EQUIFLOW: FEW-STEP 3D MRI RECONSTRUCTION WITH AN EQUIVARIANT MEANFLOW FIELD

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing an accelerated 3D MRI volume with a generative prior costs tens to thousands of network evaluations. A patch MeanFlow prior trains from tens of volumes and makes a single jump from noise to data exact, yet used plug-and-play at one evaluation it gains 0.03 dB over the prior-free solve on knee MRI at $8\times$, because an unconditional prior has to invent a clean volume before the measurement can constrain it. We introduce EquiFlow (Equivariant MeanFlow reconstruction), a reconstructor that places a group-equivariant average-velocity field inside the inverse problem rather than in front of it: the zero-filled image is the network's input, the jump starts from that clean conditional rather than from a noised state, and a warm-started conjugate-gradient solve sits inside the training graph. Network and solve alternate for one to three passes, and drawing the number of passes at random during training yields one set of weights for every evaluation budget. With three network evaluations, EquiFlow outperforms BART compressed sensing by 0.3 to 3.1 dB on 33 held-out knee volumes and by 1.6 to 4.5 dB on a brain set held out from every design decision, from $4\times$ to $32\times$ at 30 dB matched noise. At matched parameters and training budget, the equivariant backbone, which compensates for the orientation that patch training removes, is ahead of a plain convolution at every rate on both anatomies. Ablated one at a time on the knee, the jump's starting point costs 2.7 to 5.2 dB and the warm start 0.3 to 2.0 dB. Every number is measured with measurement noise restored to benchmarks whose packaged k-space is otherwise the noiseless image of the reconstruction operator. Code will be released. Placing the network inside the solve, rather than sampling a prior in front of it, appears to bring a 3D generative reconstructor within the evaluation budget a clinical protocol can spend.

1 INTRODUCTION

Magnetic resonance imaging (MRI) is slow because k-space is sampled sequentially; accelerating a scan means *undersampling*, which makes reconstruction an ill-posed inverse problem that must be regularized. Diffusion and flow models are among the most expressive image priors available, and the standard way to use one is plug-and-play: sample a clean image from the prior, reconcile it with the measurement, repeat. That arrangement costs tens to thousands of network evaluations per volume. In this work we ask how few evaluations a 3D generative reconstructor needs.

The most direct route is a one-step generative model. A patch prior trained with MeanFlow (Geng et al., 2026), which parameterizes the average velocity of a flow so that a single jump from noise to data is exact rather than a truncated integration, gains 0.03 dB over the prior-free solve on knee MRI at $8\times$ when used plug-and-play with one evaluation, and four jumps gain 2.25 dB. The limit therefore lies in the arrangement rather than in the objective or the patches, since the same network gains 3.68 dB at one evaluation once it sits inside the solve.

To make a few evaluations sufficient, we move the network from the prior to the reconstructor. We introduce *EquiFlow* (Equivariant MeanFlow reconstruction), a reconstructor that places a group-equivariant average-velocity field inside the inverse problem rather than in front of it. Three changes make this possible, and §4.3 measures what each is worth. The zero-filled image becomes the network's *input* rather than only the quantity it is later made consistent with, so the network models what the measurement leaves undetermined. The jump is taken from that clean conditional rather

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

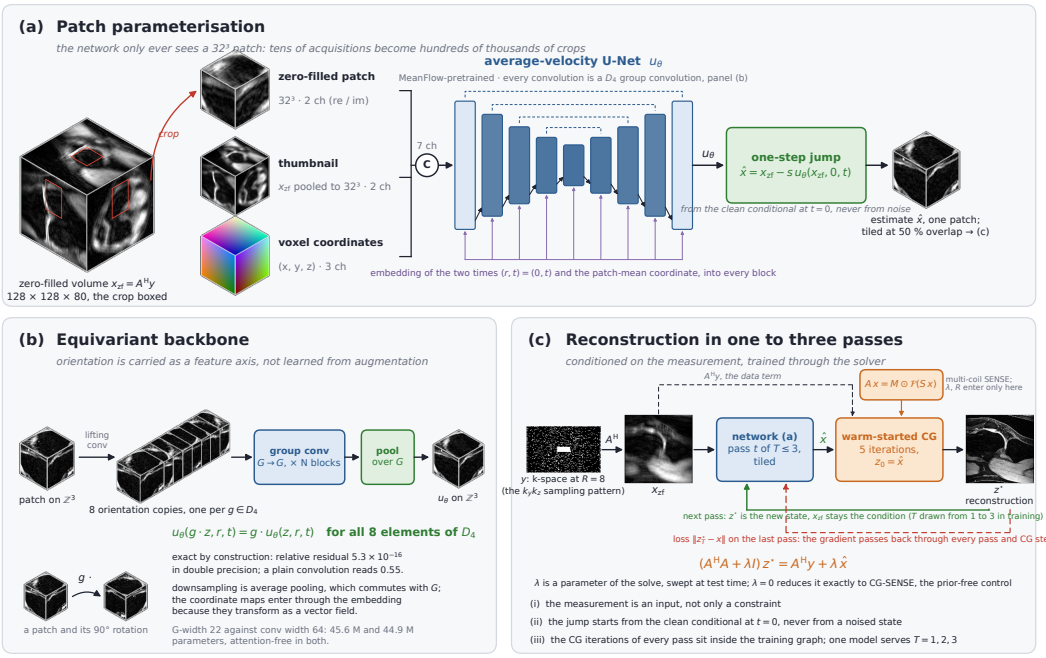

Figure 1: **Method overview. (a)** The network sees a 32^3 patch of its input, the input pooled to a 32^3 thumbnail and three coordinate channels ($2C{+}3{=}7$ inputs), and predicts the average velocity on the patch; cropping makes a 3D network trainable from a few dozen acquisitions and removes the volume's canonical orientation. **(b)** A lifting convolution carries the patch to the group and every later layer is a group convolution, so $\boldsymbol{u}_\theta(g \cdot \boldsymbol{z}) = g \cdot \boldsymbol{u}_\theta(\boldsymbol{z})$ holds for all eight elements of D_4 by construction, at matched parameter count. **(c)** Reconstruction alternates the network with a warm-started CG solve for $T \le 3$ passes: the zero-filled image is an *input* as well as the constraint, each jump starts from the clean conditional, and the CG iterations of every pass sit inside the training graph. Zeroing the prior term reduces each solve exactly to CG-SENSE.

than from a noised state, which is worth 2.7 to 5.2 dB. Finally, the short warm-started conjugate-gradient solve sits inside the training graph, so the network is trained for what the solver will do with its output; removing the warm start costs 0.3 to 2.0 dB. In this arrangement a single evaluation improves on the prior-free solve by 1.9 to 3.8 dB on knee. Because each pass returns an image consistent with the measurement, the same network can be applied to its own output: alternating network and solve for up to three passes, with the number of passes drawn at random during training, adds 0.7 to 1.8 dB, and three evaluations exceed BART compressed sensing (Lustig et al., 2007; Uecker et al., 2015) at every rate on both anatomies.

Why patches, and what they cost. A 3D model that processes whole volumes sees exactly one training example per volume, and a typical 3D MRI corpus has tens to a few hundred volumes. The patch formulation of Yang et al. (2025), developed for sparse-view CT, removes that bottleneck: the network sees only a 32^3 patch, conditioned on a downsampled thumbnail of the whole volume for global coherence and on voxel coordinates, so each volume supplies hundreds of thousands of training examples. We adapt it to multi-coil complex MRI, where the forward operator, the data-consistency solver and the network arithmetic all differ from the projection setting.

The adaptation carries a cost that, to our knowledge, has not been measured. Volumetric MRI is acquired in a canonical frame: patients are positioned by protocol, so whole volumes of the same anatomy agree on which way is up. A 32^3 patch cut from an arbitrary interior location does not inherit that agreement. On knee MRI the agreement R_4 between the dominant in-plane gradient orientations of different samples, the mean resultant length taken modulo $90°$, is 0.97 for whole volumes, 0.82 at 64^3, 0.47 at the 32^3 patches trained on, and 0.32 at 16^3. *Patch-ification* makes a 3D prior trainable from a few dozen volumes and removes the orientation prior a whole volume carries.

We therefore build the backbone from group convolutions over the in-plane roto-reflection group, which treat every orientation of a patch alike by construction rather than by augmentation.

Contributions. Two limitations set the problem: 1) a 3D prior trained on whole volumes has one example per acquisition, and 2) the one-step prior that patches make trainable is not, used plug-and-play, a few-step reconstructor. Three results follow, each resting on the one before.

- **A patch average-velocity prior for multi-coil 3D MRI.** A 32^3 patch network conditioned on a thumbnail of the volume and on voxel coordinates trains from tens of acquisitions, and the MeanFlow objective makes one jump from noise to data exact. Used as a prior it gains 0.03 dB over the prior-free solve at one evaluation, because an unconditional prior has to invent a clean volume before the measurement can constrain it.
- **A group-equivariant backbone for what patching removes.** Cropping discards the canonical orientation whole volumes share (agreement 0.97 at the volume, 0.47 at the patch), so the backbone convolves over D_4 and $\boldsymbol{u}_\theta(g \cdot \boldsymbol{z}) = g \cdot \boldsymbol{u}_\theta(\boldsymbol{z})$ holds to 5.3×10^{-16} in double precision, against 5.5×10^{-1} for a plain convolution. At matched parameters and training budget its margin over the plain convolution is 0.11 to 0.29 dB on knee at every rate and pass count but one, where the two tie, and 0.08 to 0.16 dB on brain at every rate and pass count; a convolution trained for half as many steps again still trails it on knee at three passes.
- **A few-pass reconstructor that puts the network inside the solve.** With the zero-filled image as input, the jump taken from that clean conditional, warm-started CG differentiated through, and network and solve alternated for up to three passes, three evaluations yield 0.3 to 3.1 dB over BART compressed sensing on knee and 1.6 to 4.5 dB on brain, and 3.4 to 5.5 dB over the prior-free solve on knee. Ablated one at a time, the jump's starting point costs 2.7 to 5.2 dB, the warm start 0.3 to 2.0 dB, and the explicit conditioning channels ± 0.01 dB, since at $t{=}0$ the state already is the zero-filled image.

2 Related Work

Accelerated MRI. Parallel imaging reconstructs from undersampled multi-coil k-space in image space (SENSE, Pruessmann et al., 1999) or in k-space (GRAPPA, Griswold et al., 2002), by conjugate gradients for arbitrary trajectories (Pruessmann et al., 2001), with a sparsity prior added by compressed sensing (Lustig et al., 2007). Unrolled networks (Hammernik et al., 2018; Sriram et al., 2020) learn the regulariser from paired data (Zbontar et al., 2018). With eight or more coils at low acceleration the problem is well conditioned and CG-SENSE alone is strong, so the regime in which a prior has to earn its place is heavy undersampling (Table 1).

Generative priors for inverse problems. Diffusion and score models (Ho et al., 2020; Song et al., 2020; Karras et al., 2022) are applied to inverse problems by training a task-agnostic prior and enforcing the measurement at sampling time (Chung et al., 2022; 2024), including for MRI (Chung & Ye, 2022), in the plug-and-play tradition (Venkatakrishnan et al., 2013). These priors are effective where a slice is an adequate unit of anatomy, yet they operate slice-wise in 2D or on small 3D data, and each costs tens to thousands of evaluations per volume. Flow matching (Lipman et al., 2022; Liu et al., 2022) learns a velocity field whose straighter paths need fewer integration steps, and MeanFlow (Geng et al., 2026) learns the average velocity so that a single jump is exact rather than truncated; exactness alone does not make one evaluation sufficient for reconstruction (§3.5).

Patch-based generative models. Patch diffusion trains on crops with positional conditioning for data efficiency (Wang et al., 2023), and Yang et al. (2025) carry it to 3D CT with a global-context channel. We adapt that formulation to complex multi-coil MRI, where the forward operator and the data-consistency solver differ from the projection setting.

Equivariance. Group-equivariant CNNs (Cohen & Welling, 2016) share each weight across a group of transformations, which raises weight sharing without raising the parameter count. Equivariant Imaging (Chen et al., 2021) uses the symmetry of the imaging system as a self-supervision signal when no ground truth exists, while leaving the network itself unconstrained. Ours instead

constrains the backbone to in-plane roto-reflection, so that a patch network recovers by construction the orientation that cropping removed.

3 METHOD

3.1 BACKGROUND: FLOW MATCHING AND AVERAGE VELOCITY

We work with the linear interpolant between data and noise,

$$\boldsymbol{z}_t = (1-t)\,\boldsymbol{x} + t\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I), \qquad t \in [0,1], \tag{1}$$

with $t{=}0$ the data end and $t{=}1$ pure noise. Its instantaneous velocity is $\boldsymbol{v}(\boldsymbol{z}_t, t) = \boldsymbol{\epsilon} - \boldsymbol{x}$, and flow matching (Lipman et al., 2022; Liu et al., 2022) regresses a network onto it. Sampling then integrates $\dot{\boldsymbol{z}} = \boldsymbol{v}_\theta$ at one network evaluation per step. A diffusion prior with EDM preconditioning (Karras et al., 2022) is the same arrangement in another parameterisation.

MeanFlow (Geng et al., 2026) models the *average* velocity over an interval instead,

$$\boldsymbol{u}(\boldsymbol{z}_t, r, t) = \frac{1}{t-r}\int_r^t \boldsymbol{v}(\boldsymbol{z}_\tau, \tau)\,\mathrm{d}\tau, \tag{2}$$

for which the single jump

$$\boldsymbol{z}_r = \boldsymbol{z}_t - (t-r)\,\boldsymbol{u}(\boldsymbol{z}_t, r, t) \tag{3}$$

is exact rather than a first-order step, so $t{=}1,\ r{=}0$ carries noise to data in one evaluation. Differentiating $(t-r)\,\boldsymbol{u} = \int_r^t \boldsymbol{v}$ in t gives an identity that makes $\boldsymbol{u}$ trainable without ever evaluating the integral,

$$\boldsymbol{u}(\boldsymbol{z}_t, r, t) = \boldsymbol{v}(\boldsymbol{z}_t, t) - (t-r)\,\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}(\boldsymbol{z}_t, r, t), \qquad \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u} = \partial_t \boldsymbol{u} + \boldsymbol{v}\cdot\nabla_{\boldsymbol{z}}\boldsymbol{u}, \tag{4}$$

the total derivative taken along the flow with r held fixed. Its right-hand side enters the loss as a stop-gradient target, so training stays a single regression on a single network: no teacher, no distillation, no multi-step model to collapse afterwards. The total derivative is one Jacobian-vector product, about one extra forward pass in forward mode.

3.2 PATCH MEANFLOW PRIOR

A whole-volume prior sees one training example per volume; with 86 knee and 47 brain volumes its regression target at moderate t collapses toward the mean of the data, and in our runs wider schedules and larger networks did not change this. The network therefore predicts the average velocity on a 32^3 patch. To keep patches globally coherent it also receives, concatenated along the channel axis, (i) the whole interpolated volume $\boldsymbol{z}_t$ average-pooled to 32^3, a blurry thumbnail of the entire anatomy, and (ii) three normalised voxel-coordinate channels. For complex MRI ($C{=}2$ real and imaginary channels) the input is $2C{+}3 = 7$ channels and the output is C:

$$\text{net_in} = \big[\, \boldsymbol{z}_t^{\text{patch}} \,\|\, \mathrm{pool}(\boldsymbol{z}_t) \,\|\, \text{coord} \,\big], \tag{5}$$

with the two times (r, t) entering through the embedding rather than the channel axis. There is no noise-level preconditioning to apply: the interpolant (1) is already unit-scaled in $\boldsymbol{\epsilon}$, and we run the flow on the volumes as normalised.

Training interpolates the *whole* volume at that sample's t, pools it for the thumbnail, and supervises (4) on four random patches of it. A volume of side N offers $\sim(N{-}P)^3 \approx 4\times10^5$ distinct crops, so tens of acquisitions become hundreds of thousands of examples. Without the global channel, patches do not agree on a shared anatomy.

The thumbnail is itself a function of $\boldsymbol{z}_t$, so it has to be carried in the tangent of the Jacobian-vector product, which is $(\boldsymbol{v}^{\text{patch}},\ \mathrm{pool}(\boldsymbol{v}),\ 0)$ with $\mathrm{d}t{=}1$ and r fixed. Dropping the thumbnail's tangent leaves a total derivative that is silently wrong rather than an error.

Two stabilisers introduced by Geng et al. (2026), used here with our own settings, make the objective trainable. A fraction $p_{=}{=}0.25$ of samples is drawn with $r{=}t$, where (4) reduces to plain flow matching and anchors the field on the diagonal; and each sample is weighted by its own error,

$$\mathcal{L} = \mathbb{E}\big[w\,\|\boldsymbol{u}_\theta - \mathrm{sg}[\boldsymbol{u}^\star]\|^2\big], \quad w = (\Delta + c)^{-1}, \quad \boldsymbol{u}^\star = \boldsymbol{v} - (t-r)\,\tfrac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}_\theta, \tag{6}$$

where Δ is that sample's mean squared error, w is detached, and $c{=}10^{-2}$. In our runs the $(t-r)\,\mathrm{d}\boldsymbol{u}/\mathrm{d}t$ target is heavy-tailed and an unweighted L_2 loss diverges. Times are drawn as $t = \mathrm{sigmoid}(-0.4 + \xi)$, $\xi \sim \mathcal{N}(0,1)$, with the logit-normal parameters of Geng et al. (2026), and $r \sim \mathcal{U}(0,t)$.

Headline configuration: patch $P{=}32$, four patches per volume, batch 8, AdamW at 10^{-4} with $\beta_2{=}0.95$, EMA 0.999, 30k steps.

3.3 Backbone

The velocity field $\boldsymbol{u}_\theta$ is a 3D U-Net over four resolution levels. The two times and the patch-mean coordinate enter through a shared embedding that conditions every residual block: a per-channel scale and shift in the plain backbone, and an additive per-channel bias in the equivariant one, where a conditional scale is admissible but not used. The output head is zero-initialised. Downsampling is average pooling rather than stride 2, for the reason given in §3.4. The plain variant carries a 3D self-attention block at the bottleneck; the equivariant variant cannot, because attention over the group is not equivariant here, so a matched comparison removes it from both (44.66 M parameters against 44.93 M with it). Nothing in the architecture is resolution-specific: the network always sees a 32^3 patch, whatever the size of the volume being reconstructed.

3.4 A Group-Equivariant Velocity Field

Patch-ification removes the canonical orientation a volume is acquired in: a 32^3 patch cut from an arbitrary location carries no reliable notion of which way is up, while the whole volume does. We restore it structurally rather than by augmentation, constraining the velocity field to be equivariant to the in-plane dihedral group D_4, $|G|{=}8$. A lifting convolution maps $\mathbb{Z}^3$ features to G, interior layers convolve over G, and a mean over G returns to a scalar field, so $\boldsymbol{u}_\theta(g\cdot\boldsymbol{z}, r, t) = g\cdot\boldsymbol{u}_\theta(\boldsymbol{z}, r, t)$ holds by construction to floating-point accuracy. Downsampling is average pooling rather than strided convolution, which commutes with the group action, and the three voxel-coordinate maps are routed through the embedding because they transform as a vector field under G rather than as three independent scalars. Appendix A gives the construction layer by layer, the measured equivariance residuals and the parameter matching.

Tiling a volume. The 32^3 network is applied to a whole volume by *tiling*: cover the volume with patches, evaluate each against the shared thumbnail and its own coordinates, and average the overlaps. Two tiling artifacts have to be suppressed: fixed-grid *block seams*, removed by jittering the tile origin, and *graininess*, removed by 50% overlap with count-normalized blending.

3.5 Why an Unconditional Prior Is Not Enough

Everything above describes a *prior*: a model of clean anatomy that knows nothing about the scan. Used plug-and-play (jump to a clean estimate, then reconcile it with the measurement by conjugate gradients) the one-step prior is weak. On the knee at $8\times$ it reads 30.12 dB against 30.09 for the prior-free solve at one evaluation and 32.34 at four, while the same patch prior trained as a diffusion model and sampled for 20 steps reads 31.10 (Table 1). The average-velocity objective is not what fails; the arrangement is. An unconditional prior has to invent a whole clean volume before the measurement constrains it, and one network evaluation is not enough to invent one.

We therefore move the network from the prior to the reconstructor: three changes make each network evaluation count, and a fourth spends the few evaluations a protocol can afford.

(i) The measurement is an input, not just a constraint. The zero-filled image $\boldsymbol{x}_{\mathrm{zf}} = \boldsymbol{A}^*\boldsymbol{y}$ is concatenated to both the patch and the thumbnail, so the network only has to model the part the measurement leaves open rather than reconstruct what it determines.

(ii) The jump starts from the clean conditional. Taking the MeanFlow jump from a noised state destroys exactly the measurement information (i) supplies. We therefore evaluate at $t{=}0$, where the state *is* $\boldsymbol{x}_{\mathrm{zf}}$, and give the jump a fixed scale s so that its leverage does not vanish with t:

$$\hat{\boldsymbol{x}} = \boldsymbol{x}_{\mathrm{zf}} - s\,\boldsymbol{u}_\theta\big(\boldsymbol{x}_{\mathrm{zf}},\, \boldsymbol{x}_{\mathrm{zf}},\, 0,\, t\big), \tag{7}$$

whose arguments are the state, the conditioning copy of $\boldsymbol{x}_{\mathrm{zf}}$ that (i) concatenates, r and t.

(iii) The solver is inside the training graph. Reconstruction ends with a short warm-started CG solve of

$$(\boldsymbol{A}^*\boldsymbol{A} + \lambda I)\,\boldsymbol{z}^\star = \boldsymbol{A}^*\boldsymbol{y} + \lambda\hat{\boldsymbol{x}}, \tag{8}$$

and we differentiate the training loss $\|\boldsymbol{z}^\star - \boldsymbol{x}\|$ *through* its five iterations. The network is then trained for the estimate whose projection is best rather than for a good-looking image. With the warm start replaced by CG from $\boldsymbol{z}{=}0$, the training loss does not move for ten thousand steps: a cold solver absorbs the gradient before it reaches the network.

Neither the regularization weight nor the acceleration is a network input. The weight λ is a parameter of the solve (8), changed by editing the linear system; the acceleration and the sampling pattern enter only through $\boldsymbol{A}$, so the network sees their effect already folded into $\boldsymbol{x}_{\mathrm{zf}}$. One trained model therefore covers every weight and every rate without being conditioned on either.

(iv) A few passes, one set of weights. The solve returns an image consistent with the measurement, so the network can be applied again to its own output. With $\boldsymbol{z}_0^\star = \boldsymbol{x}_{\mathrm{zf}}$, pass $k = 1, \ldots, T$ computes

$$\hat{\boldsymbol{x}}_k = \boldsymbol{z}_{k-1}^\star - s\,\boldsymbol{u}_\theta\big(\boldsymbol{z}_{k-1}^\star,\, \boldsymbol{x}_{\mathrm{zf}},\, 0,\, t\big), \qquad (\boldsymbol{A}^*\boldsymbol{A} + \lambda I)\,\boldsymbol{z}_k^\star = \boldsymbol{A}^*\boldsymbol{y} + \lambda\hat{\boldsymbol{x}}_k, \tag{9}$$

each solve warm-started at $\hat{\boldsymbol{x}}_k$, and the zero-filled image stays the condition throughout. Training draws T uniformly from $\{1, 2, 3\}$ at every step and differentiates the loss on $\boldsymbol{z}_T^\star$ through all passes, so one set of weights serves every pass count, and the later passes see the kind of input they are applied to at test time. The cost is T network evaluations and T short solves.

What MeanFlow contributes. The average-velocity parameterization makes each pass a well-defined jump rather than a truncated ODE integration. Retraining an earlier one-pass reconstructor from random weights instead of from the MeanFlow prior costs 0.1 to 0.4 dB, so the accuracy comes from (i) to (iv); MeanFlow lets them be spent in a few evaluations.

Early-stopped CG is still the regularizer. Within the solve we exploit semiconvergence (Hansen, 2010): the iterate first approaches a well-regularized solution and only later begins to fit the acquired noise, so the iteration count *is* the regularization parameter. Removing the prior term reduces (8) exactly to CG-SENSE (Pruessmann et al., 2001), the control every gain we report is measured against. The weight λ has to be swept rather than fixed: the same checkpoint can gain more than a decibel on the prior-free control at one value and fall behind it at another. Because the network only ever sees 32^3 patches, the same reconstructor applies to a volume of any size.

4 Experiments

4.1 Setup

Datasets. **SKM-TEA** knee (Desai et al., 2022) (real 8- or 16-coil) at quarter resolution, $128 \times 128 \times 80$: 86 volumes for training, 33 held out. **Calgary-Campinas CC-359** brain (Souza et al., 2018) (real 12-coil) at $128 \times 104 \times 80$, 20 volumes held out. CC-359 was added after the method was frozen and nothing in the recipe was tuned on it. SKM-TEA ships JSENSE sensitivity maps, which we use as given; for CC-359 the maps are ESPIRiT (Uecker et al., 2014) from BART (Uecker et al., 2015) on a 24-wide calibration region. An ESPIRiT eigenvector carries an arbitrary phase at every voxel, and a prior trained under one phase gauge does not transfer to another, so each CC-359 map set is rotated voxelwise to the gauge of the low-resolution maps the priors were trained on, which changes no magnitude and leaves CG-SENSE invariant.

Training. Every model is built in two stages. A patch prior is trained on target images for 30k steps with the MeanFlow objective of §3.2; the reconstructor is then fine-tuned from it with the measurement as input and the solver in the graph (§3.5), 4k steps with one pass and 4k steps with the number of passes drawn from $\{1, 2, 3\}$, sampling $R \in \{4, 8, 16, 32\}$ and λ log-uniformly per step so one set of weights covers every rate, weight and pass count. Neither λ nor R is a network input. The plain-convolution control is trained with the same budget from its own prior.

Baselines. **CG-SENSE** (Pruessmann et al., 2001) is the decisive control: removing the prior term reduces our solve to it exactly, on the same maps and the same code path, so the gap between the two

Table 1: **Knee, 33 held-out volumes, 30 dB matched noise.** PSNR in dB; NFE counts network evaluations. Every row runs on the same volumes, masks, noise level and forward operator, each method at its own best settings. Ours is one set of weights evaluated with 1, 2 or 3 passes and 20 CG iterations per pass; the plain-convolution row is the parameter-matched control trained with the same budget. †Still rising at the largest iteration count we ran (1600).

Method	NFE	4×	8×	16×	32×
CG-SENSE (prior-free)	0	37.94	30.09	26.64	25.47
BART pics (ℓ_1-wavelet)	0	41.02	33.64	28.67	26.08
3D U-Net	1	30.29	28.47	27.00	25.93
E2E-VarNet 3D	1	34.37	31.39	29.03	27.17
PnP-Flow, 2D rectified flow	≥ 100	35.71†	31.35	28.16	25.55
2D EDM prior, plug-and-play	20	38.67	31.32	27.08	25.14
DDS, 2D EDM prior	5	38.78	31.40	27.51	25.50
Patch diffusion (EDM)	20	38.83	31.10	26.97	25.33
Our prior, plug-and-play	1	38.01	30.12	26.59	25.37
Our prior, plug-and-play	4	39.58	32.34	27.81	25.73
Ours (plain convolution)	3	41.16	35.40	31.51	28.58
Ours (G-equivariant)	1	39.88	33.77	30.42	28.15
Ours (G-equivariant)	2	41.06	35.16	31.39	28.64
Ours (G-equivariant)	3	**41.29**	**35.57**	**31.76**	**28.82**

is what the network contributes. It is given its own best iteration count. **BART pics** is ℓ_1-wavelet compressed sensing (Lustig et al., 2007) as clinical collaborators run it, sixty iterations with the weight swept per rate until the optimum is interior, on our measurement and maps. **Patch diffusion** is the same patch prior trained as an EDM diffusion model (Karras et al., 2022) and run plug-and-play for 20 sampler steps, retrained and re-measured here rather than quoted. **Our own prior** used plug-and-play at one and four evaluations isolates the objective from the arrangement. Two supervised one-pass networks, a 3D U-Net and E2E-VarNet (Sriram et al., 2020) lifted to 3D, cost a single forward pass. The fast **2D** family is the cheap-reconstruction comparison: a slice-wise EDM prior driven plug-and-play and by DDS (Chung et al., 2024), and a slice-wise rectified-flow prior (Liu et al., 2022) driven by PnP-Flow (Martin et al., 2025), all trained on the same targets, so the comparison is 2D versus 3D.

Protocol. Whole-volume PSNR, SSIM (Wang et al., 2004) and fastMRI NMSE. *Every number is measured with complex measurement noise added at 30 dB k-space SNR*, because the packaged evaluation k-space of the public sets is operator-synthesised, $y = A(x^\star)$ exactly, which makes reconstruction a noiseless inversion and flatters every iterative solver while leaving feed-forward networks untouched. Under that unrepaired protocol prior-free CG-SENSE alone reaches 55.8 dB on knee half-resolution at 4×, and 43.15 dB once the noise is restored; that 12.6 dB is the size of the distortion. Every method is swept over its own knobs with the optimum bracketed: λ and the CG iterations per pass (5 or 20) for ours, the weight for pics, the step count and step size for the samplers. Three mask-and-noise draws moved every headline entry of the earlier one-pass arm and of the baselines by at most 0.05 dB, so the tables are quoted at one draw; the draw is seeded, but a seeded GPU draw also depends on the device model, so rows scored on different GPUs see different draws of the same noise level.

4.2 Reconstruction in One to Three Passes

Three network evaluations outperform BART compressed sensing at every rate on both anatomies, by 0.3 to 3.1 dB on knee and 1.6 to 4.5 dB on brain, and the prior-free solve by 3.4 to 5.5 dB on knee and 4.2 to 5.9 dB on brain. The margin over pics is smallest at 4×, where eight or more coils leave the problem well conditioned and a converged ℓ_1-wavelet solve is hard to improve on, and it widens with acceleration. The second pass carries most of what further passes add: on knee it gains 0.5 to 1.4 dB over one pass and the third 0.2 to 0.4 dB more. One pass already exceeds every baseline from 8× on and trails pics by 1.1 dB at 4×; two passes match it there. At 8× that single pass reads +3.68 dB against the prior-free solve, whereas the patch prior used plug-and-play at one evaluation reads +0.03.

At matched parameters and training budget, the equivariant backbone leads the plain convolution on knee by 0.11 to 0.29 dB at every rate and pass count but one, where the two tie (−0.01 dB); a

Table 2: **Brain, CC-359,** 20 **held-out volumes,** 30 **dB matched noise.** No hyperparameter was tuned on this set. Every row is on the ESPIRiT calibration of §4.1 with the phase gauge aligned; the supervised baselines and our arms are retrained on it, and ours run five CG iterations per pass; twenty moved the one-pass arm by at most 0.09 dB. PnP-Flow is swept over $\gamma \in \{2, \ldots, 8, 10\}$ and 100 to 1600 iterations, with the optimum interior on both axes.

Method	NFE	4×	8×	16×	32×
CG-SENSE (prior-free)	0	33.39	28.69	25.27	23.55
BART pics (ℓ_1-wavelet)	0	36.03	32.07	27.89	24.43
3D U-Net	1	28.13	27.04	26.41	26.09
E2E-VarNet 3D	1	31.63	29.14	27.74	26.98
DDS, 2D EDM prior	3	32.85	29.58	26.35	24.00
2D EDM prior, plug-and-play	20	34.38	29.87	26.28	23.87
PnP-Flow, 2D rectified flow	400	35.62	31.96	28.70	25.60
Patch diffusion (EDM)	20	34.75	30.17	26.46	24.06
Our prior, plug-and-play	4	34.95	30.82	26.88	24.10
Ours (plain convolution)	3	37.55	34.05	31.03	28.81
Ours (G-equivariant)	1	35.86	32.24	29.92	28.51
Ours (G-equivariant)	2	37.36	33.78	30.87	28.87
Ours (G-equivariant)	3	**37.63**	**34.17**	**31.16**	**28.88**

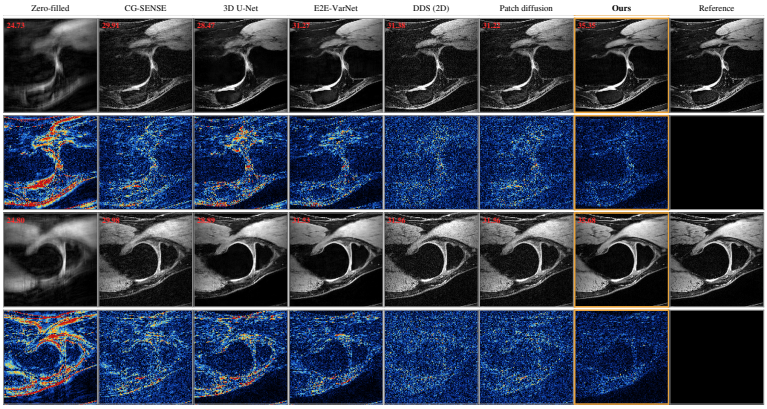

Figure 2: **Every method on two knee slices**, volume 0 at 8× with 30 dB matched noise, each method at its own best setting; slice PSNR in red, and below each image the absolute error on a scale common to its row. The two are the middle ones of four slices, spread through the volume, chosen where our column leads the strongest other method by the most. Our column is the bold row of Table 1, three passes of the final weights; DDS is shown at 20 steps. Table 1 gives every method over all 33 held-out volumes.

convolution trained for half as many steps again still trails it at three passes (41.27, 35.52, 31.67, 28.71). On brain it is ahead at every rate and pass count, by 0.08 to 0.16 dB.

Both supervised one-pass baselines fall below prior-free CG-SENSE at 4×, by 3.6 and 7.7 dB on knee, and overtake it only once the operator is ill-posed enough to need a prior. Between the two plug-and-play prior rows of Table 1, the MeanFlow prior at four evaluations is 0.4 to 1.2 dB above the diffusion prior at twenty, with a fifth of the evaluations. That gain from the objective is well below the 2.5 to 4.8 dB that three passes inside the solve gain over the same diffusion prior.

4.3 What Each Ingredient Is Worth

Table 3 removes one ingredient at a time from the one-pass reconstructor. Each arm is trained for 10k steps from the same MeanFlow prior (its attention variant for the attention row) and *scored under the protocol it was trained for*; scoring a cold-solver model warm, or a noised-jump model at t=0, measures a distribution shift rather than the ablation.

The jump has to start from the clean conditional: taking it from a noised state costs 2.7 to 5.2 dB, the largest effect we measure, because the noise removes the measurement information the condi-

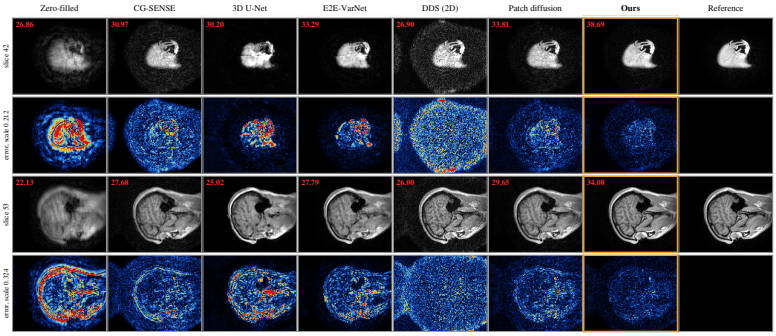

Figure 3: **The same comparison on brain**, CC-359 volume 0 at 8× on the ESPIRiT calibration of Table 2, with the layout and slice rule of Figure 2. Our column is the bold row of Table 2, three passes of the final weights; DDS is shown at 20 steps, where the table's row uses three. Table 2 gives every method over all 20 held-out volumes.

Table 3: **Ablations, knee,** 33 **held-out volumes,** 30 **dB, one pass, plain convolution.** PSNR in dB; the first row keeps all three ingredients and each later row removes or adds one. Each arm is at its own best weight: the warm-started arms on the plateau below 10^{-4}, the cold-solver arms at λ=0.1, interior to their grids.

Configuration	4×	8×	16×	32×
All three ingredients	38.29	33.29	30.29	28.08
Jump from a noised state (t_{start}=0.5)	33.07	28.79	26.54	25.39
Cold solver, trained and tested cold	36.29	32.45	29.86	27.82
Cold solver at test time only	36.42	32.56	30.02	27.97
Without the explicit ZF channels	38.29	33.29	30.28	28.08
With bottleneck self-attention	38.31	33.30	30.30	28.08

tioning supplied. **The warm start has to be inside the training graph**: removing it costs 0.3 to 2.0 dB when the network is retrained for the cold solver and 0.1 to 1.9 dB when it is removed at test time only, so retraining for a cold solver recovers nothing; the cost is largest at 4×. **The explicit conditioning channels are redundant**: once the jump starts at t=0, the state the network receives already is the zero-filled image.

5 Conclusion

A patch MeanFlow model used as a prior gains 0.03 dB over the prior-free solve on knee MRI at 8× at one evaluation. Fine-tuned inside the inverse problem, with the measurement as its input, the jump taken from that clean conditional and a warm-started solver in the training graph, the same network gains 3.68 dB at one evaluation, and alternated with the solve for three passes the reconstructor overtakes BART compressed sensing at every rate on knee and brain. A group-equivariant backbone compensates for the orientation that patch training removes and improves on a parameter-matched plain convolution by 0.11 to 0.29 dB on knee and 0.08 to 0.16 dB on brain, while a convolution trained half as long again still trails it on knee. Placing the network inside the solve therefore appears to bring a 3D generative reconstructor within the few network evaluations a clinical protocol can spend.

Limitations. Each cell is one mask-and-noise draw; the spread over three draws, at most 0.05 dB, was measured on an earlier one-pass arm. Training stops at three passes. Undersampling is retrospective and the noise simulated, so prospective data and a reader study remain to be done.

AI use statement

In this work, we used generative AI tools, an LLM coding assistant (Claude), to implement methods, to design and run experiments, to propose and refine hypotheses about the behaviour observed