SIREM: SPEECH-INFORMED MRI RECONSTRUCTION WITH LEARNED SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing speech rtMRI from highly undersampled acquisitions remains challenging because the rapid articulatory dynamics of speech demand both high spatial and temporal resolution. Yet speech rtMRI is acquired together with an information source that conventional reconstruction methods largely ignore: the synchronized acoustic signal. We introduce SIREM, a speech-informed MRI reconstruction framework that exploits this cross-modal relationship as an auxiliary prior. Because vocal-tract configurations are closely coupled to the acoustics they produce, part of the image content can be inferred from speech while the acquired k-space data provide complementary anatomical information. SIREM therefore decomposes each reconstructed frame into audio-driven and MRI-driven components that are combined through an anatomically derived spatial weighting map. The audio branch predicts articulator-related structure from synchronized speech, whereas the MRI branch recovers complementary content from the measured k-space samples. We further introduce a learnable soft weighting profile over spiral arms, allowing the contribution of individual k-space arms to be optimized jointly with the multimodal reconstruction. This provides a unified formulation for audio-guided prediction, MRI reconstruction, and k-space arm reweighting. We evaluate SIREM on the USC 75-Speaker speech rtMRI benchmark against conventional reconstruction approaches, including wavelet-based compressed sensing, total variation, and a deep learning method. SIREM achieves 44–53$\times$ higher reconstruction throughput than the iterative methods while preserving anatomically plausible vocal-tract structure, and achieves real-time computational throughput, with per-frame compute within a 30-fps budget. Our results provide an initial benchmark for speech-informed rtMRI reconstruction and demonstrate the potential of synchronized acoustics as an auxiliary prior for accelerating dynamic speech MRI.

1 INTRODUCTION

Real-time magnetic resonance imaging (rtMRI) is an imaging modality for visualizing the dynamics of speech production (Nayak et al., 2022). Unlike electromagnetic articulography (Berry, 2011) or ultrasound (Klein et al., 2013), rtMRI captures the full midsagittal vocal tract, including the tongue, lips, velum, pharynx, and laryngeal structures, without ionizing radiation or constraining natural articulation. This makes rtMRI valuable for both speech science, where it supports articulatory and phonetic analysis (Ramanarayanan et al., 2018), and clinical applications, including the assessment and treatment planning of cleft lip and palate, velopharyngeal insufficiency, dysarthria, apraxia of speech, and post-surgical speech impairments (Hagedorn et al., 2017).

A central challenge in rtMRI is reconstructing high-quality images from rapidly acquired, heavily undersampled data. To satisfy temporal-resolution constraints, speech rtMRI acquisitions often measure only a fraction of k-space per frame, turning reconstruction into a strongly ill-posed inverse problem. Parallel imaging methods such as SENSE and GRAPPA exploit coil redundancy, while compressed-sensing (CS) MRI combines data consistency with sparsity priors such as wavelets or total variation (TV) (Pruessmann et al., 1999; Griswold et al., 2002; Lustig et al., 2007). Speech-specific dynamic reconstruction methods further exploit temporal structure through low-rank, manifold, or subspace models (Fu et al., 2015; Rusho et al., 2024; Cao et al., 2025). These approaches have been effective, but they rely primarily on measurement-domain, image-domain, or temporal priors and do not explicitly use the synchronized speech waveform available during acquisition.

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.

Speech rtMRI is intrinsically multimodal: each acquisition includes a synchronized audio waveform in addition to the MRI measurements (Narayanan et al., 2014). The acoustic signal is not merely an annotation, but a physical consequence of the vocal-tract configuration that the MRI seeks to image. Articulatory studies show that the tongue, lips, jaw, and velum shape the spectral envelope, formant trajectories, and voicing structure of speech (Ramanarayanan et al., 2018), and recent audio-conditioned rtMRI synthesis and inversion studies suggest that substantial articulatory information can be recovered from speech alone (Udupa & Ghosh, 2023; Nguyen et al., 2025; Shi et al., 2025; Pérez-Toro et al., 2026). Thus, the audio channel contains structured cues about the underlying image content that are typically discarded by reconstruction pipelines.

These observations motivate a different view of speech rtMRI reconstruction: rather than treating it as a purely physics-driven inverse problem, we formulate it as a multimodal estimation problem in which synchronized speech provides an auxiliary prior on anatomically relevant image regions. Based on this perspective, we propose SIREM, a speech-informed MRI reconstruction framework that combines physics-based reconstruction with audio-driven prediction. SIREM decomposes each frame into an audio-driven component and an MRI-driven component, fuses them through an anatomically derived spatial weighting map that encodes which regions are assumed to be predictable from acoustics, and introduces a learnable soft spiral-arm profile that reweights the relative contribution of individual k-space arms to study how k-space arm usage interacts with speech-informed fusion.

Our contributions are:

- We formulate speech-informed rtMRI reconstruction as a multimodal inverse problem that combines synchronized speech with undersampled multi-coil spiral k-space.
- We introduce SIREM, a hybrid reconstruction model that fuses an audio-driven image estimate with an MRI-driven reconstruction through a spatial weighting map and a learnable soft spiral-arm reweighting profile.
- We demonstrate the feasibility of direct speech-informed rtMRI reconstruction and characterize its fidelity–efficiency trade-off relative to classical iterative and learned reconstruction methods.

SIREM is not introduced as a uniform fidelity improvement over classical iterative reconstruction. Instead, the contribution of SIREM is to show that multimodal direct reconstruction is feasible, that synchronized audio can serve as a useful auxiliary prior, and that the resulting throughput regime is qualitatively different from iterative baselines.

2 Related Work

Speech rtMRI has become a central modality for observing the moving vocal tract because it captures the full midsagittal anatomy, including structures that are difficult to measure with electromagnetic articulography or ultrasound. Public speech-rtMRI resources, particularly the USC 75-Speaker speech MRI corpus (Lim et al., 2021), provide synchronized audio, raw multi-coil non-Cartesian k-space, and reconstructed image sequences.

Classical accelerated MRI reconstruction is built on parallel imaging and compressed sensing. SENSE and GRAPPA exploit spatial encoding redundancy across receiver coils, while compressed-sensing MRI combines data consistency with sparsity-promoting priors such as wavelets and total variation (Pruessmann et al., 1999; Griswold et al., 2002; Lustig et al., 2007). Speech-specific rtMRI reconstruction has also explored dynamic priors including temporal finite-difference regularization, low-rank-plus-sparse models, manifold regularization, and subspace-based reconstruction for vocal-tract imaging (Fu et al., 2015; Rusho et al., 2024; Cao et al., 2025). These approaches improve reconstruction by exploiting measurement-domain or temporal structure, but they do not use the synchronized waveform as an explicit prior.

More recent MRI reconstruction has shifted toward physics-informed deep learning and unrolled optimization. Methods such as MoDL and variational networks (E2E-VN) combine learned image priors with data-consistency steps and have become standard references for learned reconstruction in accelerated MRI (Aggarwal et al., 2018; Hammernik et al., 2018; Sriram et al., 2020). A related direction studies joint optimization of sampling and reconstruction, treating acquisition design as a learnable component rather than a fixed heuristic (Bahadir et al., 2019; Zhang et al., 2019). SIREM is

conceptually related to this line of work, but addresses a distinct setting in which synchronized audio provides side information for reconstruction from raw spiral k-space.

Synchronized speech has also been used for vocal-tract modeling in generation, inversion, and segmentation tasks, including acoustic-to-articulatory inversion, audio-conditioned rtMRI synthesis, diffusion-based speech-to-rtMRI generation, and audio-assisted vocal-tract segmentation (Udupa & Ghosh, 2023; Nguyen et al., 2025; Pérez-Toro et al., 2026; Shi et al., 2025; Jain et al., 2024; Liu et al., 2025). These studies support the premise that speech contains substantial articulatory information, but they generally operate in image or video space rather than in the reconstruction setting. SIREM instead targets the complementary problem of using synchronized speech as an auxiliary prior for reconstructing undersampled speech rtMRI directly from raw k-space measurements.

3 SIREM

We consider speech-informed reconstruction for real-time MRI from undersampled non-Cartesian k-space measurements. Our central hypothesis is that part of the articulatory configuration is predictable from synchronized speech audio, while the remaining anatomy must be recovered from measured MRI data. Following this decomposition, we model each frame as a fusion of an audio-driven estimate and an MRI-driven reconstruction, and we learn a soft reweighting profile over the spiral arms that modulates their relative contribution to the MRI-driven estimate.

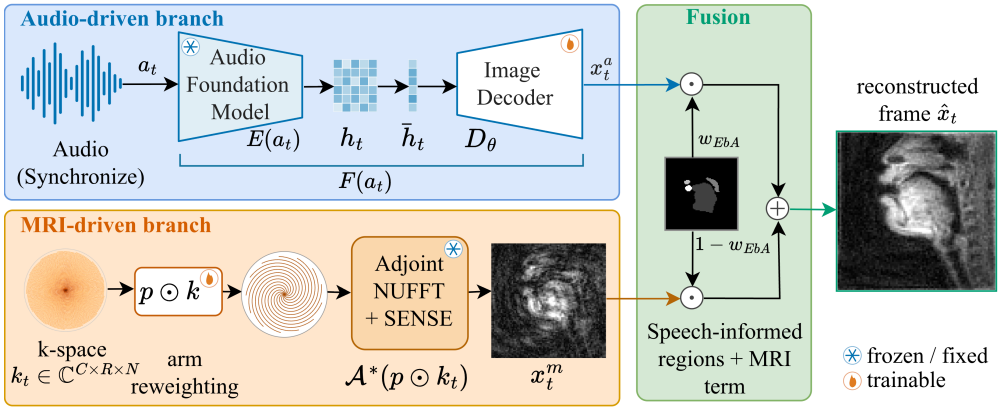


Figure 1: Overview of SIREM. The audio branch maps synchronized speech a_t to an audio-driven estimate x_t^a, while the measurement branch applies the soft spiral-arm profile p to undersampled multi-coil k-space k_t and reconstructs the MRI-driven estimate x_t^m. Pixelwise fusion under the explained-by-audio map w_{EbA} yields the final frame $\hat{x}_t$.

Figure 1 illustrates the resulting computational graph. The audio prior path maps the synchronized speech segment a_t to an audio-driven estimate x_t^a through a frozen HuBERT encoder and a trainable image decoder. In parallel, the measurement branch applies the learnable soft spiral-arm profile p to the acquired multi-coil k-space k_t and reconstructs the MRI-driven estimate $x_t^m = \mathcal{A}^*(p \odot k_t)$ by adjoint NUFFT reconstruction and SENSE coil combination. The final frame $\hat{x}_t$ is then obtained by pixelwise fusion of x_t^a and x_t^m under the explained-by-audio map w_{EbA}.

3.1 Problem Formulation

Let $x_t \in [0,1]^{H\times W}$ denote the target midsagittal frame, $a_t \in \mathbb{R}^L$ the synchronized speech segment, and $k_t \in \mathbb{C}^{C\times R\times N}$ the acquired multi-coil spiral k-space for frame t, where C is the number of coils, $R = 13$ is the number of spiral arms per full rotation, and N is the number of readout samples per arm. During training, these frame-wise quantities are batched over samples. The goal is to estimate x_t from (k_t, a_t) and a fixed map w_{EbA}.

SIREM is formulated as three coupled operators: an audio prediction branch, an MRI reconstruction branch, and a spatial fusion map. The audio branch maps synchronized speech to an audio-driven

estimate,

$$x_t^a = F(a_t), \tag{1}$$

where $F(\cdot)$ denotes the audio foundation model followed by the image decoder. The MRI branch applies a learnable soft spiral-arm profile $p \in [0,1]^R$ to the acquired measurements and reconstructs an MRI-driven estimate,

$$x_t^m = \mathcal{A}^*(p \odot k_t), \tag{2}$$

where $\odot$ denotes elementwise multiplication with broadcasting over coils and readout samples, and $\mathcal{A}^*$ denotes the SENSE-weighted adjoint NUFFT operator that maps multi-coil k-space to a single coil-combined image.

The final reconstruction is obtained by pixelwise fusion of the two estimates,

$$\hat{x}_t = w_{\mathrm{EbA}} \odot x_t^a + (1 - w_{\mathrm{EbA}}) \odot x_t^m, \tag{3}$$

where $w_{\mathrm{EbA}} \in [0,1]^{H \times W}$ is the explained-by-audio spatial weighting map. Larger values of w_{EbA} increase the contribution of the audio-driven estimate in acoustically predictable regions, whereas smaller values preserve measurement-driven recovery through x_t^m.

3.2 Dataset

Experiments are conducted on the USC 75-Speaker speech rtMRI corpus (Lim et al., 2021), which provides synchronized speech audio, raw k-space measurements, and reconstructed midsagittal vocal-tract image sequences acquired on a 1.5T scanner using a non-Cartesian spiral trajectory. We adopt a subject-independent split of 50 training, 7 validation, and 15 test subjects. Three subjects are excluded due to missing audio and corrupted/unreadable files. The validation set is used for model selection and hyperparameter tuning, whereas the test set is used only for final evaluation. Articulator masks are obtained with the segmentation network of Liu et al. (2025), retrained exclusively on external data. Since the reconstruction frames released with the dataset are themselves TV reconstructions, we use CG-SENSE reconstructions of each full-rotation 13-arm window as reference and training target (Sun et al., 2020; Maier et al., 2021; Lee et al., 2022; Yaman et al., 2020).

Preprocessing and temporal alignment. Each training instance consists of synchronized k-space measurements, trajectories, speech audio, reference reconstructions, and segmentation masks w_{EbA}. Audio is resampled to 16 kHz, and image-domain inputs are reoriented to a consistent midsagittal view, resized to 84×84, and normalized to $[0,1]$.

The acquisition uses 13 spiral arms per full rotation. In our setting, each training sample corresponds to one 13-arm acquisition window, yielding an effective temporal resolution of 12.81 fps. Since the released TV reconstructions are available at the 2-arm rate (83.28 fps), we align each 13-arm window with the temporally centered 2-arm frame. Let c_t denote the temporal center of the t-th 13-arm acquisition window. The aligned reference timestamp is defined as

$$\tau_t = \arg\min_{\tau \in \mathcal{T}_{\text{2-arm}}} |\tau - c_t| \tag{4}$$

where $\mathcal{T}_{\text{2-arm}}$ denotes the timestamps of the 2-arm reference sequence. The associated audio segment is extracted around the same center using a symmetric context window,

$$a_t = a[c_t - \delta : c_t + \delta], \tag{5}$$

with half-window δ corresponding to 20 acquisition frames at the 13-arm rate (about ± 1.56 s of context centered on c_t).

K-space representation and sensitivity estimation. For each utterance, the acquired spiral arms are grouped into consecutive blocks of 13, yielding frame-wise k-space tensors of shape $[T, C, 13, N]$, where T denotes the number of frames, C the number of coils, and N the number of readout samples per arm. For frame t, we write

$$k_t \in \mathbb{C}^{C \times R \times N}, \tag{6}$$

where the second dimension indexes the $R = 13$ spiral arms and the third dimension indexes the readout samples along each arm. The corresponding non-Cartesian sampling trajectories are arranged as

$$\omega_t \in \mathbb{R}^{13 \times N \times 2}, \tag{7}$$

where each trajectory sample specifies a two-dimensional k-space coordinate. Density compensation weights are taken from the trajectory metadata. K-space values are normalized by their maximum magnitude,

$$\tilde{k}_t = \frac{k_t}{\max_{c,i,n} |k_t[c,i,n]|}, \tag{8}$$

and trajectories are scaled to the reconstruction grid. Coil sensitivity maps are estimated once per utterance by averaging k-space over full rotations, reconstructing low-resolution coil images with an adjoint NUFFT, and applying Walsh-based sensitivity estimation (Walsh et al., 2000).

3.3 Speech-informed reconstruction model

Audio-driven branch. The audio-driven branch predicts an articulator-focused image estimate from synchronized speech. Let a_t denote the waveform segment centered at time t. We first extract contextual speech features using a frozen HuBERT encoder (Hsu et al., 2021),

$$h_t = E_{\mathrm{HuBERT}}(a_t), \qquad h_t \in \mathbb{R}^{T_h \times d}, \tag{9}$$

where T_h is the number of latent time steps and d is the feature dimension. The latent sequence is then mean-pooled over time,

$$\bar{h}_t = \frac{1}{T_h} \sum_{\ell=1}^{T_h} h_{t,\ell}, \tag{10}$$

and mapped to image space by a lightweight multilayer decoder,

$$x_t^a = F(a_t) = \mathrm{clip}_{[0,1]}\big(D_\theta(\bar{h}_t)\big), \qquad x_t^a \in [0,1]^{H \times W}, \tag{11}$$

with $H = W = 84$ in our implementation, and $\mathrm{clip}_{[0,1]}$ denoting elementwise clipping to the unit interval to keep x_t^a in the same range as x_t^m before fusion. The decoder consists of three fully connected layers with hidden dimensions 1024 and 2048, LayerNorm, GELU activations, and dropout, followed by a final linear projection to the image grid. This branch is intended to capture articulator configurations that are strongly coupled to acoustics, particularly the tongue, lips, and velum.

MRI-driven branch. The MRI-driven branch produces the estimate x_t^m from undersampled multi-coil k-space. Let $p \in (0,1)^R$ denote the learnable soft weighting profile over the $R = 13$ spiral arms. For frame t, the weights are broadcast across coils and readout samples and applied to the k-space tensor as

$$\tilde{k}_t[c,i,n] = p_i \, k_t[c,i,n], \tag{12}$$

where c indexes coils, i indexes spiral arms, and n indexes readout samples along each arm. The weighted measurements are then mapped to image space with an adjoint NUFFT-based gridding operator (Ong & Lustig, 2019). Using coil sensitivity maps S_c, the MRI-driven reconstruction is written as

$$x_t^m = \mathcal{A}^*(p \odot k_t) = \sum_{c=1}^{C} S_c^* \, \mathcal{N}^*\left(\tilde{k}_t^{(c)}\right), \tag{13}$$

where $\mathcal{N}^*$ denotes the adjoint non-Cartesian Fourier operator and S_c^* is the complex conjugate of the sensitivity map for coil c. The resulting reconstruction is normalized to $[0,1]$ on a per-sample basis before fusion with the audio-driven estimate. This branch is responsible for preserving image content that remains anchored to measured MRI data.

Spatial weighting map. The explained-by-audio map modulates the relative contribution of the audio-driven estimate x_t^a and the MRI-driven estimate x_t^m at each pixel. Let

$$w_{\mathrm{EbA}} \in [0,1]^{H \times W} \tag{14}$$

denote the explained-by-audio map, where larger values indicate regions assumed to be more predictable from synchronized speech. The final reconstruction is obtained by pixelwise fusion,

$$\hat{x}_t = w_{\mathrm{EbA}} \odot x_t^a + (1 - w_{\mathrm{EbA}}) \odot x_t^m, \tag{15}$$

where $\odot$ denotes elementwise multiplication. Larger values of w_{EbA} increase reliance on the acoustically predictable component, whereas smaller values preserve measurement-driven recovery

through x_t^m. In the reported experiments, w_{EbA} is not learned but is constructed once per utterance from anatomical masks, as follows. (i) The initial full-rotation 13-arm window $k_1^{(u)}$ of utterance u is reconstructed using the same coil sensitivities and settings as for the reference reconstructions. (ii) This frame is segmented with the network of Liu et al. (2025). (iii) $w_{\mathrm{EbA}}^{(u)}$ is set to the union of the masks of the articulators assumed to be acoustically predictable and is kept fixed for all frames of u. Thus, only the initial frame of each utterance is reconstructed and segmented to build the map automatically; all frames are then reconstructed by SIREM. This imposes an explicit anatomically interpretable prior, stabilizes training, and isolates the contribution of speech-informed fusion. Although the implementation also supports a learnable predictor for w_{EbA} from fused audio and MRI features, that variant is not used in the reported experiments.

3.4 Learnable sampling

The learnable spiral-arm profile p is optimized jointly with the reconstruction model and enters the method through the MRI-driven pathway $x_t^m = \mathcal{A}^*(p \odot k_t)$. We parameterize this profile with trainable logits $\ell \in \mathbb{R}^{13}$ and define

$$p = \sigma(\ell), \qquad p \in (0,1)^{13}, \tag{16}$$

where $\sigma(\cdot)$ denotes the sigmoid function applied elementwise. The resulting vector assigns a continuous weight to each of the 13 spiral arms and provides a differentiable mechanism for modulating their relative contribution to reconstruction. In the reported experiments, p is applied retrospectively to already acquired k-space data and should therefore be interpreted as a differentiable surrogate for arm importance rather than as a prospectively deployed acquisition policy. Although the implementation also supports hard top-k arm selection via a straight-through estimator, all reported results use the soft weighting profile throughout training and inference, so no arm is discarded.

3.5 Training objective

Let $\mathcal{D} = \{(k_t, a_t, x_t, w_{\mathrm{EbA}})\}_{t=1}^T$ denote the training set, where k_t is the multi-coil k-space measurement, a_t is the synchronized audio segment, x_t is the reference frame, and w_{EbA} is the explained-by-audio spatial weighting map. Here w_{EbA} denotes the map of the utterance containing frame t. The objective optimizes the fused reconstruction $\hat{x}_t$ while regularizing the spiral-arm profile p; since w_{EbA} is fixed, it enters the objective only through $\hat{x}_t$. We minimize the expected loss

$$\mathcal{L} = \mathbb{E}_{(k_t,a_t,x_t,w_{\mathrm{EbA}})\sim\mathcal{D}}\left[\mathcal{L}_{\mathrm{recon}} + \alpha\mathcal{L}_{\mathrm{psf}} + \beta\mathcal{L}_{\mathrm{budget}}\right]. \tag{17}$$

The reconstruction term enforces fidelity between the fused estimate $\hat{x}_t$ and the reference frame x_t:

$$\mathcal{L}_{\mathrm{recon}} = \|\hat{x}_t - x_t\|_2^2. \tag{18}$$

To regularize the learned spiral-arm weighting profile p, we penalize the non-DC energy of its 1-D inverse Fourier transform along the arm dimension. Let $\mathcal{F}_R^{-1}$ denote the 1-D inverse discrete Fourier transform over the $R = 13$ spiral-arm axis. We define

$$\mathcal{L}_{\mathrm{psf}} = \frac{1}{R}\sum_{i=1}^{R-1}\left|\left[\mathcal{F}_R^{-1}(p)\right]_i\right|^2, \tag{19}$$

which is proportional to $\sum_i (p_i - \bar{p})^2$. It discourages angular concentration of the soft profile, while the reconstruction term can still learn relative arm importance. To control the overall arm mass, we define

$$\mathcal{L}_{\mathrm{budget}} = \left(\sum_{i=1}^{13} p_i - K\right)^2 + \frac{1}{13}\sum_{i=1}^{13} p_i, \tag{20}$$

where both terms depend only on $\sum_i p_i$ and thus fix the scale of p without promoting sparsity; together with $\mathcal{L}_{\mathrm{psf}}$ they favor a uniform profile, and all 13 arms contribute ($p_i > 0$). We use $K = 2$ in all reported experiments.

The implementation also contains a smoothness regularizer on w_{EbA} intended for the learnable-map variant. Since w_{EbA} is fixed in all reported experiments, that term is constant with respect to all trainable parameters, has zero gradient, and is therefore omitted from Eq. equation 17. In all experiments, we set $\alpha = 0.1$ and $\beta = 0.01$.

Implementation details. The HuBERT encoder is frozen during training, and only the audio decoder and sampling logits are optimized. We use AdamW with learning rate 10^{-4}, weight decay 10^{-5}, batch size 16, and cosine annealing over 40 epochs. Gradients are clipped to a maximum norm of 1.0. Validation is performed after every epoch, and the final model is selected according to validation peak signal-to-noise ratio (PSNR). All experiments were conducted on a single NVIDIA Quadro RTX 5000 GPU (16 GB).

4 Experiments and Results

We evaluate and compare SIREM against iterative and deep-learning-based reconstruction methods commonly used in accelerated MRI: wavelet-based CS, TV reconstruction, and E2E-VN. These comparisons test whether synchronized speech can serve as a useful prior under undersampled acquisition. Reconstruction quality is assessed using both distortion-based and perceptual metrics, including peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) (Wang et al., 2004), normalized root mean squared error (NRMSE), high-frequency error norm (HFEN) (Ravishankar & Bresler, 2010), learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018), visual information fidelity (VIF) (Sheikh & Bovik, 2006), and Fréchet Inception Distance (FID) (Heusel et al.). Metrics are computed sequence-wise and averaged over the evaluation set. When reconstructed and reference sequences have different frame counts, the longer one is linearly resampled in time to match the shorter one before metric computation, so that all comparisons are made at a single temporal grid.

4.1 Results

We report results under two evaluation protocols. In the first, reconstructions are compared against the reference reconstruction at the native 13-arms-per-frame acquisition rate, which is also the training target of the learned methods (Table 1). In the second, reconstructions are compared against the TV reconstructions, which no method sees during training, to measure how closely learned methods match a classical iterative baseline when the reference target is changed. Additional ablations, including audio-representation comparisons, are reported in Appendix K.

Table 1: Unified comparison under two evaluation targets: (i) the CG-SENSE reference and (ii) the released TV reconstruction, used only for evaluation. Results against TV measure agreement with that reconstruction, not anatomical fidelity. Overall, the results highlight how measured reconstruction quality changes with the evaluation target.

Method	SSIM ↑	PSNR ↑	NRMSE ↓	HFEN ↓	LPIPS ↓	VIF ↑	FID ↓
Target: CG-SENSE reference							
Zero-filled	0.223	13.349	0.787	0.936	0.444	0.043	310.963
Wavelet	0.612	23.050	0.258	0.590	0.277	**0.234**	172.992
TV	**0.635**	23.032	0.258	**0.574**	0.287	0.079	98.002
E2E-VN	0.599	23.752	0.239	0.620	0.173	0.223	83.148
SIREM w/o audio	0.570	22.899	0.264	0.683	0.180	0.197	80.476
SIREM	0.606	**23.914**	**0.235**	0.612	**0.163**	0.222	**74.156**
Target: Total Variation							
Zero-filled	0.254	14.326	0.788	0.944	0.455	0.038	297.2
Wavelet	0.659	**24.488**	**0.245**	0.601	0.199	0.226	160.1
E2E-VN	0.573	22.274	0.317	0.732	0.151	0.197	169.1
SIREM w/o audio	0.649	23.821	0.269	0.536	0.077	0.293	91.6
SIREM	**0.677**	23.824	0.264	**0.535**	**0.075**	**0.297**	**91.3**

The results reveal a fidelity–throughput trade-off rather than a uniform quality gain. Table 1 reports performance under two reference targets: the CG-SENSE reference and the TV reconstruction. SIREM improves on PSNR, NRMSE, LPIPS, and FID on CG-SENSE reference reconstruction. Across both targets, iterative methods remain competitive on structural and fidelity-oriented metrics, confirming that iterative reconstruction baselines still provide strong fidelity in the current setting. However, SIREM consistently improves over its audio-ablated variant, indicating that synchronized speech contributes useful complementary information to reconstruction.

Table 2: Effect of synchronized speech conditioning, measured as the difference between SIREM and SIREM (w/o audio). Positive Δ is better for SSIM, PSNR, and VIF; negative Δ is better for HFEN, NRMSE, LPIPS, and FID. Audio conditioning improves all seven metrics under both targets.

Target	ΔSSIM	ΔPSNR	ΔNRMSE	ΔHFEN	ΔLPIPS	ΔVIF	ΔFID
Reference reconstruction	+0.036	+1.01	-0.029	-0.071	-0.017	+0.025	-6.32
Total Variation	+0.028	+0.003	-0.005	-0.001	-0.002	+0.004	-0.30

Table 2 quantifies this effect: against the CG-SENSE reference, speech conditioning improves all seven metrics (SSIM $+0.036$, PSNR $+1.01$, VIF $+0.025$) while reducing NRMSE (-0.029), HFEN (-0.071), LPIPS (-0.017), and FID (-6.32). Against the TV target, all seven metrics also improve, with the largest gain in SSIM ($+0.028$).

Taken together, these results establish an initial quantitative benchmark for speech-informed rtMRI reconstruction, showing that synchronized speech serves as a useful auxiliary prior although iterative baselines remain stronger on structural metrics.

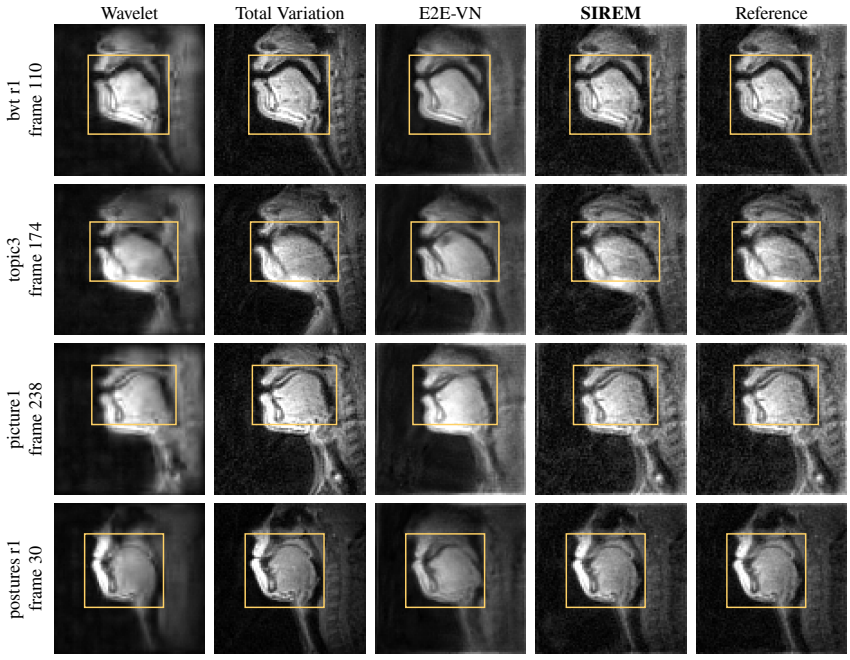


Figure 2: Qualitative comparison of four frames from one held-out test subject, each from a different speech task. Wavelet is visibly blurred, TV retains edges with grainy residuals, and E2E-VN is smooth but softens articulator boundaries. SIREM preserves clearer articulatory boundaries while using feed-forward inference. Row labels give the task and boxes mark the articulator region.

Figure 2 shows speaking a bVt word ("but"), a spoken answer to a question ("when"), a picture description ("if"), and a non-speech posture (mouth open wide), spanning dynamic and static articulator configurations. Wavelet-based compressed sensing produces smooth reconstructions but visibly blurs fine anatomical detail. Total-variation reconstruction preserves edges more faithfully while leaving a grainy residual in homogeneous regions. E2E-VN is smoother but softens the articulator boundaries. SIREM recovers sharper tongue and vocal-tract contours than E2E-VN while maintaining a direct feed-forward inference pipeline.

4.2 Efficiency analysis

Beyond reconstruction fidelity, an important objective of the proposed method is computational efficiency. Classical CS baselines such as wavelet reconstruction and total-variation reconstruction rely on iterative optimization, which increases runtime and limits throughput in accelerated settings.

In contrast, SIREM performs direct speech-informed reconstruction once the model is trained, avoiding iterative inference at test time. This makes it better aligned with real-time applications, where reconstruction speed is a practical constraint in addition to image quality.

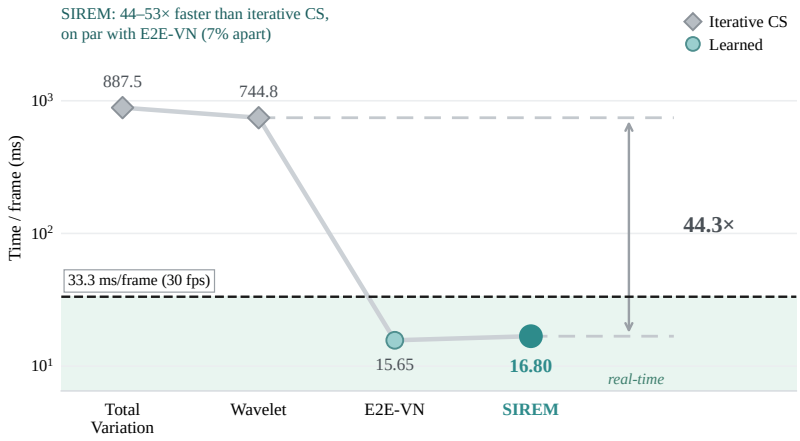


Figure 3: Runtime analysis of reconstruction methods on the test set. Markers show mean time per frame. The dashed line and shaded region mark the 30 fps real-time budget (33.3 ms/frame). Both learned methods, E2E-VN and SIREM, reconstruct within that budget, whereas the iterative CS methods are 44–53× slower.

Figure 3 reports per-frame reconstruction time and inference type for each method. SIREM reconstructs each frame in ∼16.80 ms, which is ∼44× faster than Wavelet (∼744.8 ms/frame), ∼53× faster than Total Variation (∼887.5 ms/frame), and matches E2E-VN to within 7%. Iterative methods require 100 optimization iterations and remain above 700 ms per frame. Among the nontrivial reconstruction methods evaluated here, only SIREM and E2E-VN achieve real-time computational throughput.

5 Discussion and Conclusions

This work introduces SIREM, a speech-informed reconstruction framework for rtMRI that uses synchronized audio as an auxiliary prior during reconstruction. Across both evaluation targets, classical compressed-sensing baselines remain strong on distortion-based metrics. SIREM improves over its audio-ablated variant on all metrics under both evaluation targets, and runs in a higher-throughput regime than iterative methods. The main contribution of SIREM is therefore not a uniform fidelity gain, but a multimodal reconstruction formulation with a distinct fidelity–efficiency trade-off.

Several limitations remain. The current model uses a fixed segmentation-derived explained-by-audio map rather than a learned fusion predictor; because the map is computed once from the initial frame of each utterance, it does not follow articulator motion within the utterance. The spiral-arm profile is evaluated as a retrospective soft reweighting of all acquired arms rather than as a sparse or prospective acquisition policy. The evaluation is limited to one public benchmark and one acquisition setting. Future work should study learned and temporally adaptive fusion maps, stronger MRI reconstruction backbones, sparsity-promoting and prospective sampling strategies, causal audio windows, and further ablations of the fusion mechanism and arm profile.

Potential clinical use also requires caution. Faster speech rtMRI reconstruction could reduce scan time and improve practicality, but learned reconstructions should not be used for clinical decision-making without prospective validation across speakers, disorders, scanners, and acquisition settings.

AI USE STATEMENT

In this work, generative artificial intelligence tools were used solely to assist with language editing and clarity of presentation. All research ideas, methodology, experiments, and interpretations were conceived and carried out by the authors, who take full responsibility for the originality, validity, and integrity of the work.

REPRODUCIBILITY STATEMENT

All experiments use the USC 75-Speaker speech rtMRI corpus, with the subject-independent train/validation/test split and the three excluded subjects described in Section 3.2. Preprocessing of k-space and audio, coil-sensitivity estimation, model architecture (audio branch, MRI branch, fusion weighting map, and learnable spiral-arm profile), and the training objective are described in Sections 3–3.5. Evaluation protocols and metrics are reported in Section 4. To facilitate exact replication, we will release the complete training and evaluation code (including data-loading scripts, configuration files, and random seeds) upon acceptance of the paper.

REFERENCES

Hemant K Aggarwal, Merry P Mani, and Mathews Jacob. Modl: Model-based deep learning architecture for inverse problems. *IEEE transactions on medical imaging*, 38(2):394–405, 2018.

Cagla Deniz Bahadir, Adrian V Dalca, and Mert R Sabuncu. Learning-based optimization of the under-sampling pattern in mri. In *international conference on information processing in medical imaging*, pp. 780–792. Springer, 2019.

Jeffrey J Berry. Accuracy of the ndi wave speech research system. *Journal of Speech, Language, and Hearing Research*, 54(5):1295–1301, 2011.

Peng Cao, Wenting Jiang, Changhe Chen, Yiang Wang, and Jonathan Havenhill. Self-navigated subspace reconstruction for real-time mr imaging of the vocal tract. *Magnetic Resonance Imaging*, 115:110243, 2025.

Maojing Fu, Bo Zhao, Christopher Carignan, Ryan K Shosted, Jamie L Perry, David P Kuehn, Zhi-Pei Liang, and Bradley P Sutton. High-resolution dynamic speech imaging with joint low-rank and sparsity constraints. *Magnetic Resonance in Medicine*, 73(5):1820–1832, 2015.

Mark A Griswold, Peter M Jakob, Robin M Heidemann, Mathias Nittka, Vladimir Jellus, Jianmin Wang, Berthold Kiefer, and Axel Haase. Generalized autocalibrating partially parallel acquisitions (grappa). *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 47(6):1202–1210, 2002.

Christina Hagedorn, Michael Proctor, Louis Goldstein, Stephen M Wilson, Bruce Miller, Maria Luisa Gorno-Tempini, and Shrikanth S Narayanan. Characterizing articulation in apraxic speech using real-time magnetic resonance imaging. *Journal of Speech, Language, and Hearing Research*, 60 (4):877–891, 2017.

Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P. Recht, Daniel K. Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated mri data. *Magnetic Resonance in Medicine*, 79(6):3055–3071, 2018. doi: 10.1002/mrm.26977.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.

Rishi Jain, Bohan Yu, Peter Wu, Tejas Prabhune, and Gopala Anumanchipalli. Multimodal segmentation for vocal tract modeling. *arXiv preprint arXiv:2406.15754*, 2024.

Harriet B Klein, Tara McAllister Byun, Lisa Davidson, and Maria I Grigos. A multidimensional investigation of children's/r/productions: Perceptual, ultrasound, and acoustic measures. *American Journal of Speech-Language Pathology*, 22(3):540–553, 2013.

Nam G Lee, Rajiv Ramasawmy, Yongwan Lim, Adrienne E Campbell-Washburn, and Krishna S Nayak. Maxgirf: image reconstruction incorporating concomitant field and gradient impulse response function effects. *Magnetic resonance in medicine*, 88(2):691–710, 2022.

Yongwan Lim, Asterios Toutios, Yannick Bliesener, Ye Tian, Sajan Goud Lingala, Colin Vaz, Tanner Sorensen, Miran Oh, Sarah Harper, Weiyi Chen, et al. A multispeaker dataset of raw and reconstructed speech production real-time mri video and 3d volumetric images. *Scientific data*, 8 (1):187, 2021.

Daiqi Liu, Johannes Enk, Maureen Stone, Fangxu Xing, Tomás Arias-Vergara, Jerry L Prince, Jana Hutter, Jonghye Woo, Andreas Maier, and Paula Andrea Pérez-Toro. Vocsegmri: Multimodal learning for precise vocal tract segmentation in real-time mri. *arXiv preprint arXiv:2509.13767*, 2025.

Michael Lustig, David Donoho, and John M Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

Oliver Maier, Steven Hubert Baete, Alexander Fyrdahl, Kerstin Hammernik, Seb Harrevelt, Lars Kasper, Agah Karakuzu, Michael Loecher, Franz Patzig, Ye Tian, et al. Cg-sense revisited: Results from the first ismrm reproducibility challenge. *Magnetic resonance in medicine*, 85(4):1821–1839, 2021.

Shrikanth Narayanan, Asterios Toutios, Vikram Ramanarayanan, Adam Lammert, Jangwon Kim, Sungbok Lee, Krishna Nayak, Yoon-Chul Kim, Yinghua Zhu, Louis Goldstein, et al. Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc). *The Journal of the Acoustical Society of America*, 136(3):1307–1311, 2014.

Krishna S Nayak, Yongwan Lim, Adrienne E Campbell-Washburn, and Jennifer Steeden. Real-time magnetic resonance imaging. *Journal of Magnetic Resonance Imaging*, 55(1):81–99, 2022.

Hong Nguyen, Sean Foley, Kevin Huang, Xuan Shi, Tiantian Feng, and Shrikanth Narayanan. Speech2rtmri: Speech-guided diffusion model for real-time mri video of the vocal tract during speech. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Frank Ong and Michael Lustig. Sigpy: A python package for high performance iterative reconstruction. In *ISMRM Workshop on Data Sampling and Image Reconstruction*, 2019.

Paula Andrea Pérez-Toro, Tomás Arias-Vergara, Fangxu Xing, Xiaofeng Liu, Maureen Stone, Jiachen Zhuo, Juan Rafael Orozco-Arroyave, Elmar Nöth, Jana Hutter, Jerry L Prince, et al. A speech-to-video synthesis approach using spatio-temporal diffusion for vocal tract mri. *Medical Image Analysis*, pp. 104053, 2026.

Klaas P Pruessmann, Markus Weiger, Markus B Scheidegger, and Peter Boesiger. Sense: sensitivity encoding for fast mri. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 42(5):952–962, 1999.

Vikram Ramanarayanan, Sam Tilsen, Michael Proctor, Johannes Töger, Louis Goldstein, Krishna S Nayak, and Shrikanth Narayanan. Analysis of speech production real-time mri. *Computer Speech & Language*, 52:1–22, 2018.

Saiprasad Ravishankar and Yoram Bresler. Mr image reconstruction from highly undersampled k-space data by dictionary learning. *IEEE transactions on medical imaging*, 30(5):1028–1041, 2010.

Rushdi Zahid Rusho, Abdul Haseeb Ahmed, Stanley Kruger, Wahidul Alam, David Meyer, David Howard, Brad Story, Mathews Jacob, and Sajan Goud Lingala. Prospectively accelerated dynamic speech magnetic resonance imaging at 3 t using a self-navigated spiral-based manifold regularized scheme. *NMR in Biomedicine*, 37(8):e5135, 2024.

Hamid R. Sheikh and Alan C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006. doi: 10.1109/TIP.2005.859378.

Xuan Shi, Tiantian Feng, Jay Park, Christina Hagedorn, Louis Goldstein, and Shrikanth Narayanan. Speech acoustics to rt-mri articulatory dynamics inversion with video diffusion model. *Computer Speech & Language*, pp. 101928, 2025.

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 *International conference on medical image computing and computer-assisted intervention*, pp. 64–73. Springer, 2020.

Changyu Sun, Yang Yang, Xiaoying Cai, Michael Salerno, Craig H Meyer, Daniel Weller, and Frederick H Epstein. Non-cartesian slice-grappa and slice-spirit reconstruction methods for multiband spiral cardiac mri. *Magnetic resonance in medicine*, 83(4):1235–1249, 2020.

Sathvik Udupa and Prasanta Kumar Ghosh. Real-time mri video synthesis from time aligned phonemes with sequence-to-sequence networks. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

David O. Walsh, Arthur F. Gmitro, and Michael W. Marcellin. Adaptive reconstruction of phased array mr imagery. *Magnetic Resonance in Medicine*, 43(5):682–690, 2000. doi: 10.1002/(SICI)1522-2594(200005)43:5<682::AID-MRM10>3.0.CO;2-G.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Burhaneddin Yaman, Seyed Amir Hossein Hosseini, Steen Moeller, Jutta Ellermann, Kâmil Uğurbil, and Mehmet Akçakaya. Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data. *Magnetic resonance in medicine*, 84(6):3172–3191, 2020.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Zizhao Zhang, Adriana Romero, Matthew J Muckley, Pascal Vincent, Lin Yang, and Michal Drozdzal. Reducing uncertainty in undersampled mri reconstruction with active acquisition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2049–2058, 2019.

A Appendix

This appendix supplements the main paper with background, additional interpretation, and further experiments. We first review the MRI reconstruction concepts used throughout the paper (k-space, spiral sampling, the adjoint NUFFT, multi-coil acquisition, and the SENSE-weighted adjoint operator) for readers outside the MRI community. We then expand on the design of SIREM: why synchronized speech is a meaningful prior for vocal-tract imaging, how the explained-by-audio fusion should be interpreted, a basic boundedness property of the fusion operator, the role of each regularization term, and why a fixed segmentation-derived weighting map is a reasonable first instantiation. This is followed by additional dataset and preprocessing details. We further report ablations on the speech representation, metrics restricted to the articulator region, and a metric-wise comparison with the iterative baseline and E2E-VN, and finally describe how the segmentation masks for the fusion map are obtained without leakage from the evaluation corpus. No new claims are made in the appendix; all conclusions in the main text are supported by the tables and figures presented there.

B Additional Motivation and Technical Details

This supplementary material expands the background, motivation, formulation, and implementation details of the proposed speech-informed reconstruction framework. In particular, it clarifies why synchronized speech constitutes a meaningful prior for vocal-tract MRI reconstruction, provides

additional interpretation of the fusion and sampling objectives, and summarizes implementation details omitted from the main paper for space reasons.

C MRI Reconstruction Concepts

This section provides brief background on MRI reconstruction concepts used throughout the paper.

K-space. MRI scanners measure samples of the object in the spatial-frequency domain, called *k-space*. In a discrete setting, if $x \in \mathbb{C}^{H \times W}$ denotes an image, its Fourier transform $\mathcal{F}(x)$ is indexed by frequency coordinates $k = (k_x, k_y)$. Cartesian MRI corresponds to sampling $\mathcal{F}(x)$ on a rectangular grid, while accelerated MRI undersamples k-space and reconstructs x by combining a forward model with priors.

Non-Cartesian spiral sampling and spiral arms. In spiral MRI, k-space is sampled along spiral trajectories rather than on a Cartesian grid. A *spiral arm* is one interleaved spiral readout; multiple arms with different angular offsets together form a full spiral rotation. Using fewer arms per frame increases temporal resolution but increases undersampling artifacts, making reconstruction more ill-posed.

NUFFT and the adjoint NUFFT. For non-Cartesian trajectories, the Fourier samples are taken at off-grid locations. The non-uniform fast Fourier transform (NUFFT) implements the forward operator that maps an image to samples along a non-Cartesian trajectory and its adjoint (often implemented via gridding) maps measured non-Cartesian k-space back to an image-like array. In the paper, $\mathcal{N}^*$ denotes this adjoint NUFFT operator; it is not an exact inverse, but a computationally efficient backprojection that is typically combined with coil combination and/or iterative or learned refinement.

Multi-coil acquisition and coil-sensitivity maps. Speech rtMRI is acquired with multiple receiver coils. Each coil observes the same underlying image modulated by a spatially varying complex sensitivity map S_c. A common signal model is

$$y_c = P\,\mathcal{F}(S_c \odot x) + \varepsilon_c, \tag{21}$$

where y_c is the k-space measured by coil c, P is the sampling operator (here, spiral sampling), and ε_c is noise. Coil-sensitivity maps are estimated from calibration data (or low-resolution reconstructions) and are used to combine coil images and enforce data consistency.

SENSE-weighted adjoint operator. Given estimated sensitivities, a standard single-step reconstruction is to apply the adjoint NUFFT to each coil and then combine coils using the conjugate sensitivities, i.e., $\sum_c S_c^* \mathcal{N}^*(y_c)$. This is the coil-combined backprojection used as $\mathcal{A}^*(\cdot)$ in the main text; higher-quality reconstructions can be obtained by iterating or unrolling this physics-based operator together with learned priors.

D Why Speech-Informed Reconstruction is Well Motivated

The proposed method is based on the observation that speech rtMRI is intrinsically multimodal: the audio waveform and the MRI measurements are recorded simultaneously and arise from the same articulatory process. The acoustic signal is therefore not a side channel, but a consequence of the underlying vocal-tract configuration. Positions of the tongue, lips, jaw, and velum are known to shape the spectral envelope, formant trajectories, and voicing structure of speech. This makes audio informative about at least part of the anatomical configuration visible in rtMRI.

At the same time, the relation between acoustics and anatomy is not uniform across the image. Dynamic articulators such as the tongue and lips are strongly coupled to the produced sound, whereas static anatomy, background tissue, and scanner-related intensity structure are less directly predictable from audio alone. This motivates a spatially varying multimodal model rather than a purely audio-driven one. In other words, synchronized speech should be treated as a structured prior on acoustically informative regions, not as a substitute for the entire reconstruction problem.

This view naturally suggests a decomposition into two complementary components: an audio-driven branch that predicts image content that is largely constrained by acoustics, and an MRI-driven branch that reconstructs image content that must remain anchored to measured k-space data. The explained-by-audio map then determines how strongly each location should depend on either source.

E Interpretation of the Fusion Formulation

Let x_t denote the target frame at time t. In SIREM, the audio branch produces the estimate

$$x_t^a = F(a_t), \tag{22}$$

while the MRI branch produces the estimate

$$x_t^m = \mathcal{A}^*(p \odot k_t), \tag{23}$$

where F maps synchronized speech to image space, p is the learnable soft spiral-arm profile, and $\mathcal{A}^*$ denotes the SENSE-weighted adjoint NUFFT operator that maps multi-coil k-space to a single coil-combined image.

Neither branch alone is sufficient. The MRI-only branch ignores the structured multimodal prior contained in the waveform, while the audio-only branch cannot recover image content that is weakly constrained by acoustics. We therefore introduce the spatially varying fusion model

$$\hat{x}_t = w_{\mathrm{EbA}} \odot x_t^a + (1 - w_{\mathrm{EbA}}) \odot x_t^m, \tag{24}$$

where $w_{\mathrm{EbA}} \in [0,1]^{H \times W}$ is the explained-by-audio map. Equation equation 24 can be interpreted as a convex spatial interpolation between an acoustically informed estimate and a measurement-driven reconstruction. Regions with large values of w_{EbA} are assumed to be well explained by speech, whereas regions with small values are reconstructed primarily from MRI data.

This construction imposes a useful inductive bias. Rather than forcing the model to decide globally whether a frame should be reconstructed from audio or MRI, it permits local specialization. This is particularly appropriate in speech MRI, where different anatomical regions contribute differently to the acoustic signal.

F Basic Property of the Fusion Operator

The fusion formulation in Eq. equation 24 satisfies a simple but useful boundedness property.

Proposition 1. Assume $w_{\mathrm{EbA}} \in [0,1]^{H \times W}$ and suppose both x_t^a and x_t^m take values in $[0,1]^{H \times W}$. Then the fused reconstruction $\hat{x}_t$ defined in Eq. equation 24 also lies in $[0,1]^{H \times W}$.

Proof. Consider any pixel location (i,j). Since $w_{\mathrm{EbA}}(i,j) \in [0,1]$, we can write

$$\hat{x}_t(i,j) = w_{\mathrm{EbA}}(i,j)\, x_t^a(i,j) + (1 - w_{\mathrm{EbA}}(i,j))\, x_t^m(i,j). \tag{25}$$

Because both branch outputs lie in $[0,1]$, this expression is a convex combination of two values in $[0,1]$. Therefore $\hat{x}_t(i,j) \in [0,1]$. Since this holds for every pixel, $\hat{x}_t \in [0,1]^{H \times W}$. □

This does not amount to a full theoretical analysis of the method, but it makes explicit that the fusion operator preserves normalized image ranges when both branch outputs are themselves normalized.

G Interpretation of the Learned Sampling Objective

The spiral-arm profile is parameterized by trainable logits $\ell \in \mathbb{R}^{13}$:

$$p = \sigma(\ell), \tag{26}$$

where $\sigma(\cdot)$ is the sigmoid function. This yields a differentiable soft weighting of the 13 spiral arms. In the present implementation, these weights do not alter the physical acquisition, but instead define a differentiable reweighting of the available k-space measurements during reconstruction.

The sampling-related regularizers in the training objective have complementary roles. The budget term

$$\mathcal{L}_{\text{budget}} = \left(\sum_i p_i - K \right)^2 + \text{mean}(p) \tag{27}$$

has two parts. Both parts depend only on $\sum_i p_i$, so they constrain the scale of p but do not encourage sparse arm selection.

The regularizer

$$\mathcal{L}_{\text{psf}} = \frac{1}{R} \sum_{i=1}^{R-1} \left| \left[\mathcal{F}_R^{-1}(p) \right]_i \right|^2 \tag{28}$$

penalizes the non-DC energy of the 1-D inverse Fourier transform of the sampling profile p along the arm dimension. This equals (up to a constant) the variance of p, so it favors a uniform arm weighting rather than a sparse one.

Finally, the regularizer

$$\mathcal{L}_{\text{mask}} = \frac{1}{HW} \sum_{(u,v) \neq (0,0)} \left| \left[\mathcal{F}_{\text{2D}}(1 - w_{\text{EbA}}) \right]_{u,v} \right|^2 \tag{29}$$

promotes spatial smoothness in the complement of the explained-by-audio map by penalizing all non-DC frequency components of $1 - w_{\text{EbA}}$. Since w_{EbA} determines where the model should rely on audio versus MRI measurements, this term discourages implausibly noisy or fragmented spatial switching. This term applies only to a learnable-map variant; with the fixed w_{EbA} used here it has zero gradient and is not used.

H Why a Fixed Weighting Map is a Reasonable Starting Point

In the main experiments, the explained-by-audio map w_{EbA} is not learned jointly but is provided by segmentation-derived masks, computed once per utterance from a reference reconstruction (Section 3.3). This choice was made for two reasons. First, learning both the image prediction branch and the fusion map jointly would introduce additional ambiguity in an already ill-posed reconstruction. Second, the fixed masks impose an interpretable prior: regions associated with dynamic articulators are explicitly marked as more likely to be predictable from acoustics.

This design should be understood as a first instantiation of the multimodal reconstruction idea rather than as its final form. A learned weighting map may ultimately be preferable, but the fixed-mask setting is a useful starting point because it isolates the multimodal fusion mechanism from the challenge of simultaneously discovering where that fusion should occur. Its main limitation is that it is static within an utterance and therefore encodes articulator location rather than frame-by-frame articulator motion.

I Additional Dataset Details

All experiments in the main paper use the USC 75-Speaker speech rtMRI corpus (Lim et al., 2021) with the subject-independent split described in Section 3.2 (50 training, 7 validation, and 15 test subjects). In addition, three subjects are excluded due to missing audio and unreadable files. The dataset provides synchronized speech audio, raw multi-coil non-Cartesian k-space, and reconstructed image sequences acquired on a 1.5T scanner using a spiral trajectory. The data include a wide range of linguistically motivated speech tasks and capture the full moving vocal tract at high temporal resolution. Anatomical masks used to construct w_{EbA} are obtained using the method described in Section 3.2.

J Additional Preprocessing Details

The released TV reconstructions and our wavelet reconstructions are available at the 2-arm rate and are used only for evaluation; each full-rotation 13-arm window is aligned with the temporally

centered 2-arm frame. The CG-SENSE reference is computed per 13-arm window and alignment is not required. Articulator masks are obtained with the segmentation network of Liu et al. (2025), retrained exclusively on external data (USC-TIMIT (Narayanan et al., 2014)). At test time, this network is applied only to the initial frame of each utterance, and that frame is excluded from evaluation. This mirrors standard practice in iterative reconstruction: the released TV pipeline similarly uses the initial frame for calibration and scales its regularization weight by the maximum magnitude of the corresponding zero-filled reconstruction (Lim et al., 2021). In both cases, only measured k-space data from the corresponding utterance are used, and no evaluation target enters the reconstruction.

All image-domain inputs are resized to 84×84 and normalized to $[0, 1]$. K-space values are normalized by their maximum magnitude before frame formation, and trajectories are scaled to the reconstruction grid. Coil sensitivity maps are estimated once per utterance and reused across frames of that utterance.

K Additional Ablation Studies

This appendix complements the main paper with additional ablation experiments. Section K.1 studies speech representations for the audio branch. Section K.2 evaluates reconstruction quality inside the articulator mask. Section K.3 compares SIREM with the strongest reported iterative baseline for each metric and with E2E-VN.

K.1 Audio representation ablation

We keep the MRI branch, fusion map w_{EbA}, and training setup fixed and vary only the speech-feature input. The no-audio setting removes speech features and corresponds to SIREM w/o audio in Tables 1–2. Table 3 compares MFCCs, HuBERT[1], and wav2vec 2.0[2]. Both self-supervised encoders are frozen. MFCCs improve SSIM, PSNR, and NRMSE over no audio but not HFEN; HuBERT and wav2vec 2.0 are nearly identical. We use HuBERT in the main experiments.

Table 3: Audio-representation ablation against the CG-SENSE reference.

Audio input	**SSIM↑**	**PSNR↑**	**NRMSE↓**	**HFEN↓**
No audio	0.570	22.899	0.264	0.683
MFCC	0.587	23.189	0.251	0.692
HuBERT	0.606	**23.914**	**0.235**	0.612
wav2vec 2.0	**0.607**	23.900	**0.235**	**0.610**

K.2 Evaluation inside the articulator mask

To localize the audio contribution, Table 4 reports metrics over pixels inside the articulator regions. Regional SSIM and HFEN use local windows and filters and are not directly comparable with full-image values. SIREM has favorable point estimates over its audio-ablated variant on all four regional metrics: SSIM $+0.020$, PSNR $+0.41$ dB, NRMSE -0.010, and HFEN -0.021.

Table 4: Metrics inside the articulator mask against the CG-SENSE reference.

Model	**SSIM↑**	**PSNR↑**	**NRMSE↓**	**HFEN↓**
SIREM w/o audio	0.725	22.038	0.159	0.484
SIREM	**0.745**	**22.446**	**0.149**	**0.463**

Table 5: SIREM relative to the best iterative baseline per metric and to E2E-VN, evaluated against the CG-SENSE reference. Changes vs. the iterative baseline are relative, except PSNR (absolute, in dB); changes vs. E2E-VN are absolute for all metrics.

Metric	Best iterative	SIREM	Δ vs. iterative	Δ vs. E2E-VN
SSIM↑	0.635 (TV)	0.606	-4.6%	$+0.007$
PSNR↑	23.050 (Wav.)	23.914	$+0.86$ dB	$+0.16$ dB
NRMSE↓	0.258 (both)	0.235	-8.9%	-0.004
HFEN↓	0.574 (TV)	0.612	$+6.6\%$	-0.008
LPIPS↓	0.277 (Wav.)	0.163	-41.2%	-0.010
VIF↑	0.234 (Wav.)	0.222	-5.1%	-0.001
FID↓	98.00 (TV)	74.16	-24.3%	-8.99

K.3 Metric-wise comparison with reported baselines

Table 5 summarizes SIREM relative to the strongest reported iterative point estimate for each metric and to E2E-VN. The table is derived from Table 1; it is not an additional matched-input experiment because the iterative reconstructions use their released native protocol.

L Segmentation Masks for the Fusion Map

We compute w_{EbA} from articulator masks produced by the segmentation network of Liu et al. (2025). To prevent leakage from the USC 75-Speaker corpus, the segmentation model is retrained exclusively on external USC-TIMIT data; no USC 75-Speaker frames, annotations, or checkpoints are used to train it. At test time, only the first measured window of each utterance is reconstructed and segmented, and that frame is excluded from evaluation, as detailed in Section 3.3.

[1] https://huggingface.co/facebook/hubert-base-ls960

[2] https://huggingface.co/facebook/wav2vec2-base