TUBEFM: CONDITIONAL FLOWS IN LIKELIHOOD-SCALED COORDINATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Posterior samplers for noisy inverse problems must represent uncertainty in measured as well as unmeasured directions. TubeFM learns conditional flows in an affine chart comprising a fiber coordinate and a noise-standardized measurement residual. We use this chart to connect posterior approximation, velocity regression, and numerical integration. An exact Wasserstein decomposition quantifies the error imposed by hard measurement constraints; a flow-stability bound transfers velocity error to image space; and explicit prior assumptions yield noise-uniform posterior curvature bounds. Gaussian analysis separates this target geometry from the dynamics of the chosen probability path. With the source, path, and error criterion transformed together, an affine change of coordinates preserves the mapped Runge–Kutta trajectory and evaluation count in exact arithmetic; changing the source can alter both. Experiments on 135 Gaussian problems, real handwritten images, procedural Fourier imaging, and an exploratory fastMRI pilot test these distinctions. Three-seed learned-flow comparisons show that a source-matched image model can outperform TubeFM in reconstruction, while hard projection removes measured-coordinate uncertainty. The MRI pilot reproduces this uncertainty distinction with nearly unchanged PSNR on two validation volumes. We also quantify the finite-ensemble effect in raw percentile coverage. Together, the analysis and controls identify what likelihood scaling contributes and separate it from source selection, posterior calibration, and solver efficiency.

1 INTRODUCTION

For the scalar model $Y = X + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $\sigma > 0$, setting every posterior draw to Y confuses an observation with its unknown noise-free value. In a partially observed inverse problem, the same operation can preserve diversity in unmeasured directions while eliminating uncertainty in measured directions. This distinction can be nearly invisible in a posterior mean image. It becomes explicit when the output is evaluated as a distribution.

Our starting point is measurement consistency under a noisy likelihood. Exact projection is appropriate for a noiseless model. With positive noise and a prior density, a distribution confined to an observation fiber is singular with respect to the posterior. This standard support fact, recalled in the supplement, has a concrete consequence: variability in unmeasured directions cannot compensate for missing measured variability. The relevant comparison must therefore specify the observation model and the posterior being approximated.

We consider an affine chart whose two blocks are position along an observation fiber and standardized displacement away from it. The likelihood becomes a unit quadratic in the second block, while the prior retains the dependence between blocks. TUBEFM learns a conditional flow in this chart. The posterior target is unchanged, but the source, regression geometry, and numerical error scale become explicit modeling choices. Spectral noisy inverse sampling already appears in diffusion methods such as DDRM (Kawar et al., 2022), and likelihood-informed coordinates are established in Bayesian computation (Cui et al., 2014). Here we connect this geometry to approximation error and conditional-flow dynamics, then test which effects survive a controlled comparison.

We develop three connected results. First, an exact Wasserstein decomposition separates the unavoidable innovation error of a hard constraint from the remaining error in the fiber marginal. Second, a flow-stability estimate carries population velocity error through the inverse chart, exposing the

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.

dependence on rank, noise, and operator singular values. Third, a Gaussian calculation shows when likelihood scaling bounds the conditional-flow Jacobian and when a claimed solver improvement instead comes from changing the source or path. Standard support, change-of-variables, and Bayes-risk arguments appear in the supplement.

The experiments separate three questions: whether the geometry is correct, whether a learned sampler benefits from it, and whether its uncertainty is useful under misspecification. Exact Gaussian problems provide a numerical control; real handwritten images support repeated training and held-out evaluation; procedural Fourier imaging tests calibration and lesion errors. A NYU fastMRI pilot adds acquired anatomy using one trained model and two validation volumes. The resulting evidence is deliberately mixed: source matching can improve a competing image-domain flow, and marginal image calibration can coexist with poor lesion coverage. These controls are essential to interpreting the benefits of likelihood scaling.

2 Related work

Learning probability flows. Flow matching learns a marginal velocity through regression on conditional paths (Lipman et al., 2023); generalized conditional flow matching permits non-Gaussian sources and different endpoint couplings (Tong et al., 2024). Stochastic interpolants provide a related construction of deterministic and stochastic transports (Albergo & Vanden-Eijnden, 2023; Albergo et al., 2025). TubeFM applies this transport framework in likelihood-scaled coordinates. Its population endpoint argument follows existing theory; our analysis concerns the resulting approximation metric, conditioning, and uncertainty. Translating velocity error into distributional error still requires regularity of the learned dynamics (Benton et al., 2024).

Geometry of noisy inverse problems. Likelihood-informed subspaces distinguish directions strongly informed by observations from directions dominated by a prior (Cui et al., 2014). Optimal low-rank Gaussian posterior approximations likewise depend on the interaction between likelihood and prior precision (Spantini et al., 2015). Our affine chart retains every direction and normalizes the observed residual by its noise scale. Spectral posterior sampling has an established precedent in DDRM, which constructs a diffusion in singular-vector coordinates and accounts explicitly for noisy observed coefficients (Kawar et al., 2022). More recently, Blumenthal et al. (2026) combine a complex Gaussian MRI likelihood with diffusion priors and preconditioned Langevin updates. TubeFM studies how residual centering and likelihood scaling affect conditional velocity learning and its error geometry. An empirical advantage attributable to these coordinates requires controls with matched source distributions, losses, and computational budgets.

Hard constraints and likelihood-based reconstruction. MRI-CNF generates a null-space component and combines it with measured data (Wen et al., 2023). NullFlow learns a one-step flow on a measurement fiber; its principal theory assumes noiseless measurements and its reported experiment is natural-image inpainting (Shi et al., 2026). Measured-Subspace Consistency explicitly studies a hard observation model (Lee & Choi, 2026). Our positive-noise analysis concerns a different target distribution and is compatible with those noiseless guarantees. DPS incorporates noisy likelihood information without a strict projection (Chung et al., 2023). FLOWER combines a pretrained flow with measurement-aware destination refinement and Bayesian approximations (Pourya et al., 2026). UPMRI learns from undersampled multi-coil data and uses dual-space integration with practical consistency updates; its exact-consistency proposition takes the measurement noise to zero (Luo et al., 2025). Meaningful comparisons must therefore specify the likelihood, projection implementation, and training information.

Reconstruction quality and uncertainty. End-to-end variational networks provide strong supervised MRI reconstruction baselines (Sriram et al., 2020); FLAT uses flow-inspired training to improve unrolled reconstruction trajectories (Qi et al., 2025). Their point estimates address a different output than a posterior sample ensemble. Conformal image-to-image regression provides distribution-free uncertainty procedures (Angelopoulos et al., 2022), with recent MRI work using conformal quantile regression (Giannakopoulos et al., 2026). Such coverage does not by itself establish posterior accuracy or structural reliability: voxel intensity variability can be a poor surrogate for morphological uncertainty (Trautmann et al., 2025). We therefore distinguish reconstruction error, residual disper-

sion, interval coverage, and interval width. A coordinate change alone supplies neither conditional coverage nor a guarantee of clinical reliability.

3 CONDITIONAL TRANSPORT IN LIKELIHOOD COORDINATES

Let $Y = AX + \epsilon$, where $A \in \mathbb{R}^{m\times n}$ has rank q, $0 < q < n$, and $\epsilon \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \succ 0$. The prior has a Lebesgue density p_0. Put $B = \Sigma^{-1/2}A$, $b = \Sigma^{-1/2}y$, and write a compact SVD $B = U_q S V_q^\top$. Let V_0 be an orthonormal basis of $\ker B$. All probability comparisons below condition on (y, A, Σ) unless otherwise stated. Complex systems are represented by their real and imaginary components with the corresponding real noise covariance.

Chart and posterior. Define

$$C_y(x) = (u, r) = \big(V_0^\top x,\ U_q^\top (Bx - b)\big), \tag{1}$$

$$J_y(u, r) = V_0 u + V_q S^{-1}(U_q^\top b + r) = Dc + d_y. \tag{2}$$

These affine maps are inverses. Orthogonality gives $\|Bx - b\|^2 = \|r\|^2 + \big\|(I - U_q U_q^\top)b\big\|^2$, so the transformed posterior has density

$$\pi_y(u, r) \ \propto\ p_0(J_y(u, r)) \exp(-\|r\|^2/2)\,|\det S|^{-1}. \tag{3}$$

The second term in the orthogonal decomposition is constant in x, including when the exact observation fiber is empty. The set $r = 0$ is then a least-squares slice. The Gaussian likelihood factor in (3) does *not* make the posterior residual standard normal or independent of u.

Two related metrics. The unweighted tube distance is $d_{\mathrm{tube}}(x, x') = \|C_y(x) - C_y(x')\|$. For regression we use $M = \mathrm{diag}(I_{n-q}/(n-q), I_q/q)$ and $d_M(x, x') = \|C_y(x) - C_y(x')\|_M$. These metrics differ by dimension-dependent factors. The block-normalized form avoids making the larger block dominate solely through its coordinate count; it is a modeling choice, not an additional likelihood term. If $q = n$, omit the empty fiber block instead of dividing by zero.

Training and sampling. For each training image and simulated observation, draw independent endpoints

$$C_0 \sim \mathcal{N}(0, I_n), \qquad C_1 = C_Y(X), \qquad C_t = (1-t)C_0 + tC_1, \quad t \sim \mathcal{U}[0, 1]. \tag{4}$$

Jointly drawing X and $Y \mid X$ supplies the correct conditional target in population. A conditional velocity network minimizes

$$\mathcal{L}(\theta) = \mathbb{E}\,\|v_\theta(C_t, t; Y, A, \Sigma) - (C_1 - C_0)\|_M^2\,. \tag{5}$$

Its population minimizer is the conditional expectation of the target velocity. Under integrability, nonexplosion, and uniqueness of the corresponding continuity-equation solution, its ODE carries the source to π_y (Lipman et al., 2023). Sampling draws C_0, integrates the learned field, and returns $J_y(C_1)$. All empirical claims depend on the learned approximation and numerical integration, not merely this population identity.

For an embedded solver, measure the local discrepancy δc in M-norm. A coordinate comparison must pull back the same tolerance together with the state and source. Using a standard normal independently in image and tube coordinates changes the physical source. We report that comparison separately from the same-source control. Likewise, a post-sampling hard projection $J_y(u, r) \mapsto J_y(u, 0)$ is an ablation, not a step of the positive-noise sampler.

Cartesian Fourier implementation. For unitary F, a binary Fourier mask H, and independent component noise of scale σ, combine the two supported blocks into

$$z = (I - H)Fx + H(Fx - y)/\sigma, \qquad x = F^{-1}\big[(I - H)z + H(y + \sigma z)\big]. \tag{6}$$

This requires FFTs and masking. It is a bijection on complex images with the specified noise model. For real images, conjugate symmetry can make the realified observations rank deficient; the procedural experiment therefore uses a rank-aware real chart. Multi-coil and non-Cartesian operators need their own validated operator factorization or solver. Equation (6) is not a justification for substituting a coil-combined image in those cases.

4 APPROXIMATION AND CONDITIONING

We state the main consequences here and give proofs and qualified extensions in the supplement. The assumptions are part of the results; they are not assertions about arbitrary learned image priors.

Theorem 1 (Projection error decomposition)**.** *Let P_y, Q have finite second moments in tube coordinates. Write $(U, R) = C_y(X)$ for $X \sim P_y$. If Q is supported on $r = 0$ and its fiber marginal is Q_U, then*

$$W_{2,\mathrm{tube}}^2(P_y, Q) = \mathbb{E}\,\|R\|^2 + W_2^2(P_U, Q_U). \tag{7}$$

The unique minimizing hard-supported law is $(J_y(\cdot, 0))_{\#} P_U$.

The residual term is paid by every coupling; an optimal fiber coupling can be lifted using the conditional law of $R \mid U$. Thus hard projection loses a precise part of posterior uncertainty even if its fiber marginal is exact. Its TV distance is one under the density assumption, and its sample-derived innovation intervals have zero coverage for continuous innovation marginals. These standard observations do not rule out externally calibrated wider intervals. Also, small physical noise need not mean small standardized error: with scalar unit Gaussian prior and $y = 0$, $\mathbb{E}R^2 = (1 + \sigma^2)^{-1} \to 1$.

Theorem 2 (Population flow-error transfer)**.** *Let p_t be the law of a reference ODE with field v_* and finite second moments. Suppose the learned field is globally Lipschitz in M-norm with integrable constant $L(t)$ and both flows exist uniquely and are nonexplosive. Set $e(t)^2 = \mathbb{E}_{p_t} \|v_\theta - v_*\|_M^2$ and assume $e \in L^2[0, 1]$. Starting from the same source, the endpoint law $\widehat{p}_1$ satisfies*

$$W_{2,M}(\widehat{p}_1, p_1) \le \int_0^1 e^{\int_t^1 L(s)\, ds} e(t)\, dt. \tag{8}$$

The Euclidean image-space bound is multiplied by $\left\|DM^{-1/2}\right\|_{\mathrm{op}} = \max\{\sqrt{n-q}, \sqrt{q}/s_{\min}(B)\}$.

This is a Gronwall stability argument, related to general flow-matching error bounds (Benton et al., 2024). It controls population approximation error; empirical regression loss includes irreducible conditional variance and is not itself $e(t)^2$. Numerical endpoint error adds a further coupling term.

Curvature and Gaussian dynamics. For orthogonal sampling with $\Sigma = \sigma^2 I$ and prior potential ϕ satisfying $mI \preceq \nabla^2 \phi \preceq LI$, $m > 0$, the tube posterior Hessian has eigenvalues in

$$[\min\{m, 1 + m\sigma^2\},\ \max\{L, 1 + L\sigma^2\}], \qquad 0 < \sigma \le 1. \tag{9}$$

This removes the inverse-noise growth of likelihood curvature. It does not bound an arbitrary network Jacobian. For a Gaussian tube target $\mathcal{N}(\mu, G)$, independent-endpoint flow matching instead gives the explicit field

$$v_*(c, t) = \mu + [tG - (1 - t)I]\left[(1 - t)^2 I + t^2 G\right]^{-1}(c - t\mu). \tag{10}$$

It is affine in space but generally varies in time. If the eigenvalues of G stay within fixed positive bounds, its Jacobian is uniformly bounded. An image-target eigenvalue of order σ^2 produces a peak of order σ^{-1} for this path, not a universal σ^{-2} law. A covariance-linear probability path has different dynamics. Posterior conditioning alone does not determine solver cost.

5 EXPERIMENTS

The experiments test exact identities, learned conditional sampling, and misspecification separately. Gaussian and mixture-posterior methods sample analytic posteriors of fitted priors; neural TUBEFM learns a velocity field. Methods within each comparison share observations. The digit and procedural studies use held-out test sets, while the MRI pilot uses a predeclared validation subset. We report the unit of replication and distinguish raw sample-quantile coverage from calibrated coverage; finite-ensemble effects are analyzed in Appendix H.1.

Table 1: Gaussian integrations averaged over 135 settings. The same-source control isolates representation from source construction. Wall times were measured in a shared CPU environment.

Source / representation / error	NFE	Time (ms)	Tube W_2
Tube / tube / tube	348.12	15.57	6.1e-05
Same tube / image / tube	348.12	22.54	6.1e-05
Image / image / tube	714.44	38.08	0.00014
Image / image / image	567.85	23.41	0.0055
Null-matched / tube / tube	298.09	13.29	4.1e-05
Gaussian OT oracle / tube / tube	8.00	0.41	4.1e-08

5.1 Gaussian identities and a numerical control

We use 135 problems in dimension 32, crossing ranks 8/16/24, prior precision condition numbers 1/10/100, noise levels $0.003/0.01/0.03/0.1/0.3$, and three operator seeds. Prior and measurement eigenvectors are independently rotated. Exact conditioning provides a reference posterior for evaluating distributional error. Across 4,096 joint prior–noise cases per setting, the posterior's measured-coordinate 90% intervals cover 90.02%; the hard-fiber oracle covers 0%. The sharp squared-Wasserstein identity agrees to 2.3×10^{-7}, including numerical error from singular matrix square roots.

We then integrate the independent-endpoint Gaussian CFM field with adaptive Euler–Heun, counting rejected evaluations. The 810 main integrations use absolute tolerance 10^{-6} and relative tolerance 10^{-4}; a further 360 integrations vary tolerance. Thirty-three affine probes recover endpoint moments without sampling error. Table 1 separates the source, representation, and error criterion. Holding the physical source and path fixed produces *identical NFE in all 135 settings*. Standardizing the source separately in image coordinates increases mean NFE from 348.12 to 714.44 under the same tube error criterion. At $\sigma = 0.003$, the corresponding means are 322.00 and 820.37. The numerical benefit therefore comes from the source and path construction; a change of coordinates alone does not reduce NFE. The Gaussian OT row uses a known optimal coupling; it is not a trained OT-CFM baseline.

5.2 Real handwritten images with fitted continuous priors

We use the 1,797 real 8×8 images distributed by `load_digits`, derived from the UCI optical-digits test partition (Alpaydin & Kaynak, 1998; scikit-learn developers, 2026). A fixed 360-image test set is held out. For each of three fitting seeds, the remainder is split into 1,077 fitting and 360 validation images. These are new sample-level splits within the available writer cohort, not the original writer-disjoint benchmark. Four DCT masks retaining 16 or 32 coefficients are crossed with $\sigma \in \{0.02, 0.08, 0.20\}$. We fit a shrinkage Gaussian and an unsupervised full-covariance Gaussian mixture; mixture size and regularization are chosen by validation likelihood. Each posterior supplies 256 draws per test case. Quantized digit observations do not satisfy the continuous-prior assumptions literally; the density-based theory here applies to the fitted models.

The mixture posterior obtains 19.38 dB versus 18.62 dB after hard projection. The paired difference is 0.757 dB, with a 95% image-cluster bootstrap interval $[0.728, 0.790]$ after averaging repeated conditions and seeds within each image. Measured-coordinate 90% coverage is 89.0% versus 0%, with finite-draw and prior-misspecification effects. Pixel CRPS improves from 0.0692 to 0.0566. This comparison isolates the effect of hard projection within a fitted prior family. Analytic Gaussian CFM also exposes numerical overdispersion: at 16 evaluations the image-source flow has 96.0% measured coverage, while the tube-source flow has 89.5%. Higher coverage alone would therefore rank a less accurate posterior favorably. Full interval widths, seed variation, and numerical errors appear in the supplement.

5.3 Learned conditional flows on real images

For a separate pixel-observation experiment, we train conditional MLP flows on a fixed 970/108/359/360 optimization/validation/calibration/test split of the same dataset. A fixed mask observes 32 pixels with $\sigma \in \{0.05, 0.1, 0.2\}$. Each of three seeds trains TubeFM, standard-source image CFM, and image CFM with the same physical source as TubeFM. All models have two 192-

Table 2: Learned flows on real digit images. Means over three training seeds, 360 test images per noise level, and 64 draws per image. Innovation coverage is for nominal 90% intervals; width is in noise units. Higher coverage is not always better. The matched-source image baseline has the highest PSNR here.

σ	Method	PSNR	Innovation cov.	Width
0.05	Image CFM	18.78	99.4%	8.827
0.05	Image CFM, matched source	20.04	80.9%	2.658
0.05	TubeFM	19.28	85.2%	2.558
0.05	TubeFM + hard projection	19.19	0.0%	0.000
0.1	Image CFM	18.16	97.9%	4.494
0.1	Image CFM, matched source	19.04	88.6%	2.435
0.1	TubeFM	18.32	89.6%	2.453
0.1	TubeFM + hard projection	17.80	0.0%	0.000
0.2	Image CFM	16.25	90.9%	2.375
0.2	Image CFM, matched source	16.40	91.4%	2.404
0.2	TubeFM	16.17	90.9%	2.298
0.2	TubeFM + hard projection	14.62	0.0%	0.000

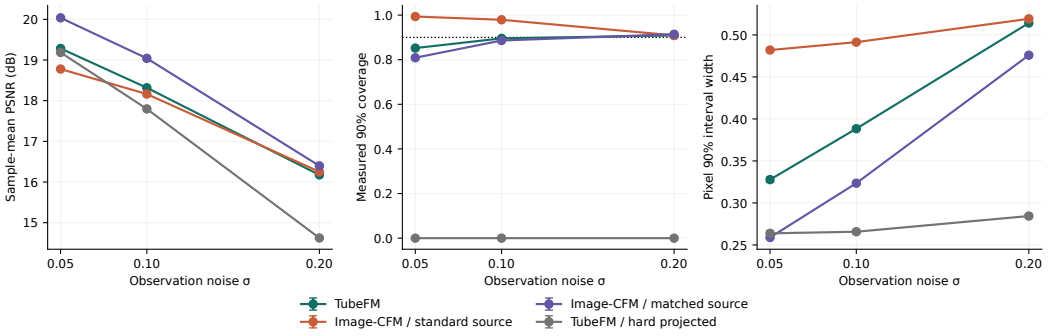


Figure 1: Learned conditional flows: reconstruction, measured coverage, and pixel interval width give different rankings. Error bars show SD across three training seeds. The source-matched image baseline has better PSNR than TubeFM at all tested noise levels; hard projection collapses measured intervals.

unit hidden layers, 69,952 parameters, and 10,000 Adam updates; validation selects the checkpoint before calibration. Each held-out conditional ensemble contains 64 draws integrated with 64 Heun steps. The source-matched baseline tests whether an apparent coordinate advantage persists when the physical source is shared. Coordinate scaling still changes the regression geometry and finite-capacity optimization.

Table 2 shows a trade-off rather than uniform dominance. At $\sigma = 0.05$, TubeFM improves PSNR over standard-source image CFM by 0.506 dB (paired image-bootstrap 95% interval $[0.454, 0.560]$), but its measured coverage is 85.2%, below the nominal level. Standard-source image CFM has 99.4% coverage with much wider innovation bands. The matched-source image baseline reaches 20.04 dB versus 19.28 dB for TubeFM, and also has lower measured coverage, 80.9%. At $\sigma = 0.2$, TubeFM does not improve PSNR over either image baseline. The hard-projection ablation has zero measured coverage at every noise level. Together, these controls show that preserving measured uncertainty and improving point reconstruction are distinct objectives. The proposed coordinates achieve the former without establishing an advantage over the source-matched image model on the latter.

Split-conformal calibration targets the event that at least 90% of image pixels are covered, using 359 calibration cases. TubeFM's observed case coverages are 89.4%, 89.3%, and 90.0% across increasing noise. The matched-source baseline also has narrower calibrated pixel bands at all three levels. Such finite-test coverage deviations do not contradict a marginal guarantee and do not establish exact-posterior recovery. The supplement reports widths, per-seed results, digit subgroups, and a 128-step solver check.

Table 3: Procedural Fourier imaging. Values are means over acquisitions; brackets are SD across three independent generator/fitting seeds. The soft Gaussian posterior improves overall reconstruction but worsens lesion error. These are synthetic, analytic-posterior results.

Method	PSNR ↑	SSIM ↑	NMSE ↓	Lesion MAE ↓
Soft Gaussian	24.35 [0.31]	0.744	0.0529	0.207
Hard conditional	22.66 [0.35]	0.688	0.0822	0.180
Zero-filled	19.06 [0.26]	0.538	0.1605	0.354
Ridge	20.69 [0.06]	0.602	0.1169	0.250

Table 4: Exploratory MRI validation: one trained seed, two volumes, three slices per volume, six conditions. Scores average slices within volume/condition, then volumes and conditions equally. Coverage is for raw empirical 5th–95th percentile bands; finite-ensemble effects are discussed in Appendix H.1. Dashes indicate a point reconstructor. No confidence intervals are inferred from two volumes.

Method	PSNR (dB)	SSIM	Pixel coverage (%)	Measured coverage (%)
TubeFM	19.575	0.3473	79.13	85.30
Hard-projected TubeFM	19.549	0.3526	78.79	< 0.001
Zero-fill	18.643	0.3401	—	—

5.4 Procedural Fourier imaging and structural failures

Three independent seeds generate 32×32 real phantoms, each with 640 training, 128 calibration, and 192 test images. Four Cartesian masks at nominal $4 \times /8\times$ acceleration and three noise levels produce 6,912 paired acquisition cases from 576 independent test images. Noise standard deviation is defined per real/imaginary component. Realification and rank reduction handle conjugate redundancy. A rank-64 PCA prior with a positive diagonal residual is fit on 544 training images; 96 separate images tune the ridge baseline. The posterior is analytic and is sampled 64 times per test case.

Table 3 shows the reconstruction trade-off. Here the hard baseline is the zero-noise conditional of the same prior, rather than the optimal hard projection of the noisy posterior. The soft endpoint improves PSNR by 1.70 dB, but lesion MAE rises from 0.180 to 0.207. Both posteriors receive the same split-conformal procedure on a fixed central ROI, using the 117th of 128 calibration scores. The criterion is that at least 90% of ROI pixels are covered in a case. Soft/hard case coverage is 90.38%/90.41%, at mean widths 0.308/0.426. However, the corresponding lesion-region case coverages are only 19.49%/56.35%. Marginal ROI calibration does not establish lesion coverage. Raw and calibrated intervals, all settings, and representative failures are retained in the supplement. Neither the lesion generator nor these learned Gaussian priors support clinical claims.

5.5 An acquired-MRI pilot on two validation volumes

We examine whether the implementation extends to acquired complex images using a pilot on NYU fastMRI knee data (Zbontar et al., 2018; Knoll et al., 2020). The 1,172 eligible volumes are assigned to a custom 821/117/117/117 training/validation/calibration/test split. The split is at scan level; patient independence is unverified. We therefore report this as a custom retrospective task. We crop complex references to 320×320 and simulate Cartesian measurements at $4 \times /8\times$ acceleration with noise $\sigma \in \{0.01, 0.03, 0.05\}$ per real/imaginary component under one training-only normalization. Source acquisitions already contain scanner noise.

We train one 77,122-parameter residual CNN for 50,000 updates with seed 11. Before evaluation, we fix a subset comprising the first two validation volumes in manifest order, with three slices per volume and all six conditions. It uses 32 draws and 64 Heun steps (128 NFE per draw). These 36 acquisitions do not constitute 36 independent scans. Calibration and test volumes are not evaluated. Both selected volumes belong to the validation split, although neither appears in the fixed batches used to select the checkpoint.

Table 4 shows nearly equal mean PSNR for TUBEFM and its post-sampling hard projection, with slightly higher SSIM after projection. The learned flow exceeds zero filling by 0.932 dB on these cases, but this comparison does not establish an advantage over learned MRI baselines. Hard projection collapses measured-coordinate intervals; the unprojected sampler retains nonzero width and 85.30% measured-coordinate coverage. With only 32 draws, this percentage alone cannot establish posterior miscalibration. Per-volume scores, proper scores, training diagnostics, and runtime measurements appear in Appendix H. Only computational profiles are available for the additional learned MRI controls, so this pilot does not isolate the effect of the coordinate system.

6 DISCUSSION AND LIMITATIONS

Likelihood scaling makes the measured component of posterior error explicit. It also exposes a distinction that is easy to miss in comparisons of learned samplers: changing coordinates, changing the physical source, and changing the regression metric are different interventions. Our Gaussian controls isolate these effects exactly. The learned-image comparisons show why this distinction matters in practice: the source-matched image flow achieves better reconstruction than TUBEFM in the tested setting. Retaining measured uncertainty is necessary for a positive-noise posterior, but is insufficient to establish posterior accuracy.

The theory has corresponding limits. Strong log-concavity in (9) excludes general multimodal image priors, and bounded target curvature alone does not bound a learned field or its integration cost. Empirically, the digit splits do not certify unseen-writer generalization, and the procedural study uses a fitted Gaussian prior. The acquired-MRI pilot is restricted to two validation volumes, one seed, and uncertified patient grouping. Its terminal validation minimum does not establish convergence; the other learned configurations have throughput profiles only, and no external MRI baseline or held-out MRI test result is available. Broader comparisons require patient-aware grouping where possible, validation-selected baselines, and structural endpoints. Within these limits, TUBEFM provides a concrete way to study noisy consistency while keeping reconstruction, distributional error, and calibration distinct.

AI USE STATEMENT

The authors conceived the central idea, led the research, evaluated the findings, and determined the paper's claims and conclusions. Under their direction, generative AI tools helped identify related literature, edit and polish manuscript drafts, and generate the procedural synthetic phantoms and prepare software smoke tests. The authors reviewed the resulting material and take responsibility for the final submission.

ETHICS STATEMENT

We evaluate TubeFM on public handwritten digits, procedural phantoms, and deidentified knee MRI accessed through NYU fastMRI. Together, these experiments examine reconstruction behavior, posterior variability, and conformal coverage for the statistical events defined in the paper. The fastMRI data remain subject to their dataset agreement, and raw volumes are not redistributed. NYU fastMRI investigators provided the data but did not participate in the analysis or writing.

REPRODUCIBILITY STATEMENT

The anonymous code supplement includes experiment code, configurations, numerical results, and scripts to reproduce the figures and tables. Appendices A–H provide proofs, protocols, and additional results. The MRI results come from one training run and a two-volume validation pilot; reproducing those experiments requires access to the restricted data. MRI checkpoints and the original split manifest are retained by the authors and are not included in the supplement.

REFERENCES

Michael Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025. URL https://www.jmlr.org/papers/v26/23-1605.html.

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.15571.

E. Alpaydin and C. Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998. URL https://archive.ics.uci.edu/dataset/80/optical+recognition+of+handwritten+digits.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. URL https://arxiv.org/abs/2107.07511.

Anastasios N. Angelopoulos, Amit Pal Kohli, Stephen Bates, Michael Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 717–730, 2022. URL https://proceedings.mlr.press/v162/angelopoulos22a.html.

Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=uqQPyWFDhY.

Moritz Blumenthal, Tina Holliber, Jonathan I. Tamir, and Martin Uecker. Fast and robust diffusion posterior sampling for MR image reconstruction using the preconditioned unadjusted langevin algorithm. *Magnetic Resonance in Medicine*, 96(3):1323–1332, 2026. doi: 10.1002/mrm.70416. URL https://onlinelibrary.wiley.com/doi/10.1002/mrm.70416.

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.

Tiangang Cui, James Martin, Youssef M. Marzouk, Antti Solonen, and Alessio Spantini. Likelihood-informed dimension reduction for nonlinear inverse problems. *Inverse Problems*, 30(11):114015, 2014. doi: 10.1088/0266-5611/30/11/114015. URL https://arxiv.org/abs/1403.4680.

Ilias I. Giannakopoulos, Lokesh B. Gautham Muthukumar, Yvonne W. Lui, and Riccardo Lattanzi. Pixel-wise uncertainty quantification of accelerated MRI reconstruction. *Magnetic Resonance in Medicine*, 2026. doi: 10.1002/mrm.70526. URL https://onlinelibrary.wiley.com/doi/10.1002/mrm.70526.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2): 217–288, 2011. doi: 10.1137/090771806. URL https://arxiv.org/abs/0909.4061.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2201.11793.

Florian Knoll, Jure Zbontar, Anuroop Sriram, Matthew J. Muckley, et al. fastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning. *Radiology: Artificial Intelligence*, 2(1):e190007, 2020. doi: 10.1148/ryai.2020190007.

Junhyeok Lee and Kyu Sung Choi. Measured-subspace consistency: A plug-and-play operator for diffusion posterior sampling in accelerated MRI reconstruction. *arXiv preprint arXiv:2606.28448*, 2026. URL https://arxiv.org/abs/2606.28448.

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://arxiv.org/abs/2210.02747.

Xinzhe Luo, Yingzhen Li, and Chen Qin. UPMRI: Unsupervised parallel MRI reconstruction via projected conditional flow matching. *arXiv preprint arXiv:2512.17493*, 2025. URL https://arxiv.org/abs/2512.17493.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990. doi: 10.1214/aop/1176990746.

Mehrsa Pourya, Bassam El Rawas, and Michael Unser. Flower: A flow-matching solver for inverse problems. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=QGd34p02mI.

Kehan Qi, Saumya Gupta, Xiaoling Hu, Qingqiao Hu, Weimin Lyu, Yicun Wang, and Chao Chen. FLAT: FLow-aligned training of unrolled networks for MRI reconstruction. *arXiv preprint arXiv:2512.03020*, 2025. URL https://arxiv.org/abs/2512.03020. Version 3, revised August 2026.

scikit-learn developers. `load_digits`: Dataset documentation, 2026. URL https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html. Accessed September 16, 2026.

Xiao Shi, Edward P. Chandler, Chicago Y. Park, Shirin Shoushtari, and Ulugbek S. Kamilov. NullFlow: One-step generative reconstruction. *arXiv preprint arXiv:2606.22696*, 2026. URL https://arxiv.org/html/2606.22696v1.

Alessio Spantini, Antti Solonen, Tiangang Cui, James Martin, Luis Tenorio, and Youssef Marzouk. Optimal low-rank approximations of Bayesian linear inverse problems. *SIAM Journal on Scientific Computing*, 37(6):A2451–A2487, 2015. doi: 10.1137/140977308. URL https://arxiv.org/abs/1407.3463.

Anuroop Sriram, Jure Zbontar, Tullie Murrell, Aaron Defazio, C. Lawrence Zitnick, Nafissa Yakubova, Florian Knoll, and Patricia Johnson. End-to-end variational networks for accelerated MRI reconstruction. In *Medical Image Computing and Computer Assisted Intervention*, pp. 64–73, 2020. URL https://arxiv.org/abs/2004.06688.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=CD9Snc73AW.

Luca L. C. Trautmann, Peter A. Wijeratne, Itamar Ronen, and Ivor J. A. Simpson. Evaluating structural uncertainty in accelerated MRI: Are voxelwise measures useful surrogates? *arXiv preprint arXiv:2503.10527*, 2025. URL https://arxiv.org/abs/2503.10527.

Jeffrey Wen, Rizwan Ahmad, and Philip Schniter. A conditional normalizing flow for accelerated multi-coil MR imaging. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 36926–36939, 2023. URL https://proceedings.mlr.press/v202/wen23a.html.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J. Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, et al. fastMRI: An open dataset and benchmarks for accelerated MRI. *arXiv preprint arXiv:1811.08839*, 2018. URL https://arxiv.org/abs/1811.08839.

A GEOMETRY, POSTERIOR DENSITY, AND HARD CONSTRAINTS

We give complete finite-dimensional arguments and separate statements about posterior geometry from statements about a chosen transport. Throughout, $A \in \mathbb{R}^{m\times n}$ has rank $0 < q < n$, $\Sigma \succ 0$, and the prior has a Lebesgue density p_0. The full-rank case uses the same formulas after deleting the empty null block; no division by $n-q$ is then made. Let $B = \Sigma^{-1/2}A = U_q S V_q^\top$ be a compact SVD, let V_0 be an orthonormal basis of $\ker B$, and set $b = \Sigma^{-1/2}y$. Define

$$C_y(x) = (V_0^\top x, U_q^\top(Bx - b)), \qquad D = [V_0, V_q S^{-1}], \tag{11}$$

$$d_y = V_q S^{-1} U_q^\top b, \qquad J_y(c) = Dc + d_y. \tag{12}$$

All posterior distributions below condition on the fixed observation y. Where Wasserstein distances are used, assume the corresponding distributions have finite second moments. Total variation uses the convention $\mathrm{TV}(P,Q) = \sup_E |P(E) - Q(E)|$.

A.1 EXACT CHART AND ITS MEANING

The identities $V_0V_0^\top + V_qV_q^\top = I$, $BV_0 = 0$, and $BV_qS^{-1} = U_q$ give $J_y(C_y(x)) = x$ and $C_y(J_y(c)) = c$. Furthermore,

$$Bx - b = U_q r - (I - U_qU_q^\top)b, \qquad \|Bx-b\|^2 = \|r\|^2 + \|(I-U_qU_q^\top)b\|^2. \tag{13}$$

The second summand is independent of x. Since $|\det D| = \det(S)^{-1}$, change of variables yields

$$\pi_y(u,r) = \frac{1}{Z_c} p_0(J_y(u,r)) \exp(-\|r\|^2/2) \det(S)^{-1}, \tag{14}$$

where Z_c is the integral of the displayed numerator. In particular, the presence of a Gaussian likelihood factor does not imply that the posterior marginal of R is standard Gaussian, or that R is independent of U. The prior term generally couples both blocks.

The unweighted tube distance is the pullback of Euclidean chart distance:

$$d_{\mathrm{tube}}(x,z)^2 = \|V_0^\top(x-z)\|^2 + \|U_q^\top B(x-z)\|^2 = \|C_y(x) - C_y(z)\|^2. \tag{15}$$

Consequently $W_{2,\mathrm{tube}}(P,Q) = W_2((C_y)_\# P, (C_y)_\# Q)$. Orthogonal changes of basis within the two coordinate blocks preserve these distances. An invertible change of measurement units, applied jointly to A, y, Σ, also preserves the likelihood quadratic form. These are geometric invariances; a neural architecture trained independently in each representation need not inherit them without a corresponding transformation of its source and vector field.

A.2 PROJECTION ONTO THE HARD SLICE

Let $H_y = \{x : U_q^\top(Bx - b) = 0\}$. This is the least-squares slice; it agrees with the exact measurement fiber only if $b \in \operatorname{range} B$. For $X \sim P$, write $(U,R) = C_y(X)$ and let P_U denote the marginal law of U. For Q supported on H_y, let $Q_U = (V_0^\top)_\# Q$.

Proposition 3 (Exact hard-slice distance decomposition)**.** *For P, Q with finite second moments and $Q(H_y) = 1$,*

$$W_{2,\mathrm{tube}}^2(P,Q) = \mathbb{E}_P\|R\|^2 + W_2^2(P_U, Q_U). \tag{16}$$

Thus the unique closest law supported on H_y is $Q^* = (x \mapsto J_y(V_0^\top x, 0))_\# P$, and

$$W_{2,\mathrm{tube}}^2(P,Q) = W_{2,\mathrm{tube}}^2(P,Q^*) + W_{2,\mathrm{tube}}^2(Q^*,Q). \tag{17}$$

If P has a Lebesgue density, then $\mathrm{TV}(P,Q) = 1$.

Proof. Every coupling of $(U,R) \sim (C_y)_\# P$ and $(U',0) \sim (C_y)_\# Q$ has cost $\mathbb{E}\|U - U'\|^2 + \mathbb{E}\|R\|^2$. Its first term is at least $W_2^2(P_U, Q_U)$. Conversely, take an optimal coupling of P_U and Q_U, and conditionally on its first coordinate sample R from the regular conditional distribution $P(dr \mid U)$. This produces a coupling of the full chart laws with exactly the stated cost. The optimal-coupling existence and conditional distribution used here hold for finite-second-moment laws on Euclidean spaces. The minimum over Q is attained precisely when $Q_U = P_U$, and the map $u \mapsto J_y(u,0)$ is one-to-one. This proves uniqueness and the second identity. Finally, H_y is a codimension-q affine subspace and has Lebesgue measure zero. Therefore $P(H_y) = 0$ whereas $Q(H_y) = 1$. □

The singularity statement is an elementary measure-theoretic consequence, not a new general principle. It applies at each fixed positive noise level. It does not say that Euclidean approximation error cannot vanish as noise decreases. In the scalar example $X \sim \mathcal{N}(0,1)$, $A = 1$, $y = 0$, and $\Sigma = \sigma^2$, the posterior variance is $\sigma^2/(1+\sigma^2)$, while

$$\mathbb{E}[R^2 \mid Y = 0] = (1+\sigma^2)^{-1} \longrightarrow 1. \tag{18}$$

The noise-dependent tube metric resolves error in units of noise standard deviation. Zero-width innovation intervals based only on hard samples have zero coverage under continuous posterior innovation marginals. Externally enlarged intervals are not covered by that statement.

For completeness, let X and $\delta(Y)$ be square-integrable. The conditional mean $\mu(Y) = \mathbb{E}[X \mid Y]$ satisfies the usual orthogonality identity: for fixed $W \succeq 0$,

$$\mathbb{E}\|X - \delta(Y)\|_W^2 = \mathbb{E}\|X - \mu(Y)\|_W^2 + \mathbb{E}\|\mu(Y) - \delta(Y)\|_W^2. \tag{19}$$

The cross term vanishes by conditioning on Y. This standard decision- theoretic identity is separate from the distributional projection result.

B Population flow matching and approximation error

Fix all conditioning variables. Specify the source as $C_0 \sim \mathcal{N}(0, I_{n-q}) \otimes \mathcal{N}(0, I_q)$, unless an experiment explicitly uses another source. Let $C_1 \sim \pi_y$ and specify their coupling. For independent-endpoint linear interpolation, $C_t = (1-t)C_0 + tC_1$ and $\dot{C}_t = C_1 - C_0$. Define

$$M = \operatorname{diag}((n-q)^{-1} I_{n-q}, q^{-1} I_q), \qquad \|z\|_M^2 = z^\top M z. \tag{20}$$

The block-normalized metric is distinct from the unweighted tube metric.

B.1 Population endpoint

Let C_t be an almost surely absolutely continuous path with $\mathbb{E}\|C_0\|^2 < \infty$ and $\mathbb{E}\int_0^1 \|\dot{C}_t\|^2 dt < \infty$. Weighted conditional least squares gives, for almost every t,

$$v^*(c,t) = \mathbb{E}[\dot{C}_t \mid C_t = c]. \tag{21}$$

Indeed, conditioning on $C_t = c$ and completing the square gives

$$\mathbb{E}[\|v - \dot{C}_t\|_M^2 \mid C_t = c] = \|v - v^*(c,t)\|_M^2 \tag{22}$$

$$+ \mathbb{E}[\|v^*(c,t) - \dot{C}_t\|_M^2 \mid C_t = c]. \tag{23}$$

The second term is independent of v. Positive fixed block weights therefore do not change the minimizer.

For a compactly supported smooth test function f, differentiation under the expectation yields

$$\frac{d}{dt}\mathbb{E} f(C_t) = \mathbb{E}[\nabla f(C_t)^\top \dot{C}_t] = \int \nabla f(c)^\top v^*(c,t) p_t(c)\, dc. \tag{24}$$

Thus the path marginals solve the weak continuity equation. Assume the velocity is measurable in time, locally Lipschitz in space with appropriate integrable bounds, has at most integrable linear growth, and that the associated continuity equation has a unique solution in the class under consideration. These assumptions ensure a nonexplosive unique flow on $[0,1]$ and identify its marginals with p_t. Hence its endpoint is π_y, and decoding by J_y returns P_y. This is the usual flow- matching argument, applied after an invertible affine change of variables (Lipman et al., 2023). Local Lipschitzness without a growth or nonexplosion condition would be insufficient.

B.2 A bound with explicit metric constants

Proposition 4 (Reference-path velocity error)**.** *Let X_t solve $\dot{X}_t = v^*(X_t, t)$ with law p_t, and let Z_t solve $\dot{Z}_t = v_\theta(Z_t, t)$ from the same source distribution. Assume both flows exist on $[0,1]$, and*

$$\|v_\theta(z,t) - v_\theta(x,t)\|_M \le L(t)\|z - x\|_M, \qquad L \in L^1([0,1]), \quad L \ge 0. \tag{25}$$

Set $e(t)^2 = \mathbb{E}_{p_t}\|v_\theta - v^*\|_M^2$ *and suppose* $e \in L^2([0,1])$. *If* q_1 *is the law of* Z_1, *then*

$$W_{2,M}(q_1, p_1) \le \int_0^1 \exp\left(\int_t^1 L(s)\,ds\right) e(t)\,dt. \tag{26}$$

In particular, if $L(t) \le L$ *and* $\int_0^1 e(t)^2 dt \le \epsilon^2$,

$$W_{2,M}(q_1, p_1) \le K_L \epsilon, \qquad K_L = \left(\frac{e^{2L}-1}{2L}\right)^{1/2}, \quad K_0 = 1. \tag{27}$$

For the decoded distributions, this implies

$$W_{2,\mathrm{tube}}((J_y)_\# q_1, P_y) \le \sqrt{\max(n-q, q)} K_L \epsilon, \tag{28}$$

$$W_2((J_y)_\# q_1, P_y) \le \max\left\{\sqrt{n-q}, \frac{\sqrt{q}}{s_{\min}(B)}\right\} K_L \epsilon. \tag{29}$$

Proof. Couple the flows with $Z_0 = X_0$. Their difference satisfies, in integral form,

$$\|Z_t - X_t\|_M \le \int_0^t L(s)\|Z_s - X_s\|_M\,ds + \int_0^t \|v_\theta(X_s, s) - v^*(X_s, s)\|_M\,ds. \tag{30}$$

Gronwall's inequality bounds the terminal difference by the second integrand weighted by $\exp(\int_s^1 L)$. Taking the L^2 norm over the common initial draw and using Minkowski's integral inequality proves (26). Cauchy–Schwarz gives K_L. For any vector z, $\|z\| \le \lambda_{\min}(M)^{-1/2}\|z\|_M$ and $\|Dz\| \le \|DM^{-1/2}\|_{\mathrm{op}}\|z\|_M$. The orthogonal blocks of D give $\|DM^{-1/2}\|_{\mathrm{op}} = \max\{\sqrt{n-q}, \sqrt{q}/s_{\min}(B)\}$. Apply these inequalities to the same terminal coupling. □

The proposition is an approximation-error bound, not a finite-sample learning rate. Its error is integrated under the exact path marginals; the empirical regression loss is not automatically an upper bound on this quantity. The conditioning theorem below does not establish a noise-uniform Lipschitz bound for an arbitrary learned velocity. These qualifications are essential when applying general flow-matching error arguments such as those of Benton et al. (2024).

C Curvature and the exactly solvable Gaussian case

Let $\phi = -\log p_0$ be twice differentiable and satisfy $mI \preceq \nabla^2\phi \preceq LI$ with $0 < m \le L < \infty$. The chart posterior potential, up to a constant, is $\Psi_c(c) = \phi(Dc + d_y) + \|r\|^2/2$, so

$$\nabla^2\Psi_c = D^\top \nabla^2\phi D + \mathrm{diag}(0, I_q). \tag{31}$$

C.1 Conditioning bounds and their scope

For orthogonal sampling with noise standard deviation σ, $S = \sigma^{-1}I_q$ and $\|D(a,b)\|^2 = \|a\|^2 + \sigma^2\|b\|^2$. Consequently

$$m\|a\|^2 + (1+m\sigma^2)\|b\|^2 \le (a,b)^\top \nabla^2\Psi_c(a,b) \le L\|a\|^2 + (1+L\sigma^2)\|b\|^2, \tag{32}$$

which proves

$$\kappa(\nabla^2\Psi_c) \le \frac{\max\{L, 1+L\sigma^2\}}{\min\{m, 1+m\sigma^2\}}. \tag{33}$$

The image-coordinate Hessian is $\nabla^2\phi + \sigma^{-2}P$. For $0 < q < n$, a measured unit vector and an unmeasured unit vector imply

$$\kappa(\nabla^2\phi + \sigma^{-2}P) \ge \frac{m + \sigma^{-2}}{L}. \tag{34}$$

This comparison uses a nonempty null space. With full-rank orthogonal sampling, all directions receive the same large curvature and the condition number need not diverge.

The same calculation extends to arbitrary fixed-rank linear operators:

$$\kappa(\nabla^2 \Psi_c) \le \frac{\max\{L, 1 + L/s_{\min}(B)^2\}}{\min\{m, 1 + m/s_{\max}(B)^2\}}. \tag{35}$$

If $\Sigma = \sigma^2 \Sigma_0$ for fixed $\Sigma_0 \succ 0$ and fixed A, the nonzero singular values of B equal those of $\Sigma_0^{-1/2} A$ divided by σ. Thus the bound remains finite as $\sigma \downarrow 0$. This is a posterior-curvature calculation. It does not establish the complexity of learning or integrating an unspecified transport, and the assumed strong log-concavity does not hold for general multimodal image priors.

For a prior with only a two-sided log-Hessian bound $\|\nabla^2 \log p_0\|_{\mathrm{op}} \le L_0$, the score identities remain

$$\nabla_x \log p_y(x) = \nabla_x \log p_0(x) + A^\top \Sigma^{-1}(y - Ax), \tag{36}$$

$$\nabla_c \log \pi_y(c) = D^\top \nabla_x \log p_0(J_y(c)) - (0, r). \tag{37}$$

The image likelihood contribution is $-V_q S r$, not $-V_q S^{-1} r$. Along a measured eigenvector of $A^\top \Sigma^{-1} A$ with eigenvalue λ, the image score Hessian has operator norm at least $\lambda - L_0$ when this is positive. In the chart,

$$\|\nabla_c^2 \log \pi_y\|_{\mathrm{op}} \le L_0 \max\{1, s_{\min}(B)^{-2}\} + 1. \tag{38}$$

Scores and CFM velocities are different objects; one cannot substitute this score-Hessian estimate into a CFM solver-complexity argument.

C.2 Gaussian posterior and independent-endpoint CFM

Let $X \sim \mathcal{N}(\mu_0, K_0)$ with $K_0 \succ 0$. Completing the square in (14) gives the tube precision and mean

$$H_c = D^\top K_0^{-1} D + \mathrm{diag}(0, I_q), \quad G = H_c^{-1}, \quad \mu = G D^\top K_0^{-1}(\mu_0 - d_y). \tag{39}$$

In block form the diagonal measured precision is $I_q + S^{-1} V_q^\top K_0^{-1} V_q S^{-1}$; the off-diagonal block is $V_0^\top K_0^{-1} V_q S^{-1}$. The inverse prior covariance is required in each block.

Take independent $C_0 \sim \mathcal{N}(0, I_n)$ and $C_1 \sim \mathcal{N}(\mu, G)$. Their linear interpolant has mean $t\mu$ and covariance $G_t = (1-t)^2 I + t^2 G$. Gaussian conditioning gives

$$\mathbb{E}[C_1 \mid C_t = c] = \mu + tG G_t^{-1}(c - t\mu), \tag{40}$$

$$\mathbb{E}[C_0 \mid C_t = c] = (1-t) G_t^{-1}(c - t\mu), \tag{41}$$

and hence

$$v^*(c, t) = \mu + A_t(c - t\mu), \qquad A_t = [tG - (1-t)I] G_t^{-1}. \tag{42}$$

The matrices G_t commute with G and their derivatives. Differentiating $G_t^{1/2}$ therefore verifies the exact flow map

$$\Phi_t(z) = t\mu + G_t^{1/2} z. \tag{43}$$

This is a closed-form Gaussian oracle. It is not evidence that a learned field can be sampled in one evaluation, nor that an arbitrary numerical solver accepts a full step.

If the eigenvalues of G lie in $[a, b] \subset (0, \infty)$, then $\lambda_{\min}(G_t) \ge a/(1+a)$ and

$$\sup_{t \in [0,1]} \|A_t\|_{\mathrm{op}} \le \max\{1, b\}(1+a)/a. \tag{44}$$

The curvature bound gives such fixed a, b in tube coordinates as noise decreases under its stated prior assumptions.

For a scalar target eigenvalue $\lambda > 0$, the corresponding coefficient is

$$a_\lambda(t) = \frac{t\lambda - (1-t)}{(1-t)^2 + t^2 \lambda}. \tag{45}$$

Set $s = t - 1/(1+\lambda)$ and $k = \sqrt{\lambda}/(1+\lambda)$. Completing the square gives $a_\lambda(t) = s/(s^2 + k^2)$. On the real line its absolute maximum is $1/(2k)$, attained at $s = \pm k$. If $\lambda \le 1$, the negative extremum lies in $[0, 1]$; if $\lambda \ge 1$, the positive extremum lies in $[0, 1]$. Hence

$$\sup_{t \in [0,1]} |a_\lambda(t)| = \frac{1 + \lambda}{2\sqrt{\lambda}}. \tag{46}$$

For measured posterior variance $\lambda = \Theta(\sigma^2)$, this is $\Theta(\sigma^{-1})$, despite the posterior score curvature being of order σ^{-2}. Equation (46) also provides a direct spectral formula for the largest spatial derivative of the independent-endpoint Gaussian flow.

The distinction depends on the probability path. A covariance-linear path $\widehat{G}_t = (1-t)I + tG$ instead has centered velocity $(G-I)\widehat{G}_t^{-1}c/2$, whose terminal derivative can be of order σ^{-2}. Results for these two paths cannot be interchanged. For an OT coupling $C_1 = \mu + G^{1/2}C_0$, trajectories are straight and one Euler step at time zero gives the endpoint. This last statement uses the known Gaussian optimal map and is not a property created by the chart.

C.3 Why no universal NFE theorem is claimed

The number of evaluations of an adaptive solver depends on the actual path, temporal derivatives, amplitudes, norm, tolerance, controller, and rejections. An upper estimate on local error provides a sufficient small step; it does not provide a necessary upper bound on every accepted step. For example, the Gaussian case $G = I$, $\mu = 0$ has $v(c, 0) = -c$ but $\Phi_1(z) = z$. A full Euler step returns zero, so the availability of (43) does not imply one-step numerical integration. Conversely, a large linear derivative can coexist with a stationary zero trajectory and zero local error. Spatial Lipschitz constants alone also do not control rapidly oscillating time-dependent translations.

There is an additional experimental control. If $x = J_y(c)$, define $v_x(x,t) = Dv_c(C_y(x), t)$ and transform the source accordingly. Affine- equivariant Runge–Kutta updates then follow the same physical trajectories. Using the pulled-back local-error norm produces the same accept/reject decisions in exact arithmetic. A measured difference in cost must therefore be attributed to a stated change in source, path, field approximation, error norm, or finite-precision behavior, rather than to an unexplained coordinate advantage.

D Perturbations of the noise model and chart

D.1 A posterior sensitivity bound

For precision matrices $K, \widetilde{K} \succ 0$, define $\ell_K(x) = \exp[-(Ax-y)^\top K(Ax-y)/2]$, $Z_K = \mathbb{E}_{p_0}\ell_K(X) > 0$, and $P_K(dx) = Z_K^{-1}p_0(x)\ell_K(x)dx$. If $\mathbb{E}_{p_0}\|AX-y\|^2 < \infty$, then

$$\mathrm{TV}(P_K, P_{\widetilde{K}}) \le \min\left\{1, \frac{\|K-\widetilde{K}\|_{\mathrm{op}}}{2Z_K}\mathbb{E}_{p_0}\|AX-y\|^2\right\}. \tag{47}$$

To prove it, put $\Delta = \mathbb{E}_{p_0}|\ell_K - \ell_{\widetilde{K}}|$. The triangle inequality and $|Z_K - Z_{\widetilde{K}}| \le \Delta$ give $\mathrm{TV}(P_K, P_{\widetilde{K}}) \le \Delta/Z_K$. Since $|e^{-s} - e^{-t}| \le |s-t|$ for $s, t \ge 0$, $\Delta \le \|K-\widetilde{K}\|_{\mathrm{op}}\mathbb{E}_{p_0}\|AX-y\|^2/2$. If $K = W^\top W$ and $\widetilde{K} = \widetilde{W}^\top\widetilde{W}$,

$$\|K-\widetilde{K}\|_{\mathrm{op}} \le (\|W\|_{\mathrm{op}} + \|\widetilde{W}\|_{\mathrm{op}})\|W-\widetilde{W}\|_{\mathrm{op}}. \tag{48}$$

The evidence denominator and moment assumption make the scope explicit. This result compares two Bayesian likelihoods. It is not the error from applying a wrong inverse to otherwise correct chart samples.

A Gaussian likelihood need not make an underdetermined posterior sub-Gaussian: an independent Cauchy prior in a null direction remains Cauchy after conditioning. Boundedness and Lipschitzness of the prior density do not exclude this example.

D.2 Approximate coordinates with the exact likelihood

Every invertible affine map is an exact coordinate system if its full transformed likelihood and Jacobian are retained. Approximation of a preferred basis by itself therefore causes no posterior bias. It may, however, degrade conditioning.

Let P be an orthogonal rank-q projector and let $[\widetilde{V}_0, \widetilde{V}_q]$ be orthogonal with

$$\|\widetilde{V}_q\widetilde{V}_q^\top - P\|_{\mathrm{op}} \le \varepsilon. \tag{49}$$

Use $\widetilde{D} = [\widetilde{V}_0, \sigma\widetilde{V}_q]$ and an arbitrary fixed offset, and retain the exact Gaussian likelihood. The transformed likelihood Hessian is

$$\sigma^{-2}\widetilde{D}^\top P\widetilde{D} = \begin{pmatrix} \sigma^{-2}\widetilde{V}_0^\top P\widetilde{V}_0 & \sigma^{-1}\widetilde{V}_0^\top P\widetilde{V}_q \\ \sigma^{-1}\widetilde{V}_q^\top P\widetilde{V}_0 & \widetilde{V}_q^\top P\widetilde{V}_q \end{pmatrix}. \tag{50}$$

Because $\|P\widetilde{V}_0\|_{\mathrm{op}} \le \varepsilon$ and $\|(I-P)\widetilde{V}_q\|_{\mathrm{op}} \le \varepsilon$, its difference from $\mathrm{diag}(0, I_q)$ has norm at most

$$\delta = \varepsilon^2/\sigma^2 + 2\varepsilon/\sigma + \varepsilon^2. \tag{51}$$

The first and last terms bound the diagonal blocks, and the middle term is a conservative bound for the two off-diagonal blocks. Meanwhile, $\|\widetilde{D}(a,b)\|^2 = \|a\|^2 + \sigma^2\|b\|^2$, so the prior curvature bounds remain valid. Rayleigh quotients therefore give

$$\kappa(\nabla^2\widetilde{\Psi}_c) \le \frac{\max\{L, 1+L\sigma^2\} + \delta}{\min\{m, 1+m\sigma^2\} - \delta}, \tag{52}$$

provided the denominator is positive. Controlling the subspace error relative to σ is necessary for this perturbative guarantee. No total-variation guarantee follows from projector error alone.

If an approximate decoder is applied to fixed chart samples instead, a direct coupling gives the valid Wasserstein bound

$$W_2((Dc+d)_{\#}\pi, (\widetilde{D}c+\widetilde{d})_{\#}\pi) \le \left(\mathbb{E}_\pi\|(D-\widetilde{D})C + d - \widetilde{d}\|^2\right)^{1/2}. \tag{53}$$

This does not imply the analogous inequality in total variation. Practical randomized factorizations require rank, accuracy, and operator- cost assumptions beyond this calculation (Halko et al., 2011).

E Calibration statements and their limitations

Freeze the score function and all trained parameters independently of calibration. Conditional on this fixed score construction, let the calibration scores $S_1, \dots, S_N$ and the independent test score S_{N+1} be iid with distribution function F. Put $k = \lceil (N+1)(1-\alpha) \rceil$ and $\widehat{q} = S_{(k)}$, with $\widehat{q} = +\infty$ if $k = N+1$. Exchangeability of calibration and test scores gives

$$\Pr(S_{N+1} \le \widehat{q}) \ge 1 - \alpha. \tag{54}$$

This probability averages over calibration and test draws (Angelopoulos & Bates, 2021). Conditional on calibration, the population coverage is $F(\widehat{q})$, not the empirical calibration CDF.

The Dvoretzky–Kiefer–Wolfowitz–Massart inequality gives

$$\Pr\left(\sup_s |F_N(s) - F(s)| > \epsilon\right) \le 2e^{-2N\epsilon^2}. \tag{55}$$

Since $F_N(\widehat{q}) \ge 1-\alpha$ for a finite conformal order statistic,

$$\Pr\{F(\widehat{q}) \ge 1 - \alpha - \epsilon\} \ge 1 - 2e^{-2N\epsilon^2}. \tag{56}$$

Thus $N \ge \log(2/\eta)/(2\epsilon^2)$ is a sufficient calibration sample size for a shortfall at most ϵ with probability at least $1-\eta$ (Massart, 1990). This is not a statement about the empirical coverage fraction in a finite test set.

For continuous F and $1 \le k \le N$, a sharper exact calculation follows from the probability integral transform:

$$F(S_{(k)}) \sim \mathrm{Beta}(k, N+1-k). \tag{57}$$

Indeed, the transformed calibration scores are iid uniform variables, whose kth order statistic has this beta density. If $k = N+1$, the $+\infty$ fallback instead has population coverage one. At $N = 64$ and $\alpha = 0.1$, $k = 59$; its fifth percentile is approximately 0.842730. These are analytic order-statistic calculations, not additional test-set observations. For comparison, the conservative two-sided DKW bound at 95% confidence has $\epsilon \simeq 0.169763$, and guaranteeing $\epsilon = 0.05$ by that bound requires $N \ge 738$.

E.1 GAUSSIAN CONDITIONAL ELLIPSOIDS

For jointly Gaussian (U, R) with mean (μ_u, μ_r) and positive definite joint covariance, write its blocks as $G_{uu}, G_{ur}, G_{ru}, G_{rr}$,

$$\mu_{r|u} = \mu_r + G_{ru}G_{uu}^{-1}(u - \mu_u), \tag{58}$$

$$G_{r|u} = G_{rr} - G_{ru}G_{uu}^{-1}G_{ur}. \tag{59}$$

The covariance is independent of u, but the conditional mean generally is not. Standardizing this Gaussian gives

$$\Pr\left\{(R - \mu_{r|u})^\top G_{r|u}^{-1}(R - \mu_{r|u}) \le \chi^2_{q,1-\alpha} \mid U = u\right\} = 1 - \alpha. \tag{60}$$

Replacing the Mahalanobis quadratic form by an unweighted squared norm is generally incorrect. Gaussian fiber and residual blocks are independent only when $G_{ur} = 0$. These are standard Gaussian identities and do not establish a coordinate-specific distribution-free coverage advantage.

For a fixed observable partition into K groups, the same conformal procedure can be applied within each group. Conditional on group counts, the iid within-group DKW bounds and a union bound show that groups with

$$N_k \ge \frac{1}{2\epsilon^2}\log\frac{2K}{\eta} \tag{61}$$

simultaneously have population coverage at least $1 - \alpha - \epsilon$ with probability at least $1 - \eta$. The groups and score construction must be fixed independently of calibration scores; insufficiently populated groups require an explicitly reported fallback. To assign a group at deployment it must be a function of observed information. The true latent fiber coordinate $V_0^\top X$ is generally unavailable in an inverse problem. Binning it is a retrospective diagnostic unless an observable assignment rule is supplied. No universal interval-width bound is claimed.

F QUALIFIED EXTENSIONS BEYOND THE IMPLEMENTED METHOD

The following calculations clarify the assumptions required by two possible extensions. They are not evidence of an implemented or evaluated sampler.

F.1 A LOCAL NON-GAUSSIAN LIKELIHOOD CHART

Let $\ell(x; y) = \log L(x; y)$ denote the log likelihood. Suppose it is constant along the V_0 directions in the stated neighborhood and can be written $\ell(x; y) = \overline{\ell}(V_q^\top(x - \widehat{x}))$. Assume $\nabla\overline{\ell}(0) = 0$ and $H = -\nabla^2\overline{\ell}(0) \succ 0$. Let $a = \lambda_{\min}(H)$, $b = \lambda_{\max}(H)$, and suppose the Hessian of $-\overline{\ell}$ is K-Lipschitz on a radius-ρ neighborhood. Use the local chart

$$x = \widehat{x} + V_0 u + V_q H^{-1/2} r. \tag{62}$$

Taylor's theorem with Lipschitz Hessian gives

$$-\overline{\ell}(z) = -\overline{\ell}(0) + \tfrac{1}{2}z^\top H z + R(z), \qquad |R(z)| \le K\|z\|^3/6. \tag{63}$$

The chart density on this neighborhood is therefore proportional to

$$p_0(\widehat{x} + V_0 u + V_q H^{-1/2} r)e^{-\|r\|^2/2 - R(H^{-1/2}r)}\det(H)^{-1/2}. \tag{64}$$

This is an exact expression with a bounded remainder, not a global Gaussian factorization. In chart coordinates, the negative-log-likelihood Hessian remainder has norm at most $K\rho/a$. Under the prior bounds $mI \preceq \nabla^2\phi \preceq LI$, combining the two contributions gives

$$\kappa_{\text{loc}} \le \frac{\max\{L, 1 + L/a\} + K\rho/a}{\min\{m, 1 + m/b\} - K\rho/a}, \tag{65}$$

when the denominator is positive and throughout the stated neighborhood. The smallest prior contribution in the measured block uses b, not a.

Twice continuous differentiability alone does not supply a cubic remainder. Nor does a Hessian null direction at one point imply that the likelihood is constant in that direction: $\ell(x_1, x_2) = -x_1^2/2 - x_2^4$ is a counterexample. These assumptions are why no unrestricted extension to all smooth log-concave likelihoods is asserted.

F.2 THE DIFFUSION DRIFT REQUIRED TO PRESERVE A FLOW PATH

Let a smooth positive density path satisfy $\partial_t p_t = -\nabla \cdot (p_t v_t)$. Let a_t be a positive definite, space-independent diffusion covariance and assume the usual existence, nonexplosion, and uniqueness conditions for the following Fokker–Planck equations. The SDE

$$dC_t = \left[v_t(C_t) + \tfrac{1}{2} a_t \nabla \log p_t(C_t)\right] dt + a_t^{1/2} dW_t \tag{66}$$

has the same marginals p_t, when initialized from p_0. Substitution into its Fokker–Planck equation proves this: the score contribution is $-\nabla \cdot (a_t \nabla p_t)/2$ and cancels the diffusion term. One may choose $a_t = \gamma(t)^2 M^{-1}$; the compensating drift must still be present. This construction belongs to the standard flow/diffusion family developed in stochastic-interpolant theory (Albergo et al., 2025).

Adding noise to the unmodified drift does not preserve the endpoint. For the stationary path $p_t = \mathcal{N}(0, I)$ and $v_t = 0$, constant uncorrected diffusion produces covariance $I + a$ at time one. For the increasing reverse clock $s = 1 - t$, the corrected reverse drift is

$$-v_{1-s}(c) + \tfrac{1}{2} a_{1-s} \nabla \log p_{1-s}(c). \tag{67}$$

The minus sign on the forward velocity is required by this clock convention.

A valid error statement can instead use relative entropy of path laws. Let P and Q be processes with the same initial law, the same positive diffusion covariance a_t, and respective drifts $b_t, \widehat{b}_t$. Assume the absolute-continuity and integrability conditions for Girsanov's formula, including the needed exponential-martingale condition. Then

$$\mathrm{KL}(P\|Q) = \frac{1}{2}\mathbb{E}_P \int_0^1 \|a_t^{-1/2}(b_t(C_t) - \widehat{b}_t(C_t))\|^2 dt. \tag{68}$$

Projection to the endpoint and data processing bound $\mathrm{KL}(p_1\|q_1)$ by the same expression. The orientation is explicit and the error is measured under the reference process. A log-Sobolev inequality for the final target alone is insufficient to assert contraction along a changing target path, and no untested robustness gain over the deterministic sampler is claimed here.

G EXPERIMENTAL PROTOCOLS AND ADDITIONAL RESULTS

The studies below serve different purposes. We report acquisition models, priors, splits, and computational budgets separately because the DCT, pixel-mask, and Fourier tasks are not directly comparable. Per-case metrics and aggregation scripts accompany the results.

G.1 GAUSSIAN STUDY

Every main configuration is the Cartesian product of three ranks, three precision condition numbers, five noise levels, and three seeds. The prior and operator orientation are drawn once per configuration, and the exact posterior is computed in float64. Independent-endpoint CFM has covariance $(1-t)^2 I + t^2 G$; rather than a linear interpolation of endpoint covariances. The adaptive controller starts with step size 0.01, uses safety factor 0.9 and growth limits $[0.2, 5]$, and counts all rejected right-hand-side calls. Affine probes use the same canonical source basis across coordinates.

The error norm for tube-controlled rows is the pulled-back block-normalized metric. The image-error row instead uses a componentwise image criterion and therefore does not target identical physical accuracy. The null-prior-matched source replaces the identity null-block covariance by the prior null marginal. The exact Gaussian OT row uses the known optimal endpoint coupling; its straight trajectories permit exact algebraic evaluation, but the specified adaptive controller still takes four accepted steps and eight evaluations. That behavior does not characterize a trained OT-CFM model.

The suite contains 810 main integrations, 360 tolerance-sensitivity integrations, and 90 instrumented allocation measurements. Solver allocation measurements exclude preconstructed matrices and native-library workspace that the tracer cannot observe. Shared-CPU wall time is descriptive. The maximum relative error in the independently checked transformed-precision identity is 7.1×10^{-12}. Near-zero Gaussian Wasserstein values can be limited by matrix-square-root cancellation; endpoint RMS and covariance errors are also recorded.

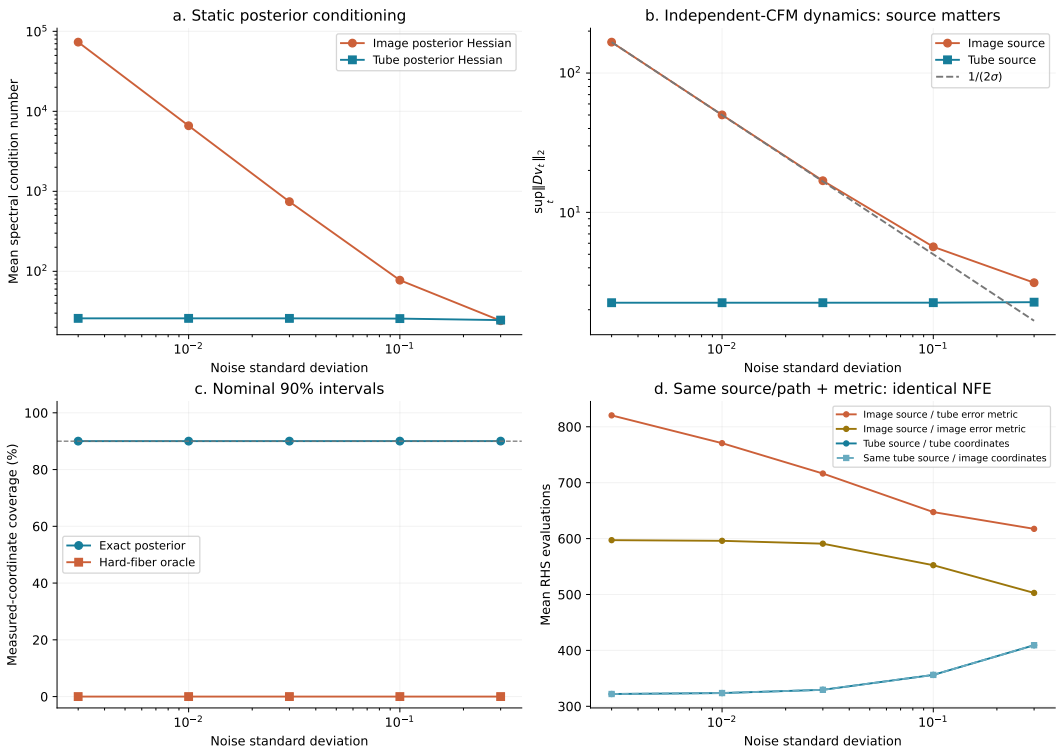


Figure 2: Gaussian theory and controls. The noise-dependent tube metric resolves measured errors in noise units. A transformed representation with identical physical source, field, and error norm preserves evaluation counts. Source changes must be evaluated separately.

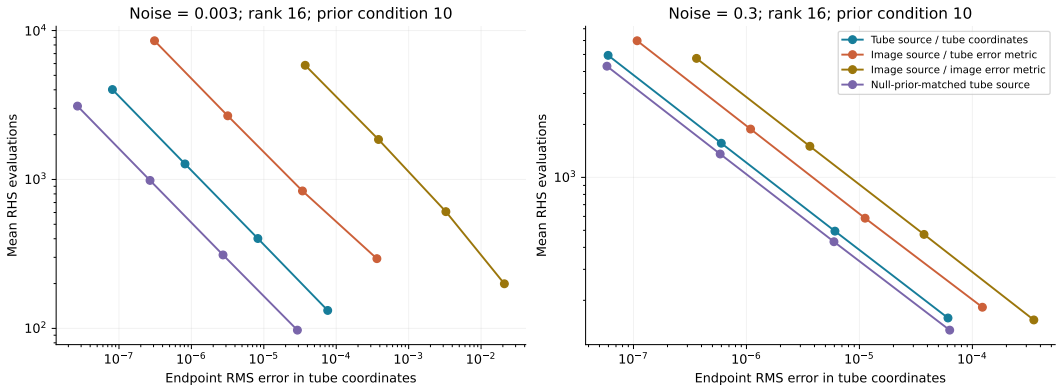


Figure 3: Numerical error versus evaluation count, including tolerance sensitivity. Different error criteria can produce different terminal accuracies; lower NFE without an accuracy comparison is insufficient.

G.2 Fitted-prior study on optical digits

Images are normalized by dividing their recorded integer counts by 16. The complete UCI dataset has a writer-based split, but the locally distributed 1,797-image subset copies its test cohort. Writer identifiers are unavailable. Our image-level resampling therefore tests within-cohort generalization and does not certify unseen-writer performance. Labels are used only for split stratification, not as conditioning variables or to fit mixture components.

The Gaussian prior uses Ledoit–Wolf covariance shrinkage. The mixture search spans component counts $\{1, 5, 10\}$ and diagonal regularization $\{0.001, 0.01\}$; the best validation-likelihood model is

frozen for evaluation. All three fitting seeds choose ten components and regularization 0.001. The observation masks are fixed low-frequency or random subsets of an orthonormal two-dimensional DCT. The same noise draw is used for all methods and fitting seeds within each measurement condition. Posterior means are analytic; sample quantiles use 256 draws. Posterior samples are not clipped to the pixel range.

Table 5: Fitted continuous priors on real optical-digit images. Measured coverage refers to marginal DCT-coefficient intervals. Confidence intervals cluster repeated settings and seeds by held-out image, conditional on the fitted models and chosen observations.

Method	PSNR ↑	Obs. cov.	Pixel cov.	CRPS ↓
Gaussian posterior	17.11	89.5%	92.6%	0.0741
Gaussian + hard	16.68	0.0%	85.8%	0.0829
GMM posterior	19.38	89.0%	91.5%	0.0566
GMM + hard	18.62	0.0%	80.7%	0.0692
Image affine FM (16 NFE)	17.11	96.0%	93.9%	0.0762
Tube affine FM (16 NFE)	17.11	89.5%	93.4%	0.0754
Image affine FM (64 NFE)	17.11	91.8%	92.7%	0.0743
Tube affine FM (64 NFE)	17.11	89.5%	92.7%	0.0742

The bootstrap uses 2,000 resamples of the 360 held-out images after averaging their repeated settings and fitting seeds. These intervals do not represent variation over independent writer cohorts or all possible training sets. Paired contrasts, seed SD, coverage at 50/80/90/95%, widths, CRPS, energy scores, and all image-level rows are included in the results archive. The empirical CRPS uses the sample-average absolute target error minus half the sample-pair absolute difference. Its sorted-sample implementation is independently checked against the direct double sum.

At noise 0.02, the Gaussian affine-flow endpoint errors in tube W_2 are 10.895 for the image source and 0.1205 for the tube source at 16 evaluations, decreasing to 1.111 and 0.00911 at 64 evaluations. These averages span four masks and three fitted priors. The exact posterior mean is shared across these affine comparisons, so their point PSNR is identical. Their difference is numerical covariance error. Standard normal sources are defined in their respective coordinates, so this is a source-and-coordinate comparison; the genuinely shared-path control is the Gaussian experiment above.

G.3 Learned conditional flows on real digit images

This experiment separates the learned method from analytic Gaussian posterior calculations. We use the 1,797 real 8×8 images packaged with `sklearn.datasets.load_digits`, which reproduces the original UCI optical digits test subset. The images are divided by 16 and repartitioned into 970 optimization, 108 validation, 359 calibration, and 360 test images, stratified by digit. These are image-level splits: writer identifiers are unavailable, so writer independence cannot be claimed. Dataset indices and the normalized-data SHA256 are provided with the experiment artifacts.

A fixed seeded mask observes 32 of 64 pixels. The model sees $y = x_{\mathrm{obs}} + \sigma\epsilon$, with independent $\epsilon \sim \mathcal{N}(0, I_{32})$ and noise levels $\sigma \in \{0.05, 0.10, 0.20\}$. Training samples these levels uniformly and resamples observation noise on every update. Calibration and test observations are held fixed across methods and training seeds. Digit labels enter only the stratification and subgroup analysis.

Each conditional velocity is a two-hidden-layer tanh MLP with width 192 and 69,952 parameters. Inputs are the 64-dimensional flow state, 32 observations, time, eight sinusoidal time features with frequencies $1, 2, 4, 8$, and $\sigma/0.2$. We train for 10,000 Adam updates of 128 tuples, using $(\beta_1, \beta_2) = (0.9, 0.999)$, numerical constant 10^{-8}, and a cosine learning rate from 10^{-3} to 10^{-4}. The independent training seeds are 11, 29, and 47. A frozen set of 2,048 noised tuples from validation images selects the checkpoint with minimum validation loss among update 1 and then every 250 updates. No calibration or test outcomes select models.

TubeFM uses state $(x_{\mathrm{unobs}}, (x_{\mathrm{obs}} - y)/\sigma)$ and an independent standard Gaussian source. Both blocks have dimension 32, so its two block-normalized losses sum to twice the coordinate-wise MSE. The standard-source image-CFM baseline uses state $(x_{\mathrm{unobs}}, x_{\mathrm{obs}})$ and a standard Gaussian source in image coordinates, with the same architecture and update budget. A second image-CFM baseline matches TubeFM's physical source by setting observed source pixels to $y + \sigma z_{\mathrm{obs}}$. It still minimizes

image-coordinate MSE. This control distinguishes source construction from the combined effect of coordinate scaling and loss weighting; neither comparison isolates a coordinate change independently of finite-capacity optimization. A hard-projection ablation assigns TubeFM's observed sample pixels to y after sampling and does not train another network.

Each model generates 64 samples per calibration and test case with 64 fixed Heun steps, requiring 128 velocity evaluations per draw. No reconstructions or samples are clipped. PSNR uses range one; NMSE is computed per image. Central 90% sample intervals use linearly interpolated 5th and 95th percentiles. Innovation intervals are computed after subtracting y and dividing by σ. These are empirical reconstruction intervals evaluated against the held-out true image, not comparisons with an available exact posterior law. All three training seeds and all three noise settings are retained in the reported tables.

For nonprojected models, calibration follows the stated 90/90 score: the 58th ordered absolute pixel error divided by sample SD, with SD floor 10^{-4}, followed by the 324th ordered score among 359 calibration cases. Its standard case-marginal guarantee requires exchangeability and does not imply all-pixel, subgroup-conditional, or exact-posterior coverage. We report observed finite-test coverage even when it is below the nominal value. The exactly degenerate measured bands of hard projection are not assigned a finite rescaling that would silently introduce nonzero posterior support.

We verify the handwritten backpropagation by a float64 central directional finite difference (relative error 8.24×10^{-10}). For each method, training seed, and noise setting, a 128-step Heun rerun uses the same source draws on the first 12 fixed test cases, assessing discretization without selecting checkpoints. Across the 27 method–seed–noise configurations, the maximum endpoint RMSE between the 64- and 128-step solutions is 1.0×10^{-4}, and the maximum posterior-mean RMSE is 3.0×10^{-5}. These checks measure sensitivity to step refinement, rather than error against an exact learned-flow solution. Paired bootstrap intervals resample 360 test-image IDs 10,000 times after averaging each image's metric across the three training seeds; they are conditional on these seeds and image-level exchangeability. Secondary comparisons are exploratory and not multiplicity adjusted.

The low resolution and fixed observation mask make this a controlled test of learned source and coordinate effects. Generalization to other architectures, unseen writers, and high-resolution inverse problems is not established.

G.4 Procedural Fourier study

Procedural Fourier stress test. We generated three independently seeded datasets of 32×32 real-valued ellipse phantoms, each containing 640 training, 128 calibration, and 192 test images. A bright synthetic lesion was included with probability 0.65. Of the training images, 544 fitted a Gaussian prior and 96 selected the ridge penalty. No calibration or test images entered either fit. The prior was $\mathcal{N}(\mu, C)$ with

$$C = U_{64}\,\mathrm{diag}(\lambda_1, \ldots, \lambda_{64})U_{64}^{\mathsf{T}} + \mathrm{diag}(d), \qquad d_j \geq 10^{-5},$$

where the principal components and their variances were estimated from the centered training images, and d was the empirical diagonal variance left after rank-64 projection, plus 10^{-5}.

Each image was observed under four Cartesian masks (4×/8×; approximately equispaced/random outer lines), each retaining two fixed central lines, and three noise levels $\sigma \in \{0.01, 0.03, 0.05\}$. The Fourier transform was orthonormal, and the real and imaginary noise components each had variance σ^2, so $\mathbb{E}|\varepsilon_k|^2 = 2\sigma^2$. For real phantoms the realified measurement operator is rank deficient because of Fourier conjugate symmetry. We therefore used its compact SVD $A = U_r \Sigma_r V_r^{\mathsf{T}}$ and the sufficient observations $z = U_r^{\mathsf{T}} y$, with $B = \Sigma_r V_r^{\mathsf{T}}$. The exact fitted-prior posterior was

$$m_\sigma = \mu + CB^{\mathsf{T}}(BCB^{\mathsf{T}} + \sigma^2 I)^{-1}(z - B\mu), \quad C_\sigma = C - CB^{\mathsf{T}}(BCB^{\mathsf{T}} + \sigma^2 I)^{-1}BC.$$

The hard-feasible ablation used the same formulas at $\sigma = 0$. It enforces $Bx = z$, equivalently $Ax = U_r U_r^{\mathsf{T}} y$; it cannot enforce all inconsistent noisy complex rows. Thus this comparison should not be read as a full-row-rank complex-image MRI experiment. Zero filling and ridge regression supplied additional point baselines.

We evaluated real, unclipped reconstructions using PSNR, SSIM, NMSE, and synthetic-lesion MAE. For each test acquisition we drew 64 exact posterior samples using conditional Gaussian sampling.

Analytic marginal standard deviations were used for the main calibration analysis. On a fixed central disk, a case was successful when its interval covered at least 90% of pixels. Each calibration score was the $\lceil 0.9P \rceil$th ordered standardized absolute error among the P evaluation pixels; the interval multiplier was the 117th of 128 calibration scores, because $\lceil (128+1)0.9 \rceil = 117$. We applied the identical split-conformal procedure to both soft and hard posteriors. This guarantees marginal case-level coverage under exchangeability, not simultaneous coverage of every pixel or coverage within the lesion subgroup. The 576 independent test images were reused across 12 acquisition conditions; we report variation across the three seeds and do not treat the resulting 6,912 acquisitions as independent subjects. This experiment evaluates analytic posterior endpoints on generated objects; its role is to expose reconstruction–uncertainty trade-offs and structural failures under prior misspecification.

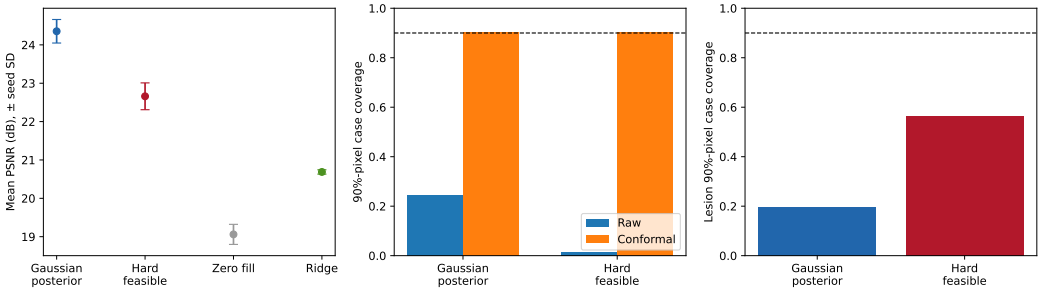


Figure 4: Procedural reconstruction and calibration. Both posteriors can be calibrated for the fixed ROI event. Lesion-region coverage remains poor, particularly for the soft Gaussian posterior, despite favorable overall image scores.

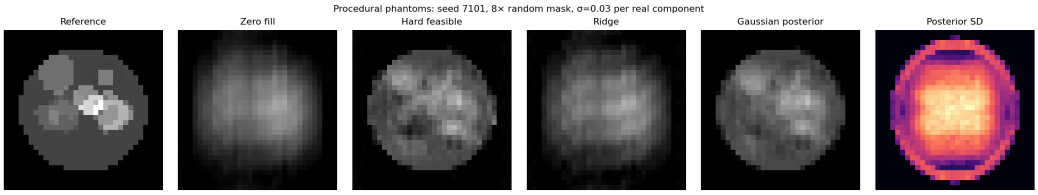


Figure 5: A retained procedural example at nominal $8\times$ acceleration. These are real-valued synthetic Fourier objects, not acquired anatomy. The reported lesion failures require evaluating structure separately from overall image similarity.

Across all settings, numerical checks verify realified forward operators, posterior conditioning, and feasibility of hard projections in the independent measured subspace. Regenerating the full-precision data and noise reproduces saved soft PSNR values to below 10^{-9} dB. The compact sample banks retain eight examples per setting, while all 192 test cases were sampled and their metrics saved. Counting repeated acquisitions as independent subjects would overstate evidence: there are 576 independent test phantoms, not 6,912 independent objects.

H ACQUIRED-MRI PILOT

Data and the custom split. Data used in this pilot were obtained from the NYU fastMRI Initiative (Zbontar et al., 2018; Knoll et al., 2020). The supplied single-coil representation is emulated from acquired multicoil knee data. We pool 973 volumes from the distributed single-coil training archive and 199 from its validation archive, recording their original archive provenance. The resulting 1,172-volume cohort is repartitioned for this custom retrospective task into 821 training, 117 validation, 117 calibration, and 117 test volumes. This reuse of the distributed partitions precludes an official fastMRI benchmark claim. Data preparation checks found no duplicate file bytes, invalid dimensions, or empty/nonfinite k-space slices. The manifest records each file's checksum and split. Reliable patient linkage was unavailable; different scans from one patient cannot be certified to occupy the same partition.

For each slice, we reconstruct a complex reference by a centered unitary inverse FFT and take a central 320×320 crop. The common normalization scale is $1.1578165868 \times 10^{-4}$, the median of the central-slice 99th magnitude percentiles from the first 32 sorted training volumes. No validation or test image sets this scale. It is a global intensity scale, not a per-image rescaling or clipping operation. Inspected training slices have 99th magnitude percentiles from 0.512 to 4.054 in normalized units. The same absolute added-noise scale therefore represents different relative noise strengths across scans.

Measurement and training protocol. The crop is Fourier transformed again to construct new retrospective observations. Each Cartesian mask retains 80 or 40 of the 320 columns, corresponding to nominal acceleration 4 or 8. It includes 26 central columns and uniformly samples the remaining retained columns without replacement. Independent Gaussian noise is added per real and imaginary component at scale 0.01, 0.03, or 0.05 after the global normalization. The reference retains acquisition noise; these simulated observations do not provide a noise-free clinical ground truth.

The completed TUBEFM model uses the native tube source, block-balanced velocity loss, and a residual CNN of width 32 and depth 4, totaling 77,122 trainable parameters. Training uses AdamW with learning rate 2×10^{-4}, weight decay 10^{-5}, batch size 2, gradient clipping at norm 1, and 50,000 updates. Volume, slice, acceleration, noise level, source draws, and interpolation time are sampled during training. The same fixed panel of sixteen validation batches is evaluated every 1,000 updates; the checkpoint with the smallest validation objective is retained. With seed 11, this minimum occurs at update 50,000. Thus completion of the allocated training schedule is not evidence of convergence. No other learned MRI configuration or training seed is evaluated.

Exploratory evaluation. The two-volume subset and its settings were fixed before evaluation. The subset consists of the first two validation volumes in the existing manifest, identified here as V1 and V2. Each contributes the first, middle, and final slice: indices $(0, 16, 33)$ for V1 and $(0, 20, 41)$ for V2. This is a deterministic convenience subset, not a representative anatomical sampling claim. The evaluation uses the selected checkpoint at 50,000 updates, 32 independent source draws, and 64 Heun steps, with draws processed in batches of two. A slice/condition has a deterministic observation seed shared by the methods. Hard projection is applied to the same generated Fourier draws; zero filling uses the same observations. These are within-model ablations and a point baseline, not independent trained comparators. No calibration factors are fitted and no held-out test volume is evaluated. The fixed checkpoint-selection panel contains 28 distinct volumes; neither pilot volume appears in that panel. Both belong to the designated validation split, and the convenience subset remains exploratory.

Metrics and aggregation. Point estimates are means of the sampled magnitude images. PSNR uses fixed normalized peak 1, and SSIM uses a uniform 7×7 window with sample-covariance correction and fixed data range 1. Images are not clipped even when their magnitudes exceed 1. Thus absolute PSNR and SSIM are not interchangeable with official benchmark scores or per-volume range conventions. NMSE divides squared magnitude error by reference magnitude energy. Pixel CRPS uses the empirical ensemble; the energy score uses independent sample pairs and is normalized by the square root of the number of pixels. Pixel metrics use the full crop, with no anatomical foreground mask; SSIM additionally omits its three-pixel boundary. Edge slices and background can therefore materially affect the scores. All these definitions are fixed in the saved evaluation implementation.

Raw pixel intervals and measured real/imaginary Fourier-component intervals use the linearly interpolated empirical 5th and 95th percentiles. Coverage averages the corresponding component indicators. We first average metrics over the three slices of a volume/condition, then give equal weight to both volumes and all six conditions. The archive contains 108 method–acquisition rows and 36 method–volume–condition rows, but only two distinct validation volumes. We report no population confidence interval or significance test. Posterior uncertainty metrics for the point baseline are omitted from our tables.

The mean PSNR difference between TUBEFM and its hard projection is 0.0261 dB; it is 0.0556 dB for V1 and -0.0035 dB for V2. Hard projection has higher mean SSIM. Pixel coverage is 79.13% and 78.79%, respectively, for the raw 32-draw percentile bands. Measured coverage is 85.30% for TUBEFM and below 0.001% after projection. The tiny nonzero hard-coverage value is consistent with floating-point equality effects when continuous noisy references are compared with collapsed intervals. No calibrated-coverage result is inferred.

Table 6: MRI pilot, separately by validation volume and averaged over the six conditions and three slices. Fixed-range PSNR differs substantially between scans. The sign of the small TubeFM–hard PSNR difference is not the same for both volumes.

Volume	Method	PSNR (dB)	NMSE	SSIM
V1	TubeFM	25.155	0.10919	0.4110
V1	Hard-projected TubeFM	25.099	0.11097	0.4186
V1	Zero-fill	24.096	0.17504	0.3857
V2	TubeFM	13.996	0.09479	0.2836
V2	Hard-projected TubeFM	13.999	0.09461	0.2866
V2	Zero-fill	13.191	0.16921	0.2945

Table 7: Descriptive posterior scores on the MRI pilot. Width is mean pixel-band width in globally normalized intensity units; lower CRPS and energy score are preferable. These small within-model differences on two validation volumes are not evidence of population superiority.

Method	Pixel CRPS	Energy score	Pixel width
TubeFM	0.07488	0.10118	0.28900
Hard-projected TubeFM	0.07501	0.10133	0.28817

Table 8: All six MRI conditions. Each entry averages two validation volumes and three slices per volume. All intervals use 32 draws; measured coverage is descriptive. No condition is selected or omitted based on its outcome.

R	σ	Method	PSNR	NMSE	SSIM	Measured cov. (%)
4	0.01	TubeFM	20.093	0.09287	0.3989	85.08
4	0.01	Hard-projected TubeFM	20.093	0.09286	0.4002	< 0.001
4	0.01	Zero-fill	19.193	0.15261	0.3969	—
4	0.03	TubeFM	19.962	0.09421	0.3834	85.23
4	0.03	Hard-projected TubeFM	19.947	0.09465	0.3909	< 0.001
4	0.03	Zero-fill	19.209	0.14909	0.3908	—
4	0.05	TubeFM	19.705	0.09822	0.3650	85.37
4	0.05	Hard-projected TubeFM	19.591	0.10202	0.3769	< 0.001
4	0.05	Zero-fill	19.093	0.14467	0.3786	—
8	0.01	TubeFM	19.341	0.10721	0.3194	85.15
8	0.01	Hard-projected TubeFM	19.342	0.10715	0.3200	< 0.001
8	0.01	Zero-fill	18.052	0.20363	0.2908	—
8	0.03	TubeFM	19.240	0.10890	0.3130	85.26
8	0.03	Hard-projected TubeFM	19.239	0.10882	0.3170	< 0.001
8	0.03	Zero-fill	18.132	0.19528	0.2921	—
8	0.05	TubeFM	19.109	0.11054	0.3038	85.70
8	0.05	Hard-projected TubeFM	19.081	0.11124	0.3106	< 0.001
8	0.05	Zero-fill	18.180	0.18749	0.2913	—

H.1 Finite-ensemble quantile coverage

The empirical percentile bands above need not have exactly 90% expected coverage even for an exact posterior sampler. The following standard order-statistic calculation makes this distinction explicit. Fix an observation and a continuous scalar posterior CDF F. Let $X_1, \dots, X_{32}$ and an independent target T be iid from F. For the implemented linear interpolation, the endpoints are

$$L = 0.45X_{(2)} + 0.55X_{(3)}, \qquad U = 0.55X_{(30)} + 0.45X_{(31)}. \tag{69}$$

Monotonicity implies $F(X_{(30)}) - F(X_{(3)}) \le F(U) - F(L) \le F(X_{(31)}) - F(X_{(2)})$. Using $\mathbb{E}F(X_{(j)}) = j/33$ therefore gives

$$\frac{27}{33} \le \Pr\{L \le T \le U\} \le \frac{29}{33}, \qquad \text{i.e., } 81.82\% \text{ to } 87.88\%. \tag{70}$$

For a uniform posterior the probability is exactly $0.9(31/33) = 84.55\%$. The probability averages over both the ensemble and the independent target. This calculation does not establish exact sampling by our network, nor does it give a finite-test guarantee for the pilot. It shows why the observed measured coverage of 85.30% cannot by itself diagnose posterior underdispersion. Model error, finite-ensemble quantile error, and variation across scans must be distinguished. A separate calibration procedure or an ensemble-size-aware evaluation is needed for a calibrated-coverage claim.

H.2 Training diagnostics and measured computation

The validation objective decreases from 1.6877 at update 1,000 to 1.5084 at update 50,000, where it reaches its minimum over the saved evaluations. Figure 6 reports the training and validation trajectories. Repeated log entries from a checkpoint restart are counted once per update. The final checkpoints include optimizer and random-number states for continuation. The terminal validation minimum leaves optimization convergence unresolved; MRI solver refinement and additional training seeds were not evaluated.

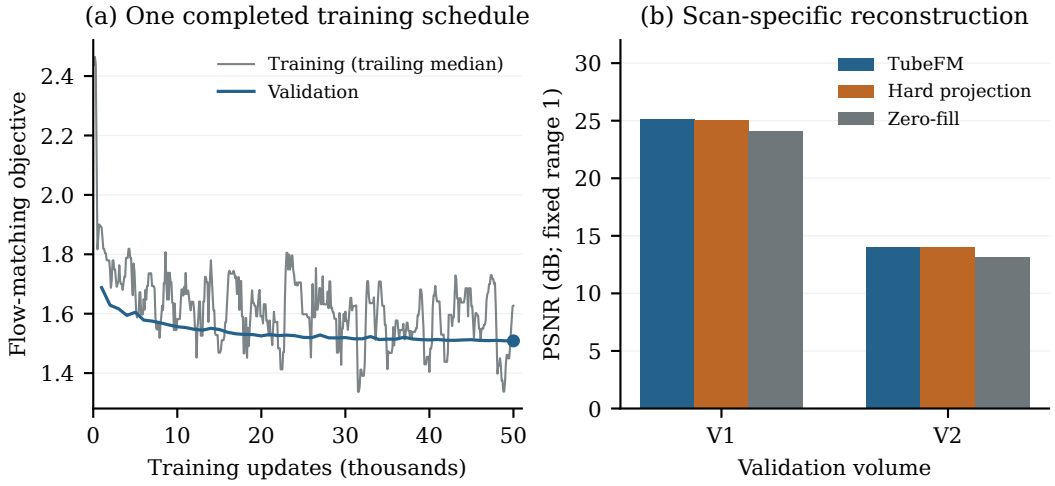


Figure 6: MRI pilot diagnostics. Left: the retained training trajectory and periodic validation objective; the training line is a trailing median over ten logged minibatches. The terminal checkpoint has the lowest observed validation objective. Right: per-volume PSNR averaged over all six conditions and three slices. V1 and V2 are two validation volumes, not repeated training seeds. No population error bars are shown.

The retained TubeFM training trajectory takes 1,625.23 seconds, excluding replayed updates, setup, and idle time. The exploratory evaluation takes 206.07 seconds; generated ensembles account for approximately 5.30 seconds per acquisition on the L40S, with the hard-projection samples reused rather than generated again. The software environment records Python 3.12.11 and PyTorch 2.7.1 with CUDA 12.8. These are observed engineering measurements, not time-to-quality comparisons.

Table 9: Short L40S profiles for the five configured representations. Each uses 20 timed training updates after five warm-up updates and three timed sampling acquisitions, with 32 draws and 128 NFE per draw. Memory is PyTorch peak allocated device memory. All five profiles use separate models after these short training sequences, not the final trained checkpoint; no accuracy-matched speed claim follows.

Configuration	ms / update	s / ensemble	Peak GiB
Fourier, matched source	34.97	5.238	0.744
Image, matched source	37.53	5.397	0.745
Image, native source	36.95	5.410	0.745
Tube, native source	37.44	5.264	0.744
Tube, physical loss	37.26	5.246	0.744

Reproducibility and scope. The accompanying code regenerates the tables from per-slice metrics and checks their agreement with per-volume aggregates. Model checkpoints, configuration files, and input checksums are retained. Raw MRI volumes are not redistributed; recomputing image-level metrics requires separate authorized access to the corresponding fastMRI data. This pilot establishes feasibility of the complex-image implementation and illustrates the loss of measured uncertainty under hard projection. Its two-volume, single-seed design does not support clinical or competitive reconstruction claims. Establishing those claims would require source-matched learned controls, additional seeds and volumes, solver refinement, appropriate external baselines, and evaluation on an independent held-out cohort.