CAST: CONDITIONAL ACQUISITION-STRUCTURED TRANSPORT FOR ACCELERATED MRI

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models are increasingly popular in accelerated MRI reconstruction. However, generative methods typically introduce acquisition operators and data consistency only at inference. We investigate how acquisition information can improve flow matching-based generative MRI reconstruction when incorporated during training. The result is Conditional Acquisition-Structured Transport (CAST), which uses acquisition operators to structure both the source distribution and the velocity field of the flow matching model. Sampling via CAST admits a corresponding k-space trajectory that preserves acquired measurements. We evaluate CAST on 2D, 3D, and cardiac cine MRI. CAST significantly improves reconstruction over network-input conditioning alone and enables the perception–distortion trade-off to be adjusted at inference. Averaging multiple samples matches or exceeds state-of-the-art unrolled networks in PSNR, closing the distortion gap that generative methods typically exhibit, while single-sample reconstructions achieve greater distributional similarity to reference images than generative baselines. On cardiac cine MRI, CAST can reconstruct fine details and reduce artifacts.

1 INTRODUCTION

Magnetic resonance imaging (MRI) is inherently slow. In static MRI, measurements are undersampled to reduce scan time. In dynamic MRI, the number of measurements available for each frame is limited by the temporal resolution (Chen et al., 2020). Reconstruction methods must recover high-quality images from such incomplete measurements, and deep neural networks achieve strong performance for this problem (Muckley et al., 2021; Hammernik et al., 2021).

Generative models, particularly diffusion and flow matching models, have become increasingly popular for image reconstruction. Generative model based approaches enable sampling which has several advantages, in particular it enables trading perceptual quality against distortion: individual samples can better match the distribution of reference images, while averaging samples can improve distortion metrics like PSNR (Blau & Michaeli, 2018; Bendel et al., 2023). This feature is relevant in MRI, where perceptual metrics based on deep features have been shown to correlate more strongly with radiologist assessments than conventional distortion metrics such as PSNR and SSIM (Adamson et al., 2025; Kastryulin et al., 2023).

However, generative methods typically trail the state-of-the-art. An hypothesis is that while state-of-the-art methods like variational networks incorporate the forward model during training, most methods based on generative priors don't: they incorporate the MRI acquisition operators and measured k-space only during reconstruction (Jalal et al., 2021; Chung & Ye, 2022; Kim et al., 2025; Martin et al., 2025; Mardani et al., 2024).

When paired undersampled measurements and fully sampled references are available, acquisition information can instead be incorporated already during training. We investigate how this information can be used to improve flow matching-based generative MRI reconstruction.

Flow matching provides a natural framework for this, as it learns a velocity field whose flow transports samples from a source distribution to the target reconstruction distribution, and both the source distribution and the velocity field can be designed explicitly (Lipman et al., 2023; Tong et al., 2024).

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.

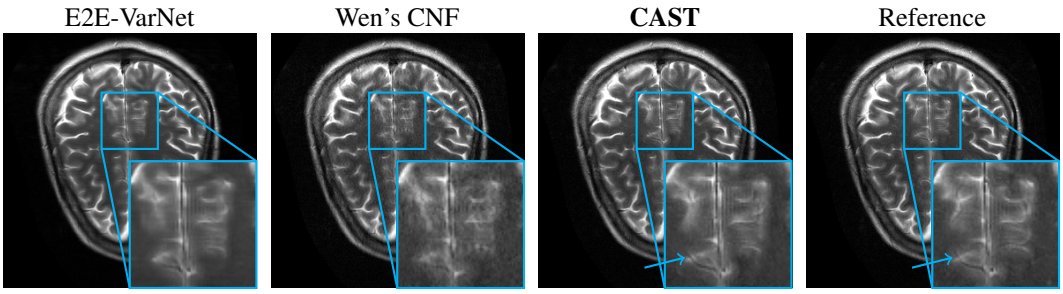

Figure 1: Example fastMRI brain reconstructions. Wen's CNF and CAST show single-sample reconstructions. CAST reconstructs finer image structure compared to VarNet and Wen's CNF.

In standard conditional flow matching, acquisition information enters only as an additional input to the velocity network. We introduce CAST (Conditional Acquisition-Structured Transport), which goes further and uses the MRI acquisition operators to structure the source distribution and velocity field themselves. CAST initializes unmeasured k-space entries with noise while retaining the acquired measurements and restricts learned updates to the unmeasured entries. Its velocity network operates in coil-combined image space and is independent of coil count, while the corresponding k-space trajectory preserves the acquired measurements throughout sampling.

Our experiments show that how acquisition information is incorporated during training substantially affects reconstruction quality. Our contributions are as follows:

- We introduce CAST, a conditional flow matching method that structures its source distribution and velocity field using MRI acquisition operators. Its network is independent of coil count and operates in coil-combined image space, and sampling induces a k-space trajectory that preserves acquired measurements.
- We show that network-input conditioning alone does not improve reconstruction quality over a pretrained prior, whereas structuring the source and velocity field with the acquisition operators improves both perceptual and distortion metrics.
- We evaluate CAST on 2D, 3D, and cardiac cine MRI. CAST improves the perception–distortion trade-off over generative baselines: single-sample reconstructions match the feature distribution of fully sampled references more closely, and averaging samples reaches PSNR and SSIM on par with state-of-the-art unrolled networks. See Figure 1, demonstrating that CAST can enable performance on par with the state-of-the-art E2E-VarNet (Muckley et al., 2021).

  On prospectively undersampled free-breathing cardiac MRI, CAST outperforms unrolled networks in held-out measurement prediction and reduces visible artifacts.

2 Related Work

Generative priors and inverse solvers. Generative priors are trained on fully sampled images and combined with acquisition operators and measurements during reconstruction through posterior sampling (Jalal et al., 2021; Chung et al., 2024; Kim et al., 2025; Lee et al., 2023), plug-and-play methods (Zhu et al., 2023; Martin et al., 2025), or variational approaches (Mardani et al., 2024; Krainovic et al., 2024). In these approaches, acquisition information is primarily introduced at inference. We instead investigate using acquisition information during training to learn the conditional reconstruction distribution directly.

Conditional generative models. Conditional generative models can learn the reconstruction distribution directly from measurement-reference pairs. Bendel et al. (2023) train a conditional GAN for posterior sampling in multicoil MRI. Wen et al. (2023) propose a conditional normalizing flow that learns the component in the measurement nullspace and combines it with measured data. CAST similarly incorporates acquisition information during training, but uses flow matching to structure

Algorithm 1 CAST training

Input: Training data $(\boldsymbol{y}^{\text{full}}, \boldsymbol{M}, \boldsymbol{S})$; Fourier operator $\boldsymbol{F}$; initial network $\boldsymbol{h_\theta}$

while training **do**
  Sample $(\boldsymbol{y}^{\text{full}}, \boldsymbol{M}, \boldsymbol{S})$; $Z \sim \mathcal{CN}(0, \boldsymbol{I})$; $t \sim \mathcal{U}[0, 1]$
  $\boldsymbol{y}^{\text{obs}} \leftarrow \boldsymbol{M}\boldsymbol{y}^{\text{full}}$; $\boldsymbol{x}^\star \leftarrow \boldsymbol{S}^{\mathsf{H}}\boldsymbol{F}^{\mathsf{H}}\boldsymbol{y}^{\text{full}}$; $\boldsymbol{B} \leftarrow (\boldsymbol{I} - \boldsymbol{M})\boldsymbol{F}\boldsymbol{S}$
  $X_0 \leftarrow \boldsymbol{S}^{\mathsf{H}}\boldsymbol{F}^{\mathsf{H}}\boldsymbol{y}^{\text{obs}} + \boldsymbol{B}^{\mathsf{H}}\boldsymbol{B}Z$
  $X_t \leftarrow (1-t)X_0 + t\boldsymbol{x}^\star$
  $\mathcal{L}(\boldsymbol{\theta}) \leftarrow \|\boldsymbol{u_\theta}(X_t, t) - (\boldsymbol{x}^\star - X_0)\|_2^2$
  $\boldsymbol{\theta} \leftarrow \text{Update}(\boldsymbol{\theta}, \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}))$
end while
return $\boldsymbol{h_\theta}$

Algorithm 2 CAST sampling

Require: $\boldsymbol{y}^{\text{obs}}$; operators $\boldsymbol{M}, \boldsymbol{F}, \boldsymbol{S}$; trained network $\boldsymbol{h_\theta}$

$\boldsymbol{B} \leftarrow (\boldsymbol{I} - \boldsymbol{M})\boldsymbol{F}\boldsymbol{S}$
Draw $Z \sim \mathcal{CN}(0, \boldsymbol{I})$
$Y_0 \leftarrow \boldsymbol{y}^{\text{obs}} + \boldsymbol{B}Z$
Integrate from $t = 0$ to $t = 1$:
$\dot{Y}_t = \boldsymbol{B}\boldsymbol{h_\theta}\big(\boldsymbol{S}^{\mathsf{H}}\boldsymbol{F}^{\mathsf{H}}Y_t, t \mid \boldsymbol{y}^{\text{obs}}, \boldsymbol{M}, \boldsymbol{S}\big)$
return $X_1 \leftarrow \boldsymbol{S}^{\mathsf{H}}\boldsymbol{F}^{\mathsf{H}}Y_1$

both its source distribution and velocity field in coil-combined image space. The velocity network has input and output dimensions independent of coil count.

Unrolled networks. Unrolled networks combine learned regularization with acquisition operators and data-consistency updates (Aggarwal et al., 2019; Sriram et al., 2020; Hammernik et al., 2021; Vornehm et al., 2025). They achieve strong performance on distortion metrics like PSNR but return a single reconstruction rather than samples from a conditional reconstruction distribution. We use unrolled networks as low-distortion references when evaluating generative reconstruction methods.

3 Problem Setup and Background

We consider multicoil accelerated MRI, where the goal is to reconstruct the ground-truth image $\boldsymbol{x}_{\text{gt}} \in \mathbb{C}^n$ from $m < n$ k-space samples per coil. The forward model is

$$\boldsymbol{y}^{\text{obs}} = \boldsymbol{M}\big(\boldsymbol{F}\boldsymbol{S}\boldsymbol{x}_{\text{gt}} + \boldsymbol{z}\big), \tag{1}$$

where $\boldsymbol{y}^{\text{obs}}$ contains the coil-wise stacked k-space measurements, $\boldsymbol{S}$ contains the sensitivity maps, and $\boldsymbol{F}$ is the discrete Fourier transform applied to each coil image. The mask $\boldsymbol{M}$ is a block-diagonal matrix selecting m k-space entries per coil, and $\boldsymbol{z}$ is white Gaussian noise.

Since ground-truth images are not available, reconstructions from fully sampled data $\boldsymbol{y}^{\text{full}}$ are often used as training and evaluation references. For multicoil acquisitions at $1.5\,\text{T}$ and above, such reconstructions typically have sufficiently high signal-to-noise ratio to serve as useful references (Zbontar et al., 2018; Muckley et al., 2021). A common choice is the minimum-variance unbiased estimate (MVUE) (Hammernik et al., 2021; Jalal et al., 2021),

$$\boldsymbol{x}^\star = \boldsymbol{S}^{\mathsf{H}}\boldsymbol{F}^{\mathsf{H}}\boldsymbol{y}^{\text{full}}. \tag{2}$$

In the noiseless case, $\boldsymbol{x}^\star = \boldsymbol{x}_{\text{gt}}$.

Let $X^\star$ denote the random MVUE reconstruction, with realizations $\boldsymbol{x}^\star$. We learn its conditional distribution,

$$X^\star \mid (\boldsymbol{y}^{\text{obs}}, \boldsymbol{M}, \boldsymbol{S}) \sim p(\boldsymbol{x} \mid \boldsymbol{y}^{\text{obs}}, \boldsymbol{M}, \boldsymbol{S}). \tag{3}$$

3.1 Conditional Flow Matching

Flow matching constructs a velocity field that transports a source distribution p_0 to a target distribution p_1 (Lipman et al., 2024). For fixed conditioning information, draw independent samples $X_0 \sim p_0$

Table 1: Datasets used for 2D, 3D, and cardiac cine MRI experiments. Training and test sizes are reported as slices/subjects.

Dataset	Contrast	Coils	Field (T)	Mask	R	Training	Test
fastMRI brain, hom.	T2	4	1.5	Equisp.	4	4496/281	128/103
fastMRI brain, diverse	T1,T1post,T2	4,12,16	1.5,3	Equisp.	4	10000/625	256/245
Stanford3D knee	PDFS	8	3	Poisson	8	3584/14	1024/4
OCMR, breath-hold	bSSFP	15–38	1.5,3	GRO	13	109/14	64/8
OCMR, free-breathing	bSSFP	15–34	1.5,3	GRO	13	–	20/20

and $X_1 \sim p_1$ and define the linear interpolation

$$X_t = (1-t)X_0 + tX_1, \qquad X_t \sim p_t, \qquad 0 \le t \le 1. \tag{4}$$

Each interpolation has constant velocity $X_1 - X_0$. For an independent $t \sim \mathcal{U}[0,1]$, the conditional flow matching loss

$$\mathcal{L}(\boldsymbol{v}) = \mathbb{E}_{t,X_0,X_1}\left[\|\boldsymbol{v}(X_t,t) - (X_1 - X_0)\|_2^2\right] \tag{5}$$

is minimized by the marginal velocity

$$\boldsymbol{u}(\boldsymbol{x},t) = \mathbb{E}\left[X_1 - X_0 \mid X_t = \boldsymbol{x}\right], \tag{6}$$

which generates the probability path p_t under the standard flow matching assumptions (Lipman et al., 2024).

4 Using Acquisition Information in Flow Matching for MRI

We consider two ways of incorporating acquisition information into conditional flow matching. First, we provide the measurements and acquisition operators only as conditioning inputs to the velocity network. We then introduce CAST, which additionally uses the MRI acquisition operators to structure the source distribution and velocity field.

4.1 Conditioning through network inputs

We first consider conditional flow matching where acquisition information is provided only through inputs to the velocity network. Let p_0 be an isotropic Gaussian source. Given an acquisition $(\boldsymbol{y}^{\mathrm{obs}}, \boldsymbol{M}, \boldsymbol{S})$, draw $X_0 \sim p_0$ and $X_1 \sim p(\boldsymbol{x} \mid \boldsymbol{y}^{\mathrm{obs}}, \boldsymbol{M}, \boldsymbol{S})$ independently and use the linear path in Equation (4). We parameterize the velocity as

$$\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x},t) = \boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}, t \mid \boldsymbol{y}^{\mathrm{obs}}, \boldsymbol{M}, \boldsymbol{S}). \tag{7}$$

We minimize the conditional flow matching loss in Equation (5). Sampling integrates $\dot{X}_t = \boldsymbol{u}_{\boldsymbol{\theta}}(X_t,t)$ from $X_0 \sim p_0$ at $t=0$ to $t=1$.

4.2 Acquisition-structured source and velocity

Conditioning through network inputs gives only the velocity network access to the acquisition information. We now introduce CAST, which additionally uses the MRI acquisition operators to structure both the source distribution and velocity field.

Define the zero-filled reconstruction $\boldsymbol{x}^{\mathrm{zf}} = \boldsymbol{S}^{\mathsf{H}}\boldsymbol{F}^{\mathsf{H}}\boldsymbol{y}^{\mathrm{obs}}$ and the complementary mask $\overline{\boldsymbol{M}} = \boldsymbol{I} - \boldsymbol{M}$, and let $\boldsymbol{B} = \overline{\boldsymbol{M}}\boldsymbol{F}\boldsymbol{S}$. The MVUE reconstruction satisfies

$$\boldsymbol{x}^{\star} = \boldsymbol{x}^{\mathrm{zf}} + \boldsymbol{B}^{\mathsf{H}}\boldsymbol{y}^{\mathrm{full}}. \tag{8}$$

We construct the source by applying the same operators to Gaussian noise:

$$X_0 = \boldsymbol{x}^{\mathrm{zf}} + \boldsymbol{B}^{\mathsf{H}}\boldsymbol{B}Z, \qquad Z \sim \mathcal{CN}(0,\boldsymbol{I}). \tag{9}$$

Table 2: Single-sample reconstruction on the homogeneous fastMRI setup. Input-only conditioning underperforms reconstruction with pretrained generative priors, while CAST substantially improves both perceptual and distortion metrics. Best scores are bold and second-best scores are underlined. KID is multiplied by 1000.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	KID ↓
Wen's CNF	35.41	0.916	0.093	9.44
FlowDPS	36.02	0.930	0.085	3.85
FlowPnP	**36.73**	**0.937**	<u>0.078</u>	<u>3.21</u>
Input-only conditioning	34.40	0.881	0.104	8.43
CAST	<u>36.26</u>	<u>0.932</u>	**0.076**	**0.50**

Draw $X_1 \sim p(\boldsymbol{x} \mid \boldsymbol{y}^{\mathrm{obs}}, \boldsymbol{M}, \boldsymbol{S})$ independently of X_0 given the acquisition and use the linear interpolant path defined in Equation (4). CAST applies the same operator $\boldsymbol{B}^{\mathsf{H}}\boldsymbol{B}$ to an image-space network $\boldsymbol{h}_{\boldsymbol{\theta}}$ to regress the velocity field

$$\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) = \boldsymbol{B}^{\mathsf{H}}\boldsymbol{B}\boldsymbol{h}_{\boldsymbol{\theta}}(\boldsymbol{x}, t \mid \boldsymbol{y}^{\mathrm{obs}}, \boldsymbol{M}, \boldsymbol{S}) \tag{10}$$

We train $\boldsymbol{u}_{\boldsymbol{\theta}}$ with the conditional flow matching loss (5). The network input and output dimensions are independent of coil count. Appendix A shows that this parameterization can represent the marginal velocity field in Equation (6). Algorithm 1 summarizes CAST training.

4.3 Measurement-preserving k-space completion

At inference, CAST samples X_0 using Equation (9) and integrates

$$\dot{X}_t = \boldsymbol{u}_{\boldsymbol{\theta}}(X_t, t) \tag{11}$$

from $t = 0$ to $t = 1$, returning X_1. To connect the image trajectory to the k-space, consider the k-space trajectory obtained by integrating the network output in unacquired k-space locations:

$$Y_t = Y_0 + \boldsymbol{B} \int_0^t \boldsymbol{h}_{\boldsymbol{\theta}}(X_s, s \mid \boldsymbol{y}^{\mathrm{obs}}, \boldsymbol{M}, \boldsymbol{S})\, ds. \tag{12}$$

where $Y_0 = \boldsymbol{y}^{\mathrm{obs}} + \boldsymbol{B}Z$. Since $\boldsymbol{M}\overline{\boldsymbol{M}} = \boldsymbol{0}$, we have $\boldsymbol{M}Y_t = \boldsymbol{y}^{\mathrm{obs}}$. Hence, the acquired k-space entries remain unchanged throughout the trajectory, while only the unacquired entries are updated. Now, applying $\boldsymbol{S}^{\mathsf{H}}\boldsymbol{F}^{\mathsf{H}}$ to Y_t yields

$$\boldsymbol{S}^{\mathsf{H}}\boldsymbol{F}^{\mathsf{H}}Y_t = X_0 + \int_0^t \boldsymbol{u}_{\boldsymbol{\theta}}(X_s, s)\, ds = X_t, \tag{13}$$

showing that Y_t induces exactly the image-space sampling trajectory in Equation (11). Algorithm 2 summarizes CAST sampling.

Measurement preservation holds for the acquired k-space entries in Y_t. Hence, the returned image need not satisfy $\boldsymbol{M}\boldsymbol{F}\boldsymbol{S}X_t = \boldsymbol{y}^{\mathrm{obs}}$. Instead, CAST targets the conditional distribution of MVUE images defined in Equation (3).

5 Experiments

We first study how acquisition information should be incorporated during training. We start with 2D accelerated MRI and compare input-only conditioning with CAST and with generative reconstruction methods based on pretrained image priors. We then examine how sampling and averaging affect the perception–distortion trade-off and compare CAST with generative reconstruction methods. Finally, we apply CAST to 3D and cardiac cine MRI.

Table 1 summarizes the datasets used in our experiments. For diffusion and flow matching models, we use the standard diffusion-model U-Net (Dhariwal & Nichol, 2021). For CAST and the input-only conditioning model, acquisition information is provided to the network through $X_t - \boldsymbol{x}^{\mathrm{zf}}$, $\boldsymbol{x}^{\mathrm{zf}}$, and the backprojected measurement residual $\boldsymbol{S}^{\mathsf{H}}\boldsymbol{F}^{\mathsf{H}}\boldsymbol{M}(\boldsymbol{F}\boldsymbol{S}X_t - \boldsymbol{y}^{\mathrm{obs}})$.

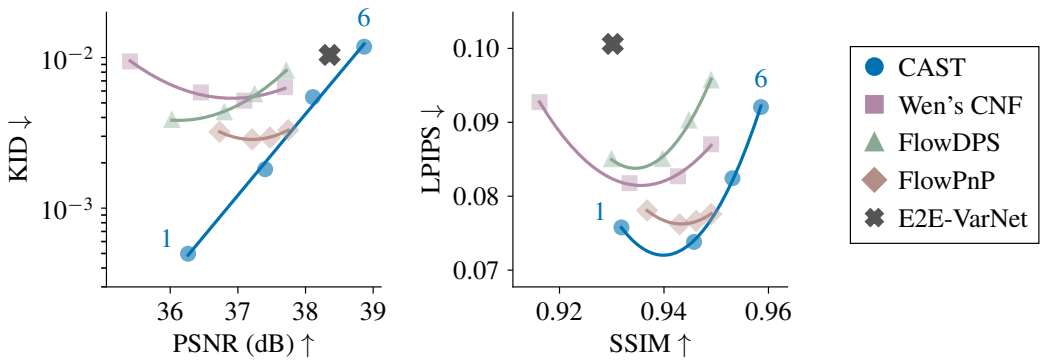

Figure 2: Perception–distortion trade-off on fastMRI brain. Markers show single-sample reconstructions and averages of two, four, and six samples; E2E-VarNet is shown as a low-distortion reference. Averaging moves the generative methods toward higher PSNR and SSIM, while single-sample CAST reconstructions achieve the lowest KID. CAST provides a more favorable trade-off than the evaluated generative baselines. Curves show quadratic fits to the measured points.

Table 3: Reconstruction performance on fastMRI brain. For each generative method and metric, the sampling or averaging procedure is selected on validation data. CAST achieves strong performance across both perceptual and distortion metrics and reaches PSNR comparable to E2E-VarNet. Best scores are bold and second-best scores are underlined within each dataset. KID is multiplied by 1000.

	Homogeneous fastMRI				Diverse fastMRI			
Method	PSNR (dB) ↑	SSIM ↑	LPIPS ↓	KID ↓	PSNR (dB) ↑	SSIM ↑	LPIPS ↓	KID ↓
E2E-VarNet	<u>38.36</u>	0.930	0.101	10.41	<u>41.58</u>	0.953	0.099	9.59
Wen	37.70	0.949	0.081	5.22	–	–	–	–
FlowDPS	37.72	0.949	0.085	3.86	40.93	0.964	0.079	2.33
FlowPnP	37.74	<u>0.949</u>	<u>0.076</u>	<u>2.82</u>	41.19	<u>0.966</u>	**0.069**	<u>1.29</u>
CAST	**38.87**	**0.959**	**0.074**	**0.44**	**41.61**	**0.971**	<u>0.071</u>	**0.37**

We report PSNR and SSIM as distortion metrics, and LPIPS and KID with VGG-16 features as perceptual metrics. KID measures the discrepancy between the feature distributions of reconstructed and fully sampled reference images. Deep-feature metrics with VGG-16 features have been shown to correlate better with radiologist assessments of MRI image quality than conventional distortion metrics like PSNR and SSIM (Kastryulin et al., 2023; Adamson et al., 2025). We use unrolled networks as low-distortion references.

5.1 Using acquisition information during training

We first study whether learning the acquisition-conditional reconstruction distribution with flow matching directly improves over reconstruction with pretrained generative priors, and whether it matters how the acquisition information is incorporated.

We consider $4\times$ accelerated fastMRI brain reconstruction with equispaced undersampling. We compare conventional flow matching (Section 3.1), where acquisition information is provided only through network inputs, with CAST, which additionally structures the source distribution and velocity field using the acquisition operators. We compare both approaches with FlowDPS (Kim et al., 2025), FlowPnP (Martin et al., 2025), and the conditional normalizing flow of Wen et al. (2023). E2E-VarNet (Sriram et al., 2020) is included as a low-distortion reference.

Table 2 compares single-sample reconstructions. Input-only conditioning underperforms the pretrained prior baselines. In contrast, incorporating the acquisition operators into the source distribution and velocity field substantially improves reconstruction quality. CAST outperforms input-only conditioning across PSNR, SSIM, LPIPS, and KID, and outperforms pretrained priors on KID.

Table 4: Reconstruction performance on Stanford 3D at $8\times$ acceleration. For each generative method and metric, the sampling or averaging procedure is selected on validation data. CAST achieves the best PSNR, SSIM, LPIPS, and KID among the evaluated methods. Best scores are bold and second-best scores are underlined. KID is multiplied by 1000.

Method	PSNR (dB) ↑	SSIM ↑	LPIPS ↓	KID ↓
E2E-VarNet	39.39	0.918	0.209	22.37
Variational	38.74	0.922	0.165	15.82
TPDM-DDS	<u>39.61</u>	<u>0.934</u>	<u>0.151</u>	<u>10.20</u>
CAST	**39.89**	**0.938**	**0.146**	**3.29**

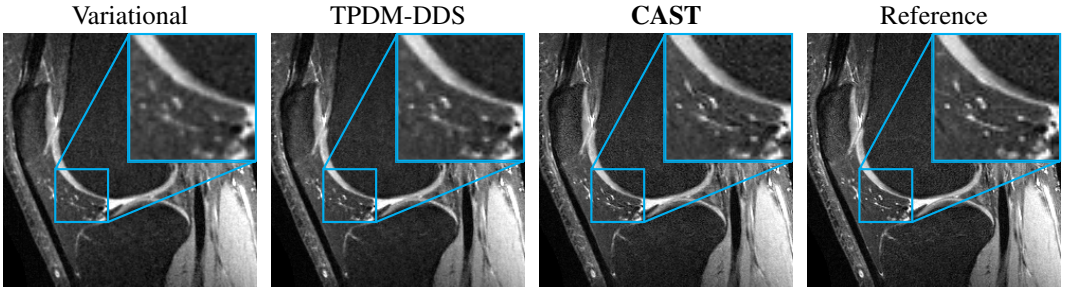

Figure 3: Example Stanford 3D reconstructions. CAST preserves finer detail than the evaluated generative-prior baselines.

However, generative reconstruction is not limited to a single output: multiple samples can also be combined to trade perceptual similarity against distortion. We next examine how this trade-off changes across methods.

5.2 Perception–Distortion Trade-off

Sampling-based reconstruction methods can trade perceptual similarity against distortion without retraining. Individual samples can retain high perceptual similarity to the reference image distribution, whereas averaging multiple samples can improve distortion metrics like PSNR at the cost of decreased perceptual similarity (Blau & Michaeli, 2018; Bendel et al., 2023). We therefore evaluate individual reconstructions and averages of multiple samples.

Figure 2 plots KID against PSNR, and LPIPS against SSIM as a function the number of samples being averaged. CAST's single-sample reconstructions yield the lowest KID, while averaging increases PSNR and SSIM at the cost of decreased perceptual similarity. Compared with the evaluated generative baselines, CAST shows higher flexibility in balancing the trade-off between perceptual and distortion metrics.

To summarize reconstruction performance, we select the sampling or averaging procedure for each method and metric on validation data and report the corresponding test performance in Table 3. CAST outperforms the evaluated generative baselines across PSNR, SSIM, and KID, with its PSNR competitive with E2E-VarNet.

Figure 1 illustrates the corresponding reconstruction differences. Single-sample CAST reconstructions retain finer image structure than the evaluated baselines.

5.3 Accelerated 3D MRI

We next evaluate CAST on 3D MRI reconstructions. We consider Stanford 3D knee data (Epperson et al., 2013) at $8\times$ acceleration with a 2D Poisson undersampling mask. The readout axis is fully sampled, such that $\boldsymbol{F}$ is the spatial 2D Fourier transform on the two phase-encoding axes and the mask $\boldsymbol{M}$ is shared across the readout axis.

Table 5: Cardiac cine reconstruction on OCMR breath-hold and free-breathing acquisitions. Free-breathing data are evaluated using hold-out k-space SER (dB) because fully sampled references are unavailable. CAST uses the validation-selected sampling procedure for each metric and achieves the best reported score across all metrics. Best scores are bold and second-best scores are underlined. KID is multiplied by 1000.

	Breath-hold				Free-breathing
Method	PSNR (dB) ↑	SSIM ↑	LPIPS ↓	KID ↓	SER (dB)↑
CineVN	<u>36.32</u>	<u>0.927</u>	<u>0.103</u>	<u>63.74</u>	15.45
PDDR	31.86	0.860	0.166	94.68	<u>16.03</u>
CAST	**37.92**	**0.952**	**0.075**	**29.49**	**16.08**

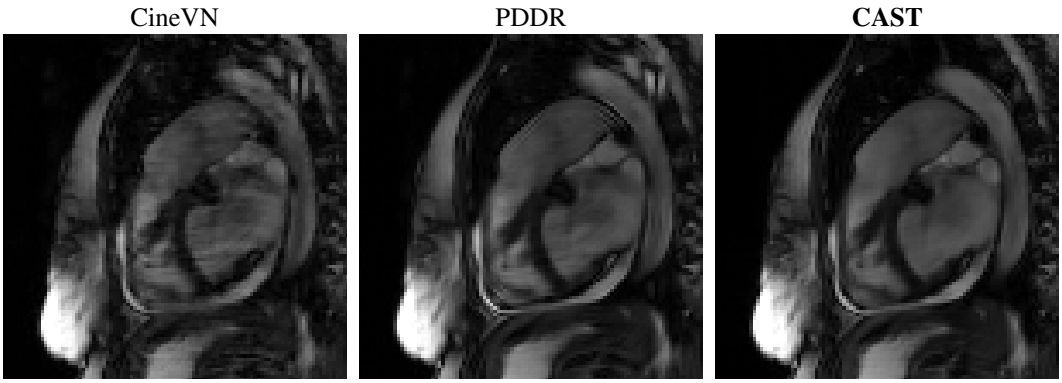

Figure 4: Example free-breathing cardiac cine reconstruction. No fully sampled reference is available. CineVN and PDDR show visible streaking. CAST reduces these artifacts substantially. Complete cine sequences are provided in the supplementary material.

CAST uses adjacent slices along the readout axis as network inputs following Fabian et al. (2022); we use the same adjacent-slice module for E2E-VarNet. We compare with two diffusion-prior methods for 3D MRI reconstruction: the variational approach of Krainovic et al. (2024) and TPDM-DDS (Lee et al., 2023; Krainovic et al., 2024). Appendix B.2 provides additional setup details.

Table 4 uses the same validation-based, per-metric selection procedure as Section 5.2. CAST achieves the best PSNR, SSIM, LPIPS, and KID among the evaluated methods, showing that the benefits observed in 2D reconstruction extend to 3D acquisitions.

Figure 3 shows example reconstructions. CAST preserves finer image structure than the evaluated generative prior baselines.

5.4 Accelerated cardiac cine MRI

We next evaluate CAST for dynamic MRI reconstruction using OCMR cardiac cine data (Chen et al., 2020), including prospectively undersampled acquisitions. Here, $\boldsymbol{F}$ denotes the two-dimensional spatial Fourier transform applied independently to each frame, and the sampling mask $\boldsymbol{M}$ varies across frames. The breath-hold test cines contain 15–25 frames and span 0.6–0.9 s, whereas the free-breathing test cines contain 30–197 frames and span 1.5–10 s.

We train on breath-hold cine, for which fully sampled k-space is available, and evaluate on both retrospectively undersampled breath-hold and prospectively undersampled free-breathing acquisitions. The free-breathing data use golden ratio offset (GRO) sampling (Joshi et al., 2022); we use the same sampling pattern when retrospectively undersampling the breath-hold training data to reduce the acquisition-related distribution shift between training and free-breathing reconstruction.

CAST is compared with the unrolled network CineVN (Vornehm et al., 2025) and the diffusion-prior method PDDR (Fürnrohr & Heckel, 2026). CAST backbone adopts the temporal U-Net architecture used by PDDR. Appendix B.3 provides additional setup details.

Table 5 reports results on breath-hold and free-breathing data. On breath-hold cine, CAST achieves the best PSNR, SSIM, LPIPS, and KID among the evaluated methods. Free-breathing acquisitions do not provide fully sampled references, so we instead hold out 10% of the acquired k-space measurements and evaluate their prediction using signal-to-error ratio (SER) (Fürnrohr & Heckel, 2026). CAST also achieves the highest hold-out SER in this setting.

Figure 4 shows an example free-breathing reconstruction. CineVN exhibits visible streaking across the heart, while CAST reduces them substantially. Complete reconstructed cine sequences are provided in the supplementary material.

6 Conclusion

We study how acquisition information can be used during training for accelerated MRI reconstruction with flow matching. In our experiments, providing this information only through network inputs does not improve over reconstruction with pretrained image priors, whereas structuring the source distribution and velocity field with acquisition operators substantially improves the conditional model. The resulting approach, CAST, matches or exceeds state-of-the-art unrolled networks in PSNR through sample averaging, while individual samples provide perceptual quality that unrolled networks cannot. This yields a high degree of flexibility in balancing perceptual similarity and distortion at inference. Qualitative results show that CAST reconstructs fine image detail and reduces visible artifacts in challenging dynamic MRI acquisitions.

AI Use Statement

In this work, we used generative AI tools to assist with setting up the code base for running experiments described in Section 5. We also used AI-tools for reviewing the proof in Appendix A, for reviewing writing, figure polishing, and identifying relevant literature and baseline methods. We inspected LLM-generated code, and reviewed writing and proof suggestions. We take full responsibility for the manuscript and associated research artifacts, including all AI-assisted content.

Ethics Statement

This work studies MRI reconstruction using existing publicly available datasets: fastMRI brain (Zbontar et al., 2018), Stanford 3D (Epperson et al., 2013), and OCMR (Chen et al., 2020). We use these datasets in accordance with their respective licenses and data-use agreements. No new participant data were collected for this study.

Reproducibility Statement

All datasets used in this work are publicly available. Appendix A provides the theoretical argument supporting the CAST flow matching construction. Appendix B describes the experimental setup, training and inference procedures, hyperparameters, and evaluation protocols. We will make the code for reproducing the experiments publicly available.

References

Philip M. Adamson, Arjun D. Desai, Jeffrey Dominic, Maya Varma, Christian Bluethgen, Jeff P. Wood, Ali B. Syed, Robert D. Boutin, Kathryn J. Stevens, Shreyas Vasanawala, John M. Pauly, Beliz Gunel, and Akshay S. Chaudhari. Using deep feature distances for evaluating the perceptual quality of MR image reconstructions. *Magnetic Resonance in Medicine*, 2025.

Hemant K. Aggarwal, Merry P. Mani, and Mathews Jacob. MoDL: Model-based deep learning architecture for inverse problems. *IEEE Transactions on Medical Imaging*, 2019.

Matthew Bendel, Rizwan Ahmad, and Philip Schniter. A regularized conditional GAN for posterior sampling in image recovery problems. In *Advances in Neural Information Processing Systems*, 2023.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Chong Chen, Yingmin Liu, Philip Schniter, Matthew Tong, Karolina Zareba, Orlando Simonetti, Lee Potter, and Rizwan Ahmad. OCMR (v1.0): Open-access multi-coil k-space dataset for cardiovascular magnetic resonance imaging. arXiv:2008.03410, 2020.

Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated MRI. *Medical Image Analysis*, 2022.

Hyungjin Chung, Suhyeon Lee, and Jong Chul Ye. Decomposed diffusion sampler for accelerating large-scale inverse problems. In *International Conference on Learning Representations*, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021.

Kevin Epperson, Anne Marie Sawyer, Michael Lustig, Marcus Alley, Martin Uecker, Patrick Virtue, Peng Lai, and Shreyas Vasanawala. Creation of fully sampled MR data repository for compressed sensing of the knee. In *SMRT Annual Meeting*, 2013.

Zalan Fabian, Berk Tinaz, and Mahdi Soltanolkotabi. HUMUS-Net: Hybrid unrolled multi-scale network architecture for accelerated MRI reconstruction. In *Advances in Neural Information Processing Systems*, 2022.

Florian Fürnrohr and Reinhard Heckel. Piecewise dynamic diffusion regularization for reconstruction of cardiac cine MRI. arXiv:2607.03299, 2026.

Kerstin Hammernik, Jo Schlemper, Chen Qin, Jinming Duan, Ronald M. Summers, and Daniel Rueckert. Systematic evaluation of iterative deep neural networks for fast parallel MRI reconstruction with sensitivity-weighted coil combination. *Magnetic Resonance in Medicine*, 2021.

Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G. Dimakis, and Jonathan I. Tamir. Robust compressed sensing MRI with deep generative priors. In *Advances in Neural Information Processing Systems*, 2021.

Mihir Joshi, Aaron Pruitt, Chong Chen, Yingmin Liu, and Rizwan Ahmad. Technical report (v1.0)–pseudo-random cartesian sampling for dynamic mri. arXiv:2206.03630, 2022.

Sergey Kastryulin, Jamil Zakirov, Nicola Pezzotti, and Dmitry V. Dylov. Image quality assessment for magnetic resonance imaging. *IEEE Access*, 2023.

Jeongsol Kim, Bryan Sangwoo Kim, and Jong Chul Ye. FlowDPS: Flow-driven posterior sampling for inverse problems. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Anselm Krainovic, Stefan Ruschke, and Reinhard Heckel. Resolution-robust 3D MRI reconstruction with 2D diffusion priors: Diverse-resolution training outperforms interpolation. arXiv:2412.18584, 2024.

Suhyeon Lee, Hyungjin Chung, Minyoung Park, Jonghyuk Park, Wi-Sun Ryu, and Jong Chul Ye. Improving 3d imaging with pre-trained perpendicular 2d diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code. arXiv:2412.06264, 2024.

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

Ségolène Martin, Anne Gagneux, Paul Hagemann, and Gabriele Steidl. PnP-Flow: Plug-and-play image restoration with flow matching. In *International Conference on Learning Representations*, 2025.

Matthew J. Muckley, Bruno Riemenschneider, Alireza Radmanesh, Sunwoo Kim, Geunu Jeong, Jingyu Ko, Yohan Jun, Hyungseob Shin, Dosik Hwang, Mahmoud Mostapha, Simon Arberet, Dominik Nickel, Zaccharie Ramzi, Philippe Ciuciu, Jean-Luc Starck, Jonas Teuwen, Dimitrios Karkalousos, Chaoping Zhang, Anuroop Sriram, Zhengnan Huang, Nafissa Yakubova, Yvonne W. Lui, and Florian Knoll. Results of the 2020 fastMRI challenge for machine learning MR image reconstruction. *IEEE Transactions on Medical Imaging*, 2021.

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 – MICCAI 2020*, 2020.

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.

Martin Uecker, Frank Ong, Jonathan I. Tamir, Dara Bahri, Patrick Virtue, Joseph Y. Cheng, Tao Zhang, and Michael Lustig. Berkeley advanced reconstruction toolbox. In *Proceedings of the International Society for Magnetic Resonance in Medicine*, 2015.

Marc Vornehm, Jens Wetzl, Daniel Giese, Florian Fürnrohr, Jianing Pang, Kelvin Chow, Rolf Gebker, Rizwan Ahmad, and Florian Knoll. CineVN: Variational network reconstruction for rapid functional cardiac cine MRI. *Magnetic Resonance in Medicine*, 2025.

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*, 2023.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J. Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, Nafissa Yakubova, James Pinkerton, Duo Wang, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastMRI: An open dataset and benchmarks for accelerated MRI. arXiv:1811.08839, 2018.

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

A Validity of the CAST Flow Matching Construction

Fix the acquisition $(\boldsymbol{y}^{\mathrm{obs}}, \boldsymbol{M}, \boldsymbol{S})$ throughout this section; all distributions and expectations below are conditional on it.

For a target sample X_1 with corresponding fully sampled data Y^{full}, the target decomposition in Equation (8) gives

$$X_1 - \boldsymbol{x}^{\mathrm{zf}} = \boldsymbol{B}^{\mathsf{H}} Y^{\mathrm{full}} \in \mathrm{range}(\boldsymbol{B}^{\mathsf{H}}). \tag{14}$$

Likewise, the CAST source in Equation (9) satisfies

$$X_0 - \boldsymbol{x}^{\mathrm{zf}} = \boldsymbol{B}^{\mathsf{H}} \boldsymbol{B} Z \in \mathrm{range}(\boldsymbol{B}^{\mathsf{H}}). \tag{15}$$

Thus both endpoints lie in the affine space $\boldsymbol{x}^{\mathrm{zf}} + \mathrm{range}(\boldsymbol{B}^{\mathsf{H}})$, and consequently

$$X_1 - X_0 \in \mathrm{range}(\boldsymbol{B}^{\mathsf{H}}). \tag{16}$$

From Section 3.1, the marginal flow matching velocity is

$$\boldsymbol{u}(\boldsymbol{x}, t) = \mathbb{E}\left[X_1 - X_0 \mid X_t = \boldsymbol{x}\right]. \tag{17}$$

Since $\mathrm{range}(\boldsymbol{B}^{\mathsf{H}})$ is a linear space, the conditional expectation also lies in this space. Therefore,

$$\boldsymbol{u}(\boldsymbol{x}, t) \in \mathrm{range}(\boldsymbol{B}^{\mathsf{H}}). \tag{18}$$

Since $\mathrm{range}(\boldsymbol{B}^{\mathsf{H}}\boldsymbol{B}) = \mathrm{range}(\boldsymbol{B}^{\mathsf{H}})$, the marginal velocity can be written in the CAST form

$$\boldsymbol{u}(\boldsymbol{x}, t) = \boldsymbol{B}^{\mathsf{H}} \boldsymbol{B} \boldsymbol{h}(\boldsymbol{x}, t) \tag{19}$$

for some image-space field $\boldsymbol{h}$.

Thus restricting the velocity to the CAST parameterization does not exclude the marginal minimizer of the conditional flow matching loss. Under the conditions stated in Section 3.1, this marginal velocity transports the CAST source distribution to the target conditional distribution.

B Experiment Details

Unrolled networks training. All experiments are performed on NVIDIA A40 GPUs.

Unrolled networks training. To have unrolled networks output accurate complex valued reconstructions, we train them with mean squared error between output and the MVUE references. SSIM is not directly optimized, so the reported SSIM may underestimate the performance achievable with an SSIM loss.

Training duration. Models train until performance on a validation set no longer improves.

Hyperparameters. Table 6 lists the inference hyperparameters and tuning grids; iterative methods based on trained image priors typically have two to four hyperparameters that should be tuned. For CAST, we tune only the number of integration steps. For static MRI, after selecting checkpoints and solver settings by validation PSNR, we select the sampling procedure separately for each metric as described in Section 5.2.

Sensitivity maps. We estimate coil sensitivity maps using BART (Uecker et al., 2015) from the central 24×24 k-space region for 2D data and the central $24 \times 24 \times 24$ region for 3D data. For cardiac cine, maps are fixed across frames. Following (Vornehm et al., 2025; Fürnrohr & Heckel, 2026), we estimate them from temporally averaged k-space.

B.1 fastMRI experiments

B.1.1 Setup Details

Models and Training. CAST has 34M parameters on both datasets. FlowPnP and FlowDPS use the same unconditional flow-matching architecture with 128M parameters.

Table 6: Inference hyperparameters and tuning grids for the evaluated methods. Braces denote tuned values and single entries denote fixed values. Diffusion time is normalized; DC denotes data consistency.

Method	Hyperparameters and tuning grids	Count
FlowPnP (Martin et al., 2025)	Iterations $\{50, 60, 75\}$; DC steps $\{1, 2\}$; initial DC stepsize 1; decay $\{0.15, 0.25\}$.	4
FlowDPS (Kim et al., 2025)	Iterations $\{60, 75, 90\}$; DC steps $\{4, 5, 6\}$; DC stepsize $\{1.7, 1.9, 2.1\}$.	3
Variational (Krainovic et al., 2024)	Early stopping iteration $\{5, 10 \ldots, 200\}$; stepsize $\{0.01, 0.03, 0.1, 0.3\}$; regularization $\{0.003, 0.01, 0.03\}$; max diffusion time 0.4.	4
TPDM-DDS (Lee et al., 2023; Krainovic et al., 2024)	Reverse steps $\{40, 60, 80, 100, 120\}$; DDIM η $\{0.5, 0.95, 1\}$; DDS γ $\{0.5, 0.95, 2\}$; CG iterations $\{2, 4, 8, 12\}$.	4
PDDR (Fürnrohr & Heckel, 2026)	Early stopping iteration $\{5, 10, \ldots, 200\}$; stepsize 1; regularization $\{0.005, 0.009, 0.03, 0.06\}$; max diffusion time 0.4.	4
CAST	Integration steps $\{3, 5, 10\}$.	1

E2E-VarNet uses the published fastMRI architecture (Sriram et al., 2020; Muckley et al., 2021) with 12 cascades and 30M parameters.

CAST, E2E-VarNet, and the unconditional prior use Adam with learning rate 10^{-4}. All models use optimizer parameters $(\beta_1, \beta_2) = (0.9, 0.999)$ and $\epsilon = 10^{-8}$, and gradient clipping at norm one.

Our implementation of Wen's CNF has 136M parameters. Following Wen et al. (2023), we first pretrain its conditioning network to predict missing k-space measurements. We then jointly train the conditioning network and flow by conditional maximum likelihood. Both stages use Adam, with learning rates $3 \cdot 10^{-3}$ and $3 \cdot 10^{-4}$, respectively.

On the homogeneous set, CAST training takes 13 GPU hours, 18 GPU hours for E2E-VarNet and the unconditional prior, and 30 GPU hours for Wen's CNF. On the diverse set, CAST training takes 30 GPU hours, while E2E-VarNet and the unconditional prior take around 40 GPU hours.

Validation. We select E2E-VarNet, Wen's CNF and CAST checkpoints by validation PSNR based on 8 to 16 validation volumes. For FlowPnP and FlowDPS, we tune inference solver settings on four to eight validation volumes. We select checkpoints separately for FlowPnP and FlowDPS based on validation PSNR.

Inference. CAST uses five Heun ODE steps per sample on both datasets because the PSNR gain from increasing to ten steps was marginal. For the homogeneous and diverse datasets, FlowPnP uses 60 and 50 iterations, respectively, with two DC steps, initial DC stepsize 1, and decay 0.25. FlowDPS uses 75 and 60 iterations, respectively, with five DC steps and DC stepsizes 1.7 and 1.9.

Reconstruction Time. One reconstruction takes 0.7 s for CAST, 0.1 s for Wen's CNF, 13.2 s for FlowDPS, 10.3 s for FlowPnP, and 0.1 s for E2E-VarNet. Times are measured per slice and per sample for sampling-based methods.

Evaluation. We compute reconstruction accuracy on a square center crop whose side length is the smaller image dimension. We first average reconstruction accuracy within each volume and then average over subjects.

B.2 STANFORD 3D

B.2.1 SETUP DETAILS

Models and Training. CAST and E2E-VarNet use adjacent slice reconstruction following Fabian et al. (2022). E2E-VarNet uses the standard fastMRI architecture (Sriram et al., 2020). The models have 34M and 30M parameters, respectively. Each network receives the center slice and its two neighbors along the readout axis, with the loss evaluated on the center slice. The diffusion prior has 34M parameters and uses the same U-Net backbone as CAST. All models use Adam with learning rate 10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and gradient clipping at norm one.

Training takes 10, 14, and 10 GPU hours for CAST, E2E-VarNet, and the diffusion prior, respectively.

Validation. We use the separate validation volume to select model checkpoints and tune inference hyperparameters by PSNR.

Inference. CAST uses five Heun ODE steps; increasing to ten steps gave marginal PSNR gains. TPDM-DDS uses 80 reverse steps, DDIM $\eta = 1$, DDS $\gamma = 2$, and four CG iterations. The variational method early stops at 135 Adam iterations, stepsize 0.1, regularization weight 0.03, maximum diffusion time 0.4, and 50 slices per view for the prior term.

Reconstruction Time. One volume takes 76.9 s per CAST sample, 688.2 s per TPDM-DDS sample, 22.6 s for E2E-VarNet, and 1189.0 s for the complete variational reconstruction.

Evaluation. We omit the first and last 40 slices along the coronal and sagittal axes and three slices along the axial axis because they are mainly empty, leaving a $176 \times 240 \times 314$ central crop. For each view, we average slice scores within each volume and then over the four volumes.

B.3 OCMR

B.3.1 SETUP DETAILS

Training Data Preparation. We form fully sampled MVUE references from breath-hold k-space and retrospectively undersample the measurements. We use golden ratio offset (GRO) sampling masks (Joshi et al., 2022) because GRO is the acquisition pattern used for free-breathing data, reducing the difference between training and free-breathing acquisition setups.

For each training cine, we precompute 32 acquisition draws in a fixed bank and sample from this bank during training. Each draw uses a GRO mask and readout partial-Fourier pattern matched to the free-breathing acquisition metadata.

For free-breathing evaluation, we hold-out 10% of acquired phase-encoding lines independently in each frame. We apply the same removal procedure to the training masks, so the training mask distribution matches the reconstruction inputs after hold-out removal. Hold-out lines are excluded from the free-breathing reconstruction input, data consistency, and sensitivity-map estimation.

We augment breath-hold cines by changing the starting frame and continuing cyclically through the complete cine. The acquisition bank fixes this temporal order for each draw. We also apply spatial reflections and a random global complex phase shared across all frames (Vornehm et al., 2025).

Models and Training. We train CAST on complete breath-hold cines using Adam with learning rate 10^{-4}.

We compare with CineVN (Vornehm et al., 2025) and PDDR (Fürnrohr & Heckel, 2026) and follow the authors' recommended architecture and training setup.

Validation. We select model checkpoints by PSNR on 10 breath-hold validation volumes.

Inference. CAST uses ten Heun ODE steps. PDDR uses 200 iterations, stepsize 1, maximum diffusion time 0.4, and regularization weights 0.009 for breath-hold and 0.03 for free-breathing.

Table 7: Paired 95% confidence intervals for **CAST** minus each baseline on fastMRI. KID differences are multiplied by 1000. Bold intervals include zero.

Homogeneous fastMRI				
Baseline	PSNR (dB)	SSIM	LPIPS	KID
E2E-VarNet	[0.46, 0.56]	[0.027, 0.030]	[-0.029, -0.025]	[-12.03, -8.11]
Wen's CNF	[1.04, 1.31]	[0.009, 0.011]	[-0.009, -0.006]	[-5.69, -3.94]
FlowDPS	[1.06, 1.25]	[0.009, 0.010]	[-0.012, -0.010]	[-4.24, -2.67]
FlowPnP	[1.03, 1.22]	[0.009, 0.010]	[-0.004, -0.002]	[-2.96, -1.86]
Diverse fastMRI				
Baseline	PSNR (dB)	SSIM	LPIPS	KID
E2E-VarNet	**[-0.01, 0.09]**	[0.016, 0.019]	[-0.030, -0.027]	[-10.35, -8.14]
FlowDPS	[0.61, 0.76]	[0.006, 0.008]	[-0.010, -0.008]	[-2.34, -1.61]
FlowPnP	[0.37, 0.48]	[0.004, 0.005]	[0.001, 0.002]	[-1.09, -0.75]

Table 8: Paired 95% confidence intervals for **CAST** minus each baseline on OCMR. KID differences are multiplied by 1000. Bold intervals include zero.

	Breath-hold				Free breathing
Baseline	PSNR (dB)	SSIM	LPIPS	KID	SER (dB)
CineVN	[0.61, 2.56]	[0.010, 0.041]	[-0.048, -0.012]	[-69.54, -7.97]	[0.36, 0.93]
PDDR	[3.39, 9.05]	[0.040, 0.156]	[-0.137, -0.055]	[-101.30, -32.41]	**[-0.16, 0.25]**

Reconstruction Time. For a breath-hold cine with 23 frames, CAST takes 13.4 s per sample, CineVN and PDDR take 2.4 s and 114.0 s, respectively. For a free-breathing cine with 92 frames, CAST takes 70.4 s per sample, CineVN takes 10.7 s and PDDR takes 246.4 s.

Evaluation. Breath-hold metrics use a fixed 120×120 center crop. We evaluate 64 slice-cines from 18 scans. We treat scans from the same acquisition session as one subject. We average frame scores within each slice-cine, slice-cines within scans, scans within subjects, and then give the eight subjects equal weight. LPIPS and KID independently min–max normalize each supported reconstruction and reference over the whole cine.

For each free-breathing scan, let $\boldsymbol{y}_{\text{hold}}$ denote the hold-out measurements and $\widehat{\boldsymbol{y}}_{\text{hold}} = \boldsymbol{M}_{\text{hold}}\boldsymbol{F}\boldsymbol{S}\widehat{\boldsymbol{x}}$ their prediction from the reconstruction $\widehat{\boldsymbol{x}}$, where $\boldsymbol{M}_{\text{hold}}$ selects the hold-out samples. We compute

$$\text{SER} = 10 \log_{10} \left(\frac{\|\boldsymbol{y}_{\text{hold}}\|_2^2}{\|\boldsymbol{y}_{\text{hold}} - \widehat{\boldsymbol{y}}_{\text{hold}}\|_2^2} \right).$$

The squared norms sum over all hold-out complex samples across frames and coils within a scan. We then average SER over the 20 subjects.

C Confidence Intervals

We compute paired bootstrap confidence intervals for the difference between CAST and each baseline, keeping the trained models and validation-selected sampling procedures fixed. We sample subjects with replacement and recompute the mean difference, using the same subjects for both methods. We repeat this process 10000 times and use the 2.5th and 97.5th percentiles as the bounds of the 95% confidence interval.

Tables 7 and 8 report CAST-minus-baseline intervals. Positive differences favor CAST for PSNR, SSIM, and SER; negative differences favor CAST for LPIPS and KID. Bold intervals contain zero, indicating that the paired 95% bootstrap confidence interval does not exclude a zero mean difference between CAST and the baseline. This does not imply that their performance is the same.

Table 9: Minimum and maximum paired differences between **CAST** and each baseline across the four Stanford3D test volumes. Each metric uses its validation-selected procedure.

Baseline	PSNR (dB)	SSIM	LPIPS
E2E-VarNet	[0.37, 0.65]	[0.013, 0.029]	[-0.067, -0.059]
Variational	[1.06, 1.25]	[0.013, 0.017]	[-0.020, -0.017]
TPDM-DDS	[0.23, 0.31]	[0.003, 0.005]	[-0.005, -0.003]

Stanford3D. Since only four test volumes are available, we report the minimum and maximum paired differences in Table 9, within each volume. CAST improves PSNR, SSIM, and LPIPS over each baseline on every volume. KID compares image features across volumes and gives one score for the full test set, so it does not provide per-volume differences.