concat	**0.004475507**	Selected for paired comparison
Poisson cross	0.06314433	Stopped; missing PAIM endpoint
Poisson full	0.1253712	Stopped; missing PAIM endpoint

Boundary anchor terms are constant because boundary values are fixed at zero. The tensor-product sine basis S diagonalizes L with eigenvalues

$$\ell_{jk} = -\frac{4}{h_y^2}\sin^2\frac{j\pi}{2(H-1)} - \frac{4}{h_x^2}\sin^2\frac{k\pi}{2(W-1)}. \tag{75}$$

Writing hats for sine coefficients and $\gamma = 2D\tau w/(ns^2)$ gives

$$\widehat{v}_{jk} = \frac{\widehat{d}_{\theta,jk} - \gamma\ell_{jk}\widehat{b}_{jk}}{1+\gamma\ell_{jk}^2}. \tag{76}$$

This is a numerical solver component. As τ grows, its output tends to $-\widehat{b}_{jk}/\ell_{jk}$ and becomes independent of the learned anchor. Near-ground-truth reconstruction in this limit validates the known forward equation and discretization, not a new generative-model advantage. Applying the same operation to TARFlow is necessary to expose this distinction.

L.2 Selection and verification

The control chooses one common strength per task by the average RMSE of both model outputs on a fixed 64-example validation prefix. It never optimizes the difference between model scores. Full validation metrics are accompanied by metrics on the remaining suffix, excluding those 64 selection examples. This suffix remains part of a previously inspected validation set; it is not a new independent test.

Focused tests compare the generic solver with an analytic scalar quadratic solution, check per-example batch independence and observed-only updates, and verify stationarity on real Darcy, Poisson and turbulent operators. The spectral implementation is checked against a manufactured discrete Poisson solution and the gradient of the same interior objective. Neither numerical stationarity nor a smaller residual establishes improved statistical generalization, and the unchanged raw NF likelihood is reported separately.

M Exploratory Noise-Compatible Physical Supervision

The following objectives are separate preliminary candidates. They do not reinterpret the reciprocal-tilt checkpoints and are not evidence of superiority across three cases. They address a specific limitation of raw per-sample physics: intentional corruption variance contributes to a squared residual even when the clean field is physically valid.

For a linear residual $r(U,c) = A(c)U - b(c)$ and a fixed positive semidefinite weight W,

$$\mathbb{E}\|r(U,c)\|_W^2 = \|\mathbb{E}r(U,c)\|_W^2 + \operatorname{tr}(W\operatorname{Cov}[r(U,c)])\,. \tag{77}$$

The mean term does not directly penalize noise variance. It also cannot detect arbitrary excess variance, so a mean-only penalty is an incomplete description of the desired law.

M.1 Mean constraints and an unbiased estimator

Suppose the clean conditional mean satisfies $A(c)\mathbb{E}[X \mid c] = b(c)$. Gaussian corruption preserves this mean. A mean-consistency objective is

$$J_{\mathrm{mean}}(p) = \frac{1}{D}\mathbb{E}_{q_\sigma}[-\log p(U \mid c)] + \lambda\|\mathbb{E}_p r(U,c)\|_W^2, \qquad \lambda \ge 0. \tag{78}$$

For fixed c, subtracting the objective at q_σ gives $J_{\mathrm{mean}}(p) - J_{\mathrm{mean}}(q_\sigma) = D^{-1}\,\mathrm{KL}(q_\sigma\|p) + \lambda\|\mathbb{E}_p r\|_W^2 \ge 0$. Thus the desired noisy density remains a population optimum; there is no reciprocal density factor to compensate. This statement assumes feasibility and finite expectations.

For M independent raw draws, an unbiased U-statistic (Hoeffding, 1948) is

$$\widehat{L}_{\mathrm{mean}} = \frac{\|\sum_m r_m\|_W^2 - \sum_m \|r_m\|_W^2}{M(M-1)}. \tag{79}$$

Its minibatch value may be negative. Clipping negative estimates or substituting correlated antithetic draws would change the estimator. Independence is conditional on the fixed condition; all draws still depend differentiably on the same network parameters.

M.2 MATCHING THE INTENDED RESIDUAL NOISE LAW

For the Poisson prototype, let $G(c)$ be the equation-derived Dirichlet solution and set $e = U - G(c)$. Under the intended deterministic-clean corruption, $e \sim \mathcal{N}(0, \sigma^2 I)$. This solution-norm residual explicitly uses a numerical solver in training; it is not a free operation or a claim of solver-free local PDE supervision.

Let v_j be fixed unit-norm low-frequency cosine modes. For frequencies t_ℓ, define

$$\phi_{j\ell}(e) = \begin{bmatrix} \cos(t_\ell v_j^\top e/\sigma) \\ \sin(t_\ell v_j^\top e/\sigma) \end{bmatrix}, \qquad m_{\sigma,j\ell} = \begin{bmatrix} \exp(-t_\ell^2/2) \\ 0 \end{bmatrix}. \tag{80}$$

The centered features $\psi = \phi - m_\sigma$ have zero expectation under the desired noise. We apply the same independent-draw U-statistic to ψ and train with

$$J_{\text{law}}(p) = D^{-1}\mathbb{E}_{q_\sigma}[-\log p(U \mid c)] + \lambda \|\mathbb{E}_p \phi(e) - m_\sigma\|_2^2. \tag{81}$$

The second term is a finite-feature kernel mean discrepancy (Gretton et al., 2012). Since it is nonnegative and vanishes at q_σ, the same population-optimum argument applies. Standard Gaussian score denoising is compatible with that optimum. Finite features do not uniquely identify a distribution, bound score error, or guarantee finite-model reconstruction gains.

The implemented probe uses 8×8 orthonormal spatial modes, frequencies $\{0.125, 0.5, 1\}$, eight independent raw draws, and 300 extra updates. The auxiliary is evaluated every ten updates with frequency compensation. Three TRAIN gradient probes fix its coefficient to target a 0.2 parameter-gradient norm ratio before validation. The paired MLE control uses identical data, corruption, initial weights and optimizer settings, and is reused exactly between the two candidates. No denoised estimate is supervised during training. Each candidate uses its own ordinary score at inference, not the earlier rational correction.

Table 13: Exploratory Poisson objectives,300 additional paired updates,seed42. Both candidates use equation-derived solution-norm supervision, not the rational-tilt correction. The MLE control is reused exactly. Results are single-seed validation, not confirmation across three physical cases.

Model	Training auxiliary	RMSE ↓	PSNR ↑	SSIM ↑	RMAE ↓
TARFlow	None	**0.004188049**	**48.78261**	**0.9920875**	**9.25938**
PAIM-Flow	Mean-risk	0.004264879	48.62025	0.9919148	9.300324
PAIM-Flow	Noise-law	0.004271581	48.60872	0.9918258	9.41793

The probes use all 2048 Poisson validation examples at one training seed. Their status is exploratory, not independent-test confirmation. Extending the argument to nonlinear residuals, approximate data constraints or the turbulent smoothness proxy requires new assumptions and tests; it cannot be inferred from the Poisson experiment.

Neither prototype improves the paired MLE-only control in this experiment. Ordinary-denoised validation RMSE is 0.00418805 for TARFlow, 0.00426488 for the mean objective and 0.00427158 for the residual-law objective. A separate TRAIN diagnostic uses 64 draws on each of 16 conditions not used to set the coefficient. The low-mode residual variance divided by σ^2 is 3.121, 3.164 and 3.149, respectively. The estimated feature discrepancy is 0.02764, 0.02926 and 0.02789. Thus the short prototypes do not demonstrate improvement even in these intended distributional diagnostics. This result limits the interpretation of the population compatibility argument and motivates variance-aware optimization checks before expansion.

N EXPLICIT CORRUPTION-COVARIANCE MATCHING

A separate candidate specifies the corruption law instead of inferring a tilt from raw-risk training. For fixed positive definite Σ_c, let $U = X + \epsilon$ with $\epsilon \mid c \sim \mathcal{N}(0, \Sigma_c)$. Differentiating the Gaussian convolution yields

$$\mathbb{E}[X \mid U = u, c] = u + \Sigma_c \nabla_u \log q_\Sigma(u \mid c). \tag{82}$$

This identity motivates covariance-matched inference after MLE training on the same corruption. It is a different objective from the model-output physical risk of the main method. It does not prove the raw-risk matching hypothesis.

The experiment uses inverse-quadratic reference-Laplacian spectral weights, normalized to keep $\mathrm{tr}(\Sigma) = D\sigma^2$. Poisson uses the Dirichlet interior basis and independent boundary noise. Darcy and turbulence use a Neumann graph-Laplacian roughness proxy, not their full physical equations. All eigenvalues remain positive. The isotropic control and both candidate strengths start from the same restored checkpoint, use paired data/noise streams, and receive identical continuation budgets. Each candidate is evaluated using its own raw samples with the isotropic correction and with its actual covariance. No denoising computation participates in training.

Table 14: Separate covariance-corruption candidates, validation seed42,1000 paired updates. Ordinary and matched denote inference on identical weights and raw samples. $\gamma = 0$ is the isotropic TARFlow control. This experiment does not use the earlier raw-risk physical loss.

Task	γ	Output	RMSE ↓	PSNR ↑	SSIM ↑	RMAE ↓
Darcy	0	score	**0.00649146**	**37.6425**	**0.927099**	**15.782**
Darcy	1	score	0.0151826	29.4162	0.567287	234.358
Darcy	1	matched	0.0187941	27.5	0.456054	314.984
Darcy	9	score	0.0373691	21.5108	0.1738	644.071
Darcy	9	matched	0.0392414	21.0768	0.177216	451.775
Poisson	0	score	**0.0041568**	**48.8543**	**0.992202**	**9.22461**
Poisson	1	score	0.00899002	41.0341	0.940164	100.485
Poisson	1	matched	0.00760234	42.6193	0.961642	71.4729
Poisson	9	score	0.0249129	32.0105	0.664057	343.75
Poisson	9	matched	0.0149894	36.6073	0.875763	81.1142
Turbulent	0	score	**0.0222559**	**32.8716**	**0.919925**	**0.00267973**
Turbulent	1	score	0.0305049	30.0867	0.845964	0.00666218
Turbulent	1	matched	0.02892	30.551	0.864978	0.00704847
Turbulent	9	score	0.053852	25.1279	0.645248	0.016841
Turbulent	9	matched	0.0438189	26.9227	0.738132	0.0109982

This control distinguishes two comparisons: whether the covariance rule helps a model trained on anisotropic corruption, and whether that model outperforms the ordinary isotropic TARFlow baseline. Only the second can establish a practical improvement over the baseline. All tested strengths and unfavorable outcomes are retained. Finite-training results do not inherit the exact-score identity automatically. Signed difference plots expose the numerical changes on fixed examples without altering predictions or the common scales of the accompanying reconstruction plates. In this continuation study, covariance matching improves some anisotropic models relative to their isotropic denoiser but no candidate outperforms the matched-budget isotropic baseline. The result does not support replacing the primary method with this candidate, and no state-of-the-art claim follows from the larger within-model correction.

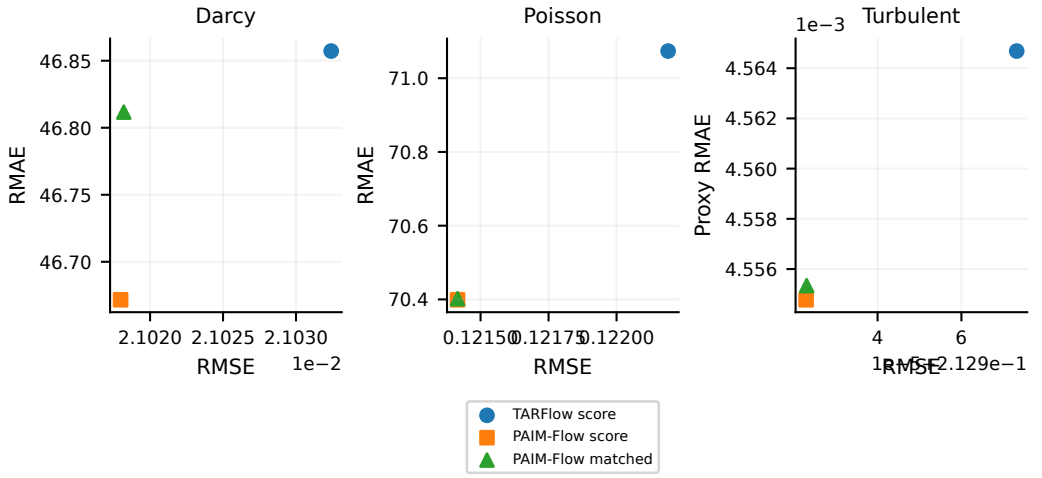


Figure 6: Seed42 validation reconstruction–residual coordinates. TARFlow ordinary, PAIM-Flow ordinary and PAIM-Flow matched outputs are measured separately. Axis offsets expose small differences; the plot should not be interpreted as a large relative improvement. The turbulent vertical axis is the declared wall/smoothness proxy.

O ADDITIONAL RECONSTRUCTION RESULTS

All panels use recovered seed42 endpoints and unmodified validation arrays.

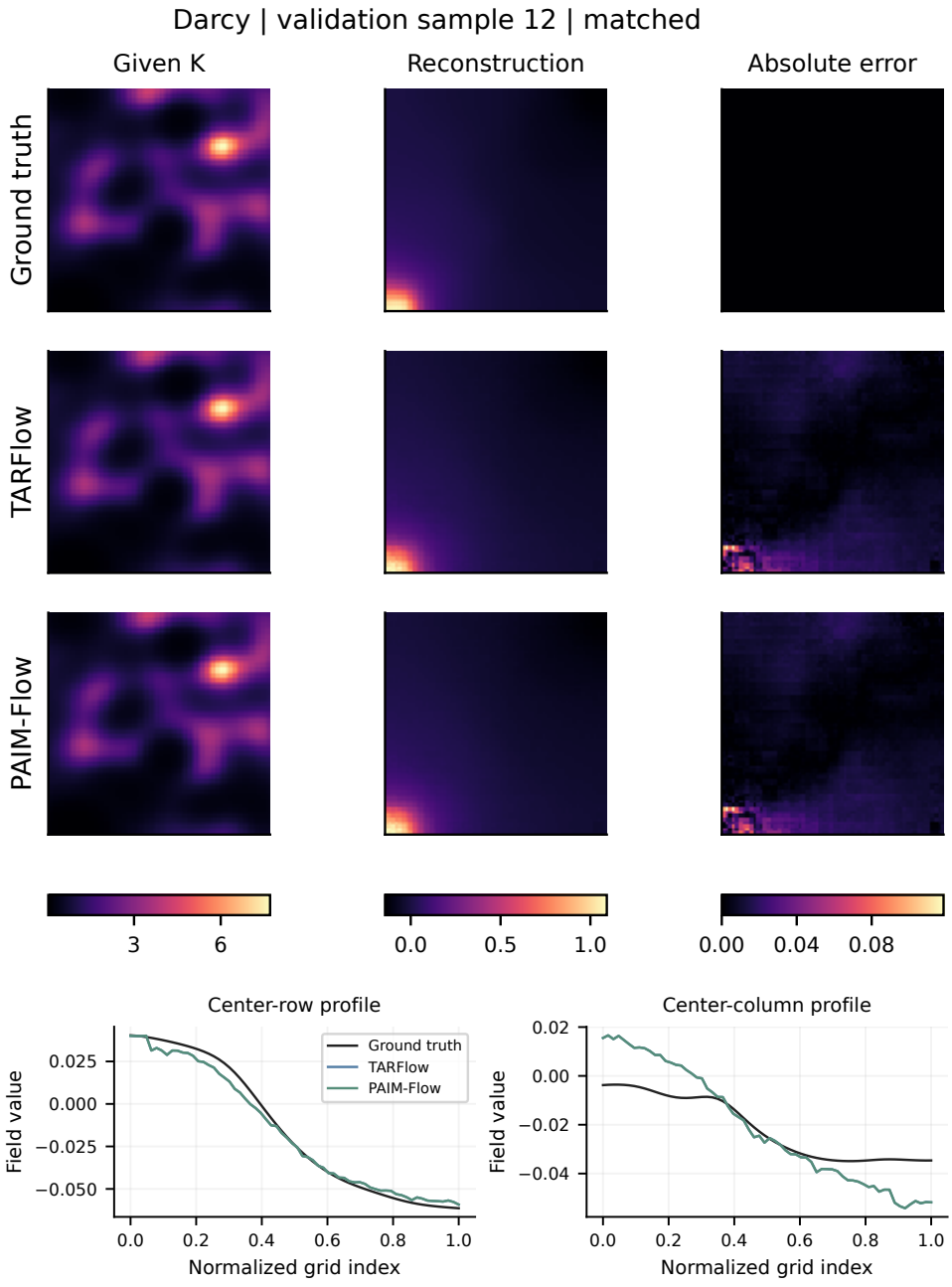


Figure 7: Darcy validation sample 12, seed 42. All predictions come from the new matched experiment. Center-line profiles compare amplitudes on the same field scale; substantial remaining reconstruction error is not hidden by separately normalized images.

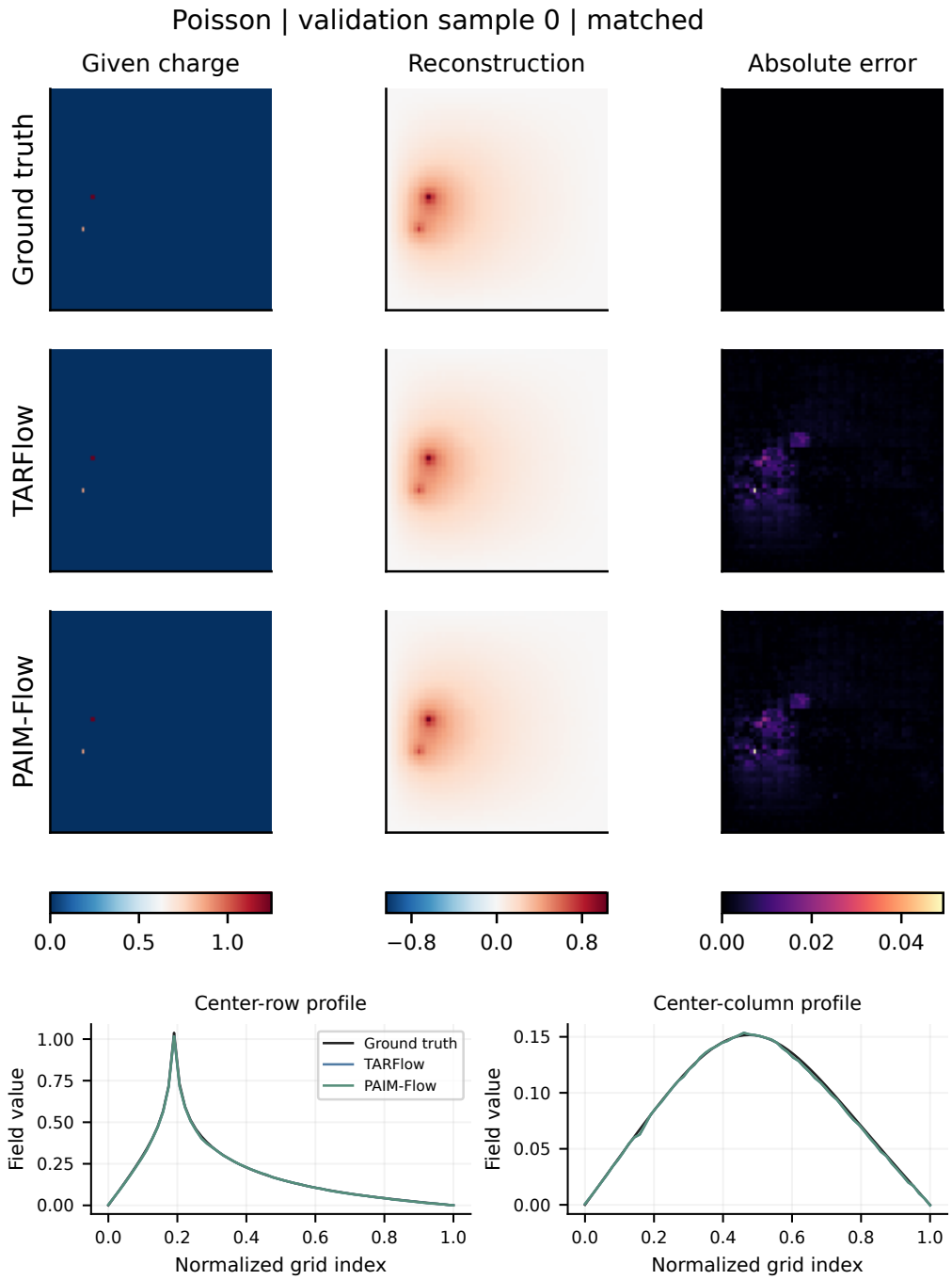


Figure 8: Conditional Poisson reconstruction, sample 0. Only ground truth and final denoised outputs are shown in three rows. Signed potential uses a shared symmetric color range. The center-line profiles show remaining low-frequency structure and amplitude errors.

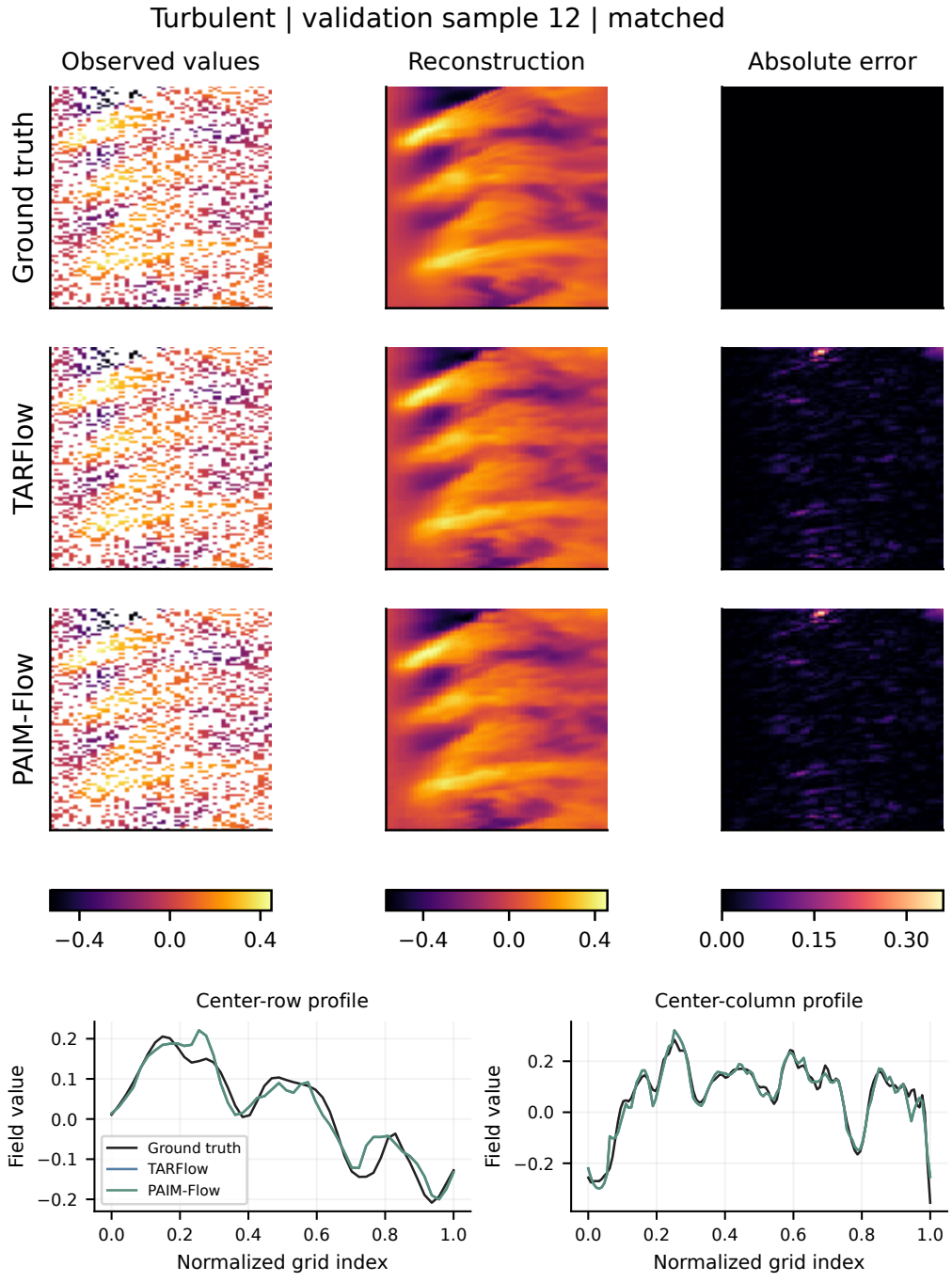


Figure 9: Turbulent sparse reconstruction, sample 12. Native evaluation remains 128×48; near-square display changes aspect only. White input locations are unobserved. Profiles use normalized grid indices and TRAIN-normalized field values.