Improving Deep Learning for Accelerated MRI With Data Filtering

Kang Lin **Anselm Krainovic** **Kun Wang** **Reinhard Heckel**
Department of Computer Engineering
Technical University of Munich
{ka.lin, anselm.krainovic, kun2000.wang, reinhard.heckel}@tum.de

Abstract

Deep neural networks achieve state-of-the-art results for accelerated MRI reconstruction. Most research on deep learning based imaging focuses on improving neural network architectures trained and evaluated on fixed and homogeneous training and evaluation data. In this work, we investigate data curation strategies for improving MRI reconstruction. We assemble a large dataset of raw k-space data from 18 public sources consisting of 1.1M images and construct a diverse evaluation set comprising 48 test sets, capturing variations in anatomy, contrast, number of coils, and other key factors. We propose and study different data filtering strategies to enhance performance of current state-of-the-art neural networks for accelerated MRI reconstruction. Our experiments show that filtering the training data leads to consistent, albeit modest, performance gains. These performance gains are robust across different training set sizes and accelerations, and we find that filtering is particularly beneficial when the proportion of in-distribution data in the unfiltered training set is low.

1 Introduction

Deep neural networks achieve state-of-the-art results for accelerated MRI reconstruction [28]. While the majority of existing literature focuses on designing better neural network architectures for improving performance in accelerated MRI [15, 39, 10], research on effective dataset design for improving performance of neural networks for image reconstruction is limited. As a result, best practices for constructing datasets to train high-performing and robust models remain largely unclear.

In contrast, recent works in computer vision and natural language processing show that carefully curated training datasets can significantly boost model performance [11, 29, 12, 14, 19, 32]. For large foundation models, filtering an initial pool of web-scraped data for high-quality samples and training on this refined subset has led to substantial improvements across benchmarks [14, 12, 19].

We treat data as a fundamental part of model development, rather than a fixed resource, and demonstrate that curating training data through filtering candidate datasets can improve performance of existing state-of-the-art neural networks for accelerated MRI. For example, Figure 1 (left) shows for 8-fold accelerated MRI that a VarNet [39] (state-of-the-art for accelerated 2D MRI) trained on a smaller filtered dataset can provide a better reconstruction than the same model trained on the much larger unfiltered dataset. Our main contributions are as follows:

- We propose and investigate a variety of data filtering methods for improving training sets for deep learning based accelerated MRI. Similar to well-performing filtering approaches in the vision-language domain [12, 14], our best performing curation technique is based on retrieving images from the initial unfiltered training set that are similar to the validation data in terms of the DreamSim metric [13].

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

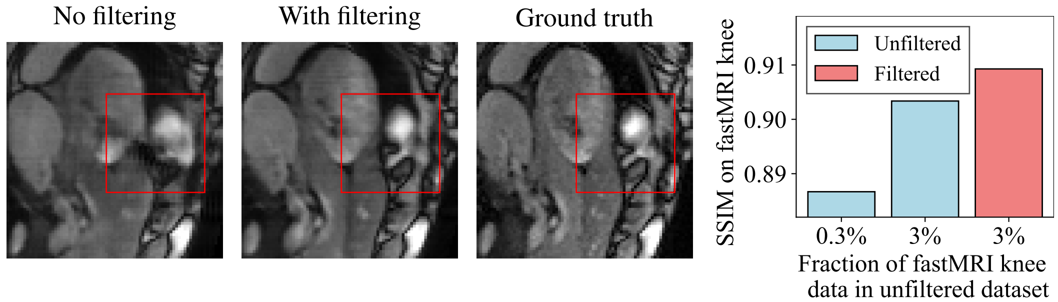

Figure 1: Performance of a VarNet [39] trained on an unfiltered dataset (120k slices) and on a filtered dataset (40k slices) for 8-fold accelerated MRI. **Left:** Cardiac MRI [44] reconstruction example showing that the VarNet trained on the filtered dataset yields a better reconstruction than the VarNet trained on the unfiltered dataset. **Right:** While a larger fraction of fastMRI knee data [47] in the training dataset results in a major performance boost on fastMRI knee test data, additionally filtering this dataset set results in further but smaller performance gains.

- We find that training on filtered datasets improves model performance compared to training on the unfiltered dataset, on both in-distribution and out-of-distribution data, with larger improvements on in-distribution data on average. However, we find that these performance gains are on average modest compared to starting with a better designed training set, such as one that includes more data from the distribution where high performance is desired. For example, as shown in Figure 1 (right), increasing the fraction of fastMRI knee data [47] in a fixed-size training set results in a major performance boost on fastMRI knee data. Applying filtering on top of this already improved dataset yields additional, but smaller gains.
- While the quantitative improvements from data filtering are modest, we find that they correspond to a visible reduction in small reconstruction artifacts and sharper details compared to training on the unfiltered dataset.
- We study how applying data filtering impacts reconstruction performance under different compositions and sizes of the unfiltered dataset, and across acceleration factors. We find that, compared to training on unfiltered data, filtering consistently leads to better performance when the unfiltered dataset contains a low fraction of in-distribution data. In our setups, the improvement from filtering is comparable to that of a 3-fold increase of unfiltered training data.

Related work. Several works show that data curation significantly impacts the performance of vision-language models (VLMs). For example, Schuhmann et al. [35] use a trained CLIP model [33] to curate a large-scale, open-source, multimodal dataset from web-scraped data. Models trained on this dataset achieve competitive results compared to state-of-the-art proprietary models. Similarly, Gadre et al. [14] investigate various data filtering approaches and propose a dataset to further improve VLM performance along with a benchmark to facilitate research in data curation. Fang et al. [12] further investigate training a model specifically for data filtering in the VLM context.

Similarly, in natural language processing, Li et al. [19] and Penedo et al. [32] show that carefully applying heuristics and machine learning models to filter large, uncurated text corpora leads to substantial gains in LLM performance.

Research on data curation for imaging is relatively limited. For natural image restoration, Yang et al. [45] and Li et al. [20] curate datasets from web-scraped images using heuristic filtering and show that training on their dataset yields slight improvements over existing datasets.

For accelerated MRI, several works introduce raw k-space datasets to facilitate machine learning research [47, 4, 23, 43, 44, 36], but do not study filtering or curation. Zbontar et al. [47] were the first to release a large, fully-sampled k-space dataset, which advanced the field. However, many subsequent works focus on improving neural networks [39, 28, 10], as opposed to the data. Lin and Heckel [22] emphasize the need for diverse k-space datasets to enhance robustness under distribution

Table 1: Fully-sampled k-space datasets used in this work. Scans containing multiple echoes, averages, or have a time component are separated as such and counted as separate volumes. Also, 3D MRI scans are converted to three individual volumes with a new slice direction depicting axial, sagittal, or coronal views.

Dataset	Anatomy	View	Image contrast	Vendor	Magnet	Coils	Vol./Subj.	Slices
fastMRI knee [47]	knee	coronal	PD, PDFS	Siemens	1.5T,3T	15	1.2k/1.2k	42k
fastMRI brain [47]	brain	axial	T1, T1POST, T2, FLAIR	Siemens	1.5T, 3T	4-20	6.4k/6.4k	100k
CMRxRecon2023 [44]	heart	various	SSFP-Balanced	Siemens	3T	10	9.3k/300	58k
M4Raw [23]	brain	axial	T1, T2, FLAIR	XGY	0.3T	4	1.4k/183	25k
SKM-TEA [7]	knee	various	qDESS	GE	3T	8, 16	930/155	338k
AHEAD [3]	brain	various	MP2RAGE-ME	Philips	7T	32	1.1k/77	315k
fastMRI breast [36]	breast	various	VIBE	Siemens	3T	16	1.8k/300	499k
Lung 3D UTE [27]	lung	various	UTE	N/A	3T	23	69/23	18k
Chirp 3D [31]	brain	various	MPRAGE	Siemens	3T	17	6/1	1.4k
Extreme MRI [30]	lung, abdomen	various	SPGR, UTE	GE	3T	8, 12	6/2	1.7k
Fruits, Phantom [46]	N/A	various	MPRAGE	Siemens	3T	58, 64	6/2	1.9k
Heart T2-mapping [49]	heart	SAX	paper	Phillips	3T	32	44/12	1k
fastMRI prostate [43]	prostate	axial	T2	Siemens	3T	10-30	312/312	9.5k
Stanford 2D [5]	various	various	various	GE	3T	3-32	89/89	2k
NYU data [15]	knee	various	PD, PDFS, T2FS	Siemens	3T	15	100/20	3.5k
M4Raw GRE [23]	brain	axial	GRE	XGY	0.3T	4	366/183	6.6k
SMURF [2]	knee, breast, abdomen	various	FSE, FatSat, WatSat, Dixon	Siemens	3T	10-20	113/11	1.3k
OCMR [4]	heart	various	SSFP	Siemens	0.5T – 3T	15-38	4.8k/165	1.3k

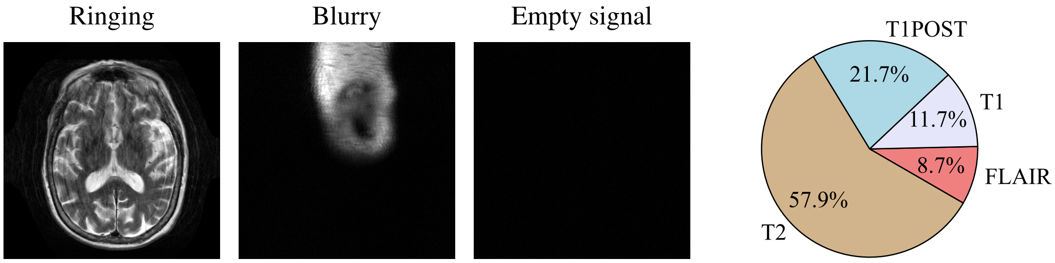

Figure 2: **Left:** Examples of low quality images within the fastMRI brain and knee test sets that we exclude from evaluation. **Right:** Skewed distribution of image contrasts within the fastMRI brain test set.

shifts but do not explore additional data curation strategies. In this work, we combine many existing open-source k-space datasets and explore data filtering for improving accelerated MRI.

2 Problem setup and data

In this section, we provide background on accelerated MRI, introduce the training data sources, the test data, as well as the models that we consider.

Accelerated MRI. We consider the problem of reconstructing a complex-valued image $\mathbf{x} \in \mathbb{C}^N$ based on undersampled measurements $\mathbf{y} \in \mathbb{C}^m$ in a multi-coil accelerated MRI setting. In this setup, C receiver coils measure electromagnetic signals, and the measurements from the i-th coil are modeled as:

$$\mathbf{y}_i = \mathbf{M}\mathbf{F}\mathbf{S}_i\mathbf{x} + \mathbf{z}_i \in \mathbb{C}^m, \quad i = 1, \dots, C, \tag{1}$$

where $\mathbf{S}_i$ is the spatial sensitivity map of the i-th coil, $\mathbf{F}$ denotes the 2D discrete Fourier transform, $\mathbf{M}$ is an undersampling mask that selects a subset of frequency components, and $\mathbf{z}_i$ is additive white Gaussian noise. The measurements $\mathbf{y}_i$ are known as k-space data. We focus on 2D MRI with Cartesian undersampling, where the central k-space region is fully sampled capturing 4%–8% of all k-space lines depending on the acceleration factor. The remaining lines are sampled equidistantly with a random offset from the start.

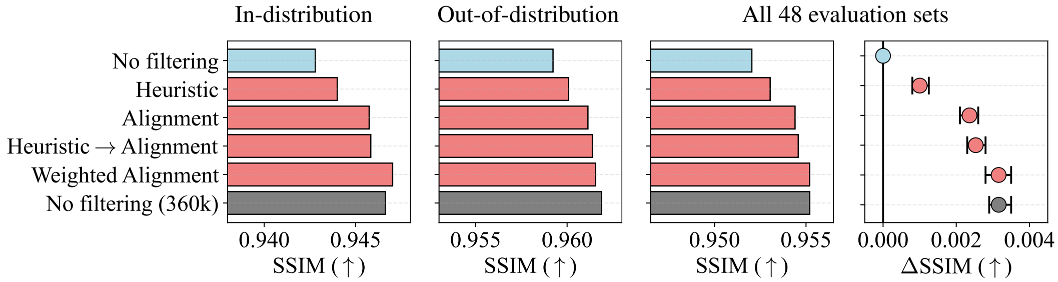

Figure 3: Data curation improves performance. For all investigated filtering methods, training on the filtered dataset improves performance over training on the unfiltered dataset (120k slices) on in-distribution and out-of-distribution evaluations. As a additional reference, the gray bar shows the performance obtained by training on a larger unfiltered dataset of 360k slices. We observe that using weighted alignment filtering matches the performance obtained by training on the larger unfiltered dataset. In the rightmost plot, we report the mean and 95% confidence intervals for performance gains over no filtering (at 120k slices), demonstrating that improvements are statistically significant.

Datasets. We utilize the data sources listed in Table 1. The first 12 sources serve as training data. Among these 12, the k-space data from fastMRI knee and brain [47], CMRxRecon2023 [44], and M4Raw [23] are acquired using 2D MRI sequences, whereas the other sources use 3D MRI sequences. Since we focus on models trained on 2D slices, we convert the 3D MRI k-space data into three separate volumes, each corresponding to axial, sagittal, or coronal views. This approach effectively increases the number of 2D slices available for training, yielding a total number of 1.1M slices for training.

Evaluation. We evaluate performance on accelerated 2D MRI. We only evaluate on data sources that are acquired with an actual 2D MRI sequence since this data enables realistic simulation of accelerated 2D MRI. Hence, in-distribution performance is evaluated on fastMRI knee, fastMRI brain, CMRxRecon2023, and M4Raw data. The last six data sources in Table 1 are used for out-of-distribution evaluation.

Many of the evaluation datasets are unbalanced in attributes such as contrast and magnetic field strength and often include lower-quality images with artifacts and noise. For example, Figure 2 illustrates that the fastMRI brain test set contains images with strong scanner artifacts and mainly T2-weighted images. To address this, we carefully curate our evaluation sets to ensure a more reliable assessment. First, we categorize the k-space data by data source, anatomy, anatomic view, contrast, number of coils, and magnetic field strength. From each group, we manually choose 5 to 24 images, ensuring diversity across subjects and slice coverage while excluding scanner artifacts. This selection process results in an evaluation suite of 48 curated test sets with 21 in-distribution test sets and 27 out-of-distribution test sets.

Models. Most of our experiments are with unrolled neural networks, specifically VarNets [39] with 80M parameters, since this type of network is the current state-of-the-art for accelerated 2D MRI reconstruction. In Appendix B, we also consider other neural networks trained end-to-end, specifically U-nets and Vision Transformers, and our conclusions for those are the same as for VarNets.

The VarNets take as input retrospectively undersampled measurements $\mathbf{y}$ and are trained with the objective to maximize SSIM between model output and the fully-sampled (i.e. $\mathbf{M}$ is identity) magnitude minimum variance unbiased estimator (MVUE) reconstructions. The sensitivity maps for computing the MVUE reconstruction are estimated with the BART toolbox [42]. The total training compute is chosen such that a model's performance saturates on a validation set that is curated in the same fashion as our evaluation suite.

Beyond networks trained end-to-end, in Section 3.6, we extend our results to diffusion-model based reconstruction methods.

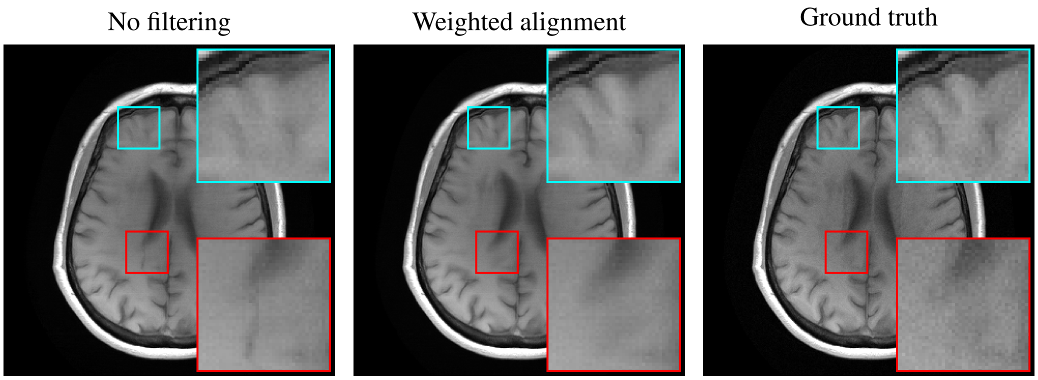

Figure 4: Reconstruction examples at 4-fold acceleration showing that the reconstruction by the model trained without filtering contains small artifacts (red), which are substantially reduced by the model trained with weighted alignment filtering, while also providing sharper details (cyan).

3 Experiments

We consider two classes of filtering approaches: heuristic filtering (Section 3.1) and alignment-based filtering (Section 3.2). In Section 3.3, we compare the performance of the considered filtering approaches. Then, in Section 3.4, we analyze how performance behaves as we vary dataset size and the difficulty of the problem by changing the acceleration factor. Section 3.5 explores the relationship between performance and train-test similarity. Finally, in Section 3.6, we demonstrate that our findings for end-to-end models generalize to diffusion model-based reconstruction methods.

3.1 Heuristic filtering

MRI scans can contain images with visual degradation such as blurriness (for example, see Figure 2). Under the hypothesis that removing such low-quality data could help the neural network learn better image priors for reconstruction, we consider removing low-quality data from the training set.

To filter a dataset, we compute a score for each image within a dataset and keep an image when the score lies above a threshold. Our heuristic filter is a composition of the two heuristic filters below:

- **Energy filtering** identifies low-energy (i.e., dark) images. For a slice, we calculate its energy-ratio score $\frac{\max(\text{slice})}{\max(\text{volume})}$, where $\max(\text{slice})$ is the slice's maximum intensity and $\max(\text{volume})$ is the maximum intensity of the entire volume. A lower energy ratio corresponds to darker images. We keep slices with a ratio above 0.11.
- **Edge-density filtering** identifies images that tend to be smooth or blurry. We first apply the Canny edge detector to compute the edges of an image. Then, the edge-density is calculated as the ratio of edge pixels to the total number of pixels in the image. We keep slices with a ratio above 0.017.

Ablation studies on the threshold choices are provided in Figure 12 in the appendix.

3.2 Alignment-based filtering

Besides heuristic filtering methods, we consider alignment-based filtering. Alignment-based filtering has been successful for vision-language data [14, 12]. Gadre et al. [14] demonstrate that filtering data by retrieving data from the data pool that are similar to the benchmark data (in their case ImageNet) and training on this retrieved data improves performance on the benchmark compared to training on the entire data pool. We explore whether similar approaches can work for accelerated MRI and introduce two variants of alignment-based filtering: a default version which we call **alignment filtering** throughout, and an alternative version called **weighted alignment filtering**.

We leverage DreamSim [13], a perceptual image similarity metric that combines embeddings from CLIP, OpenCLIP, and DINO, fine-tuned on human judgments data. This metric aligns better with

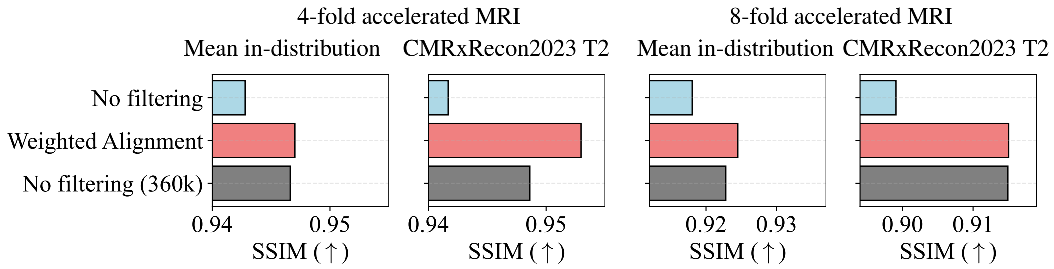

Figure 5: Compared to the mean performance gain obtained by filtering over no filtering, there exist data distributions on which filtering can significantly boost performance over no filtering (120k slices). As additional reference, the gray bar shows the performance gain obtained by training on a three times larger unfiltered dataset with 360k slices.

human image-similarity judgements than existing low-level metrics (such as PSNR, LPIPS) and semantic metrics (such as CLIP, DINO), and performs well on image retrieval tasks [41]. DreamSim computes the similarity between two images A and B as the cosine-similarity between the embeddings for images A and B.

To filter a dataset using DreamSim, we embed the magnitude of the fully-sampled MVUE reconstructions (i.e. the target images) in a dataset using the DreamSim model and do the same for each image in the validation set which is curated in the same fashion as the evaluation set. Then, for each embedding in the validation set we retrieve the images corresponding to the embeddings of the k-nearest neighbors within the dataset. Concretely, our default **alignment filtering** works as follows:

1. Preprocessing: Compute the magnitude image of the MVUE reconstruction for all slices in a dataset and in the validation set and normalize the magnitude images by the maximum magnitude pixel within their volume.
2. All magnitude images are then divided into non-overlapping image patches of size 128x128 pixels. These image-patches are embedded using the DreamSim model.
3. For each embedding, compute the cosine-similarity to the embedding belonging to the all zero image. Image patches with a similarity larger than 0.6 are discarded. This step removes image patches that are mostly empty.
4. To filter the dataset, retrieve for each embedding in the evaluation set, the images belonging to the k-nearest neighbors embeddings in the dataset.
5. Lastly, since the k-nearest neighbors of two different embeddings can contain the same image, remove all duplicates.

In our experiments, if not mentioned otherwise, we choose the number of nearest-neighbors such that 1/3 of the total number slices of the unfiltered dataset are retained. We ablate this choice on a random subset of the unfiltered data with 120k slices (see Figure 13 in the Appendix), and observed that retaining 20k to 40k slices yields similar best performance. While this choice works well in most of our experimental setups, it is not individually tuned for every setup.

Weighted alignment filtering omits Step 5 in the alignment filtering algorithm. This induces different sampling frequencies for images during training, i.e., images that are retrieved more often are also sampled more often during training. Instead of directly using the raw sampling frequencies obtained by omitting Step 5 in the alignment filtering algorithm, we take the square root of the raw sampling frequencies and use this output as sampling frequency of a slice. We observed that this approach yields slightly better performance than using the raw sampling frequencies.

Deduplication. Although our datasets do not contain slices with exact duplicates, near-duplicates occur due to very similar neighboring slices within a volume. This is often the case for the 3D MRI volumes considered here. Based on this observation, a potential caveat of alignment filtering is that the k-nearest neighbors of an embedding can contain many such near-duplicates which restricts the diversity of the retrieved dataset. To mitigate this problem, we remove near-duplicates within a volume before applying alignment filtering or weighted alignment filtering as follows: For each

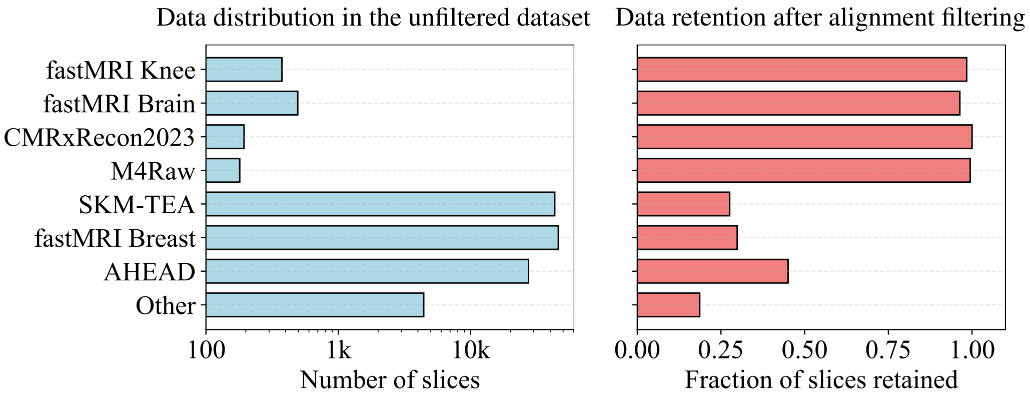

Figure 6: Number of samples for each data source in the unfiltered dataset with 120k slices in total (**left**) and fraction of samples remaining after alignment filtering 40k slices (**right**). After filtering, samples from in-distribution datasets (fastMRI knee, fastMRI brain, CMRxRecon, M4Raw) are kept almost completely.

embedding within a volume, we remove all other embeddings within the same volume if their similarity lies above a certain threshold, which we set to 0.9.

3.3 Main results

Figure 3 reports the performance of different filtering approaches. The unfiltered dataset considered for those experiments is a random subset of all volumes from the 12 training data sources totaling 120k slices. Section 3.4 reports results on the entire data pool containing 1.1M slices. Performance is reported as the average performance on all 48 test sets including in-distribution and out-of-distribution data. The main takeaways are as follows:

- All reported filtering methods improve over no filtering on both in-distribution data and out-of-distribution data.
- Alignment filtering provides better performance than heuristic filtering.
- Applying heuristic filtering first and then alignment filtering does not improve performance over only using alignment filtering.
- Weighted alignment filtering further improves performance and provides the best performance among our investigated filtering approaches.
- Overall, the mean performance gains from filtering are modest but statistically significant.

Given the modest average performance gains from filtering, a natural question is whether these improvements yield perceptible visual differences, especially when reconstructions are already mostly accurate, as in 4-fold acceleration. For example, a small numerical gain might only correspond to a slight change in brightness, which might not be perceptually significant. To investigate this, we assess how these gains appear in the test reconstructions. We find that often weighted alignment filtering reduces small reconstruction artifacts and yields sharper details compared to no filtering.

As shown in Figure 4, a model trained on the unfiltered dataset already produces an overall accurate reconstruction, but small artifacts remain. These artifacts are largely absent in the reconstructions produced by the model trained on the filtered dataset, while providing sharper details. More reconstructions are provided in the appendix in Figure 14 for 4-fold acceleration and Figure 15 for 8-fold acceleration.

Figure 5 illustrates that for certain data distributions, filtering can lead to a notable higher performance gain than the average gain. For example, on the T2-weighted cardiac images of the CMRxRecon2023 dataset [44], the performance gain (at around +0.01 SSIM) obtained by filtering is more than twice as high as the average performance gain at both 4-fold and 8-fold acceleration (a reconstruction example is shown in Figure 1). Figure 16 in the Appendix provides a detailed evaluation on all 48 test sets, where we observe that weighted alignment filtering improves on 46 out of those 48 test sets.

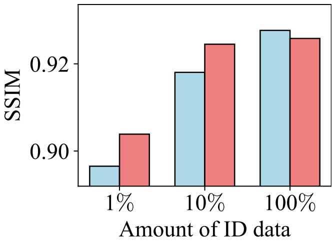

Figure 7: In-distribution (ID) performance at 8-fold acceleration as a function of the amount of in-distribution data in the unfiltered dataset. The unfiltered dataset size is fixed at 120k slices. Filtering improves performance when the fraction of in-distribution data is low. If the dataset is completely in-distribution, then filtering does not further improve performance.

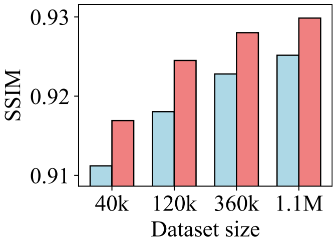

Figure 8: In-distribution performance at 8-fold acceleration as a function of the amount of total data in the unfiltered dataset. The unfiltered datasets contain 10% in-distribution data. While filtering provides consistent gains across scales, it also significantly outperforms a randomly selected subset of the same size, as seen by comparing its performance against a same-sized unfiltered dataset.

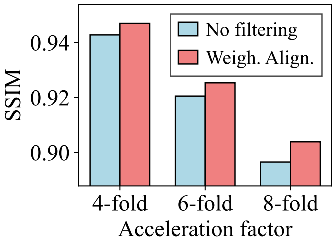

Figure 9: In-distribution performance as a function of acceleration factor. The unfiltered dataset is fixed at 120k slices containing 1% in-distribution data. Weighted alignment filtering improves performance across accelerations. Slightly larger gains are observed at higher accelerations.

Appendix B contains additional details, results for other model architectures and explores how performance changes when models are fine-tuned on the validation set used for alignment filtering. Also for those setups, we observe that weighted alignment filtering yields performance gains over no filtering.

3.4 Ablation experiments

Importance of in-distribution data in the unfiltered dataset. Figure 6 shows the data distribution across different sources before and after alignment filtering. We observe that after filtering almost all data samples from in-distribution sources, i.e., fastMRI knee, fastMRI brain, CMRxRecon2023 and M4Raw are retained. Filtering affects almost exclusively the 3D MRI data sources which are used as auxiliary training data for improving performance. This observations indicates that an effective filter identifies in-distribution data as much as possible and mainly removes data from auxiliary data sources.

Based on this observation, we now examine how the initial composition of the unfiltered dataset influences the effectiveness of alignment filtering. Figure 7 compares at 8-fold acceleration in-distribution performance between weighted alignment filtering and no filtering as a function of the fraction of in-distribution data in an unfiltered dataset of fixed size (120k slices). We observe that filtering improves performance when the fraction of in-distribution data is low, and in the case where no auxiliary data is used for training, filtering can hurt performance. This suggests that filtering is beneficial when in-distribution data is scarce.

Dataset size. Next, we study how filtering performance is related to the size of the unfiltered dataset. Figure 8 compares at 8-fold acceleration in-distribution performance between weighted alignment filtering and no filtering as a function of the unfiltered dataset size containing 10% in-distribution data. We observe that weighted alignment filtering yields similar performance improvements across different data scales. On the investigated data scale, the performance gains obtained by filtering are comparable to the gains obtained by a 3-fold increase of the unfiltered dataset.

Acceleration factor. Lastly, we investigate how filtering performance is affected by the acceleration factor, which changes the reconstruction difficulty. Figure 9 shows performance of weighted alignment filtering as a function of the acceleration factor. The unfiltered dataset size is fixed to 120k slices with 1% in-distribution data. We observe that filtering improves performance across acceleration factors with a slight tendency of larger improvements at higher accelerations, where there is more room for improvement.

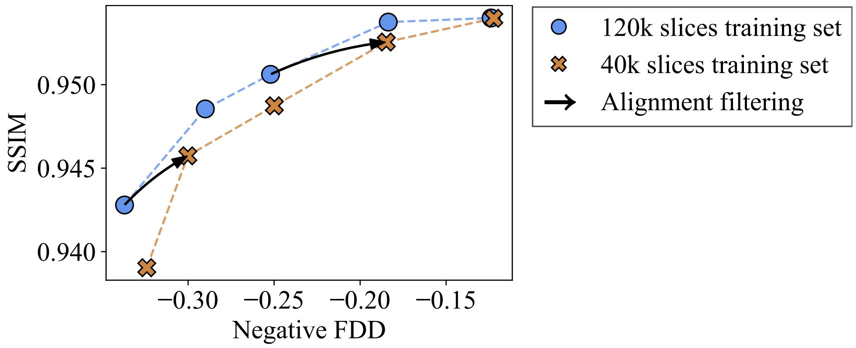

Figure 10: For a fixed training set size, similarity between training and test distribution measured as the negative Fréchet DreamSim Distance (FDD) correlates with performance on the test set. Alignment filtering reduces the dataset size but improves this similarity which relates to improved performance.

Similar qualitative results are obtained when investigating scaling for 4-fold accelerated MRI and scaling with model size. Results are provided in Appendix C.

3.5 Relation between performance and train-test set similarity

We hypothesize that alignment filtering, which selects training samples using a validation set that resembles the test set, improves performance by increasing the similarity between the resulting training and the test distribution.

To investigate this hypothesis, we quantify train-test set similarity with what we call the **Fréchet DreamSim Distance**, which is similar to the Fréchet Inception Distance (FID) [16]. Instead of using Inception-v3 embeddings, Fréchet DreamSim Distance uses DreamSim embeddings, following Stein et al. [40], who show that relying on DreamSim embeddings when computing the Fréchet distance between two datasets captures distributional similarity better than relying Inception-v3 embeddings.

Figure 10 shows this metric between training sets and the in-distribution validation sets, and we see a high correlation with reconstruction performance at 4-fold acceleration for fixed training set sizes. Alignment filtering reduces the training set size but increases the train-test similarity which relates to performance gains.

However, while we find that this similarity metric correlates well for in-distribution evaluations, we only observed weak correlation when considering out-of-distribution setups. For example, we found that taking a training set that is completely out-of-distribution relative to the test sets and substituting 1% of that training set with in-distribution data can significantly enhance performance on the in-distribution test sets. Yet, the Fréchet DreamSim Distance remains largely unchanged as only a tiny fraction of the dataset has changed. For out-of-distribution evaluation other similarity metrics, such as those relying on nearest neighbors between training and test sets [26, 22], can provider better correlation with performance.

3.6 Results for reconstruction with diffusion models

In the previous sections, we studied data filtering for models trained end-to-end. In this section, we explore whether the same filtering techniques can also benefit diffusion model-based reconstruction approaches for accelerated MRI. We compare diffusion models trained on an unfiltered dataset with those trained on the weighted alignment filtered dataset. The diffusion models are trained on MVUE reconstructions of fully sampled k-space data. For reconstruction, we consider variational optimization [25] and decomposed diffusion sampling [6] (more details are in Appendix D).

Figure 11 shows for 4-fold accelerated MRI that filtering with weighted alignment also improves the performance of diffusion models. This improvement is consistent across both sampling techniques, different sizes of unfiltered data and varying proportions of in-distribution data.

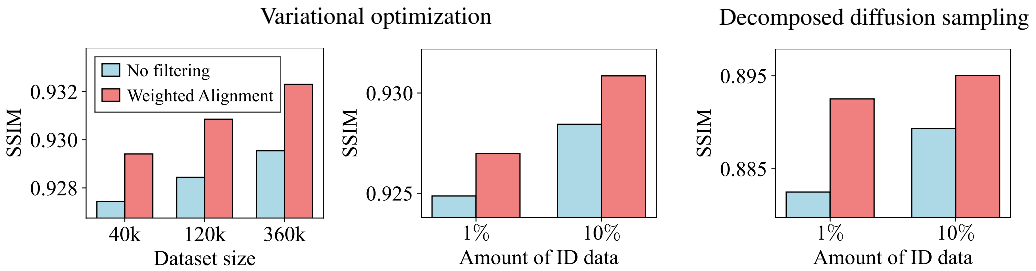

Figure 11: Performance comparison under 4-fold acceleration on in-distribution data of a diffusion model trained on a weighted alignment filtered dataset and a model trained on an unfiltered dataset. We consider variational optimization [25] and decomposed diffusion sampling [6] for reconstruction. We study performance under different sizes of the unfiltered dataset, and varying fraction of in-distribution (ID) data in a fixed size unfiltered dataset of 120k slices. Filtering improves performance of diffusion models with either reconstruction method and across dataset compositions.

4 Conclusion and limitations

This work proposes and investigates various filtering strategies for accelerated MRI and demonstrates that data filtering can advance the performance of existing state-of-the-art neural networks.

Our main finding is that data curation through filtering for accelerated 2D MRI consistently improves performance for end-to-end models as well as for diffusion models, which are currently the two most performant and widely used model classes. However, the improvements are relatively modest compared to the improvements data filtering achieves in other domains, e.g., for language models and for vision-language models. The reason could be that the quality of the images in the medical datasets we considered are already of relatively high quality.

In this work we focused on accelerated 2D MRI, while other important related reconstruction problems such as accelerated 3D MRI, motion compensated MRI reconstruction, and image reconstruction problems beyond MRI are not considered.

While refining data curation processes have become critical research areas in machine learning for computer vision and natural language processing, in imaging they received little attention. Our work is an early step towards understanding effective data filtering for imaging, in particular for accelerated MRI.

Acknowledgments and Disclosure of Funding

The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany in the programme of "Souverän. Digital. Vernetzt.". Joint project 6G-life, project identification number: 16KISK002, as well as from Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 456465471, 464123524, and 517586365.

References

[1] P. M. Adamson, A. D. Desai, J. Dominic, C. Bluethgen, J. P. Wood, A. B. Syed, R. D. Boutin, K. J. Stevens, S. Vasanawala, J. M. Pauly, A. S. Chaudhari, and B. Gunel. Using Deep Feature Distances for Evaluating The Perceptual Quality of MR Image Reconstructions. *Magnetic Resonance in Medicine*, 2025.

[2] B. Bachrata, B. Strasser, W. Bogner, A. I. Schmid, R. Korinek, M. Krššák, S. Trattnig, and S. D. Robinson. Simultaneous Multiple Resonance Frequency imaging (SMURF): Fat-water imaging using multi-band principles. *Magnetic Resonance in Medicine*, 2021.

[3] M. Caan. Quantitative motion-corrected 7T sub-millimeter raw MRI database of the adult lifespan. https://doi.org/10.34894/IHZGQM, 2022.

[4] C. Chen, Y. Liu, P. Schniter, M. Tong, K. Zareba, O. Simonetti, L. Potter, and R. Ahmad. OCMR (v1.0)–Open-Access Multi-Coil k-Space Dataset for Cardiovascular Magnetic Resonance Imaging. *arXiv:2008.03410*, 2020.

[5] J. Y. Cheng. Stanford 2D FSE. http://mridata.org/list?project=Stanford 2D FSE, 2018.

[6] H. Chung, S. Lee, and J. C. Ye. Decomposed Diffusion Sampler for Accelerating Large-Scale Inverse Problems. In *International Conference on Learning Representations*, 2023.

[7] A. D. Desai, A. M. Schmidt, E. B. Rubin, C. M. Sandino, M. S. Black, V. Mazzoli, K. J. Stevens, R. Boutin, C. Re, G. E. Gold, B. Hargreaves, and A. Chaudhari. SKM-TEA: A Dataset for Accelerated MRI Reconstruction with Dense Image Labels for Quantitative Clinical Evaluation. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[8] P. Dhariwal and A. Q. Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, 2021.

[9] K. Epperson. Creation of fully sampled mr data repository for compressed sensing of the knee. In *SMRT Annual Meeting*, 2013.

[10] Z. Fabian, B. Tinaz, and M. Soltanolkotabi. HUMUS-Net: Hybrid Unrolled Multi-scale Network Architecture for Accelerated MRI Reconstruction. In *Advances in Neural Information Processing Systems*, 2022.

[11] A. Fang, G. Ilharco, M. Wortsman, Y. Wan, V. Shankar, A. Dave, and L. Schmidt. Data Determines Distributional Robustness in Contrastive Language Image Pre-training (CLIP). In *International Conference on Machine Learning*, 2022.

[12] A. Fang, A. M. Jose, A. Jain, L. Schmidt, A. T. Toshev, and V. Shankar. Data Filtering Networks. In *International Conference on Learning Representations*, 2023.

[13] S. Fu, N. Y. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel, and P. Isola. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. In *Advances in Neural Information Processing Systems*, 2023.

[14] S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang, E. Orgad, R. Entezari, G. Daras, S. M. Pratt, V. Ramanujan, Y. Bitton, K. Marathe, S. Mussmann, R. Vencu, M. Cherti, R. Krishna, P. W. Koh, O. Saukh, A. Ratner, S. Song, H. Hajishirzi, A. Farhadi, R. Beaumont, S. Oh, A. Dimakis, J. Jitsev, Y. Carmon, V. Shankar, and L. Schmidt. DataComp: In search of the next generation of multimodal datasets. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[15] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll. Learning a Variational Network for Reconstruction of Accelerated MRI Data. *Magnetic Resonance in Medicine*, 2018.

[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

[17] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[18] A. Krainovic, S. Ruschke, and R. Heckel. Resolution-Robust 3D MRI Reconstruction with 2D Diffusion Priors: Diverse-Resolution Training Outperforms Interpolation. *arXiv:2412.18584*, 2024.

[19] J. Li, A. Fang, G. Smyrnis, M. Ivgi, M. Jordan, S. Y. Gadre, H. Bansal, E. Guha, S. S. Keh, K. Arora, S. Garg, R. Xin, N. Muennighoff, R. Heckel, J. Mercat, M. Chen, S. Gururangan, M. Wortsman, A. Albalak, Y. Bitton, M. Nezhurina, A. Abbas, C.-Y. Hsieh, D. Ghosh, J. Gardner, M. Kilian, H. Zhang, R. Shao, S. Pratt, S. Sanyal, G. Ilharco, G. Daras, K. Marathe, A. Gokaslan, J. Zhang, K. Chandu, T. Nguyen, I. Vasiljevic, S. Kakade, S. Song, S. Sanghavi, F. Faghri, S. Oh, L. Zettlemoyer, K. Lo, A. El-Nouby, H. Pouransari, A. Toshev, S. Wang, D. Groeneveld,

L. Soldaini, P. W. W. Koh, J. Jitsev, T. Kollar, A. Dimakis, Y. Carmon, A. Dave, L. Schmidt, and V. Shankar. DataComp-LM: In search of the next generation of training sets for language models. In *Advances in Neural Information Processing Systems*, 2024.

[20] Y. Li, K. Zhang, J. Liang, J. Cao, C. Liu, R. Gong, Y. Zhang, H. Tang, Y. Liu, D. Demandolx, R. Ranjan, R. Timofte, and L. Van Gool. LSDIR: A Large Scale Dataset for Image Restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

[21] K. Lin and R. Heckel. Vision Transformers Enable Fast and Robust Accelerated MRI. In *International Conference on Medical Imaging with Deep Learning*, 2022.

[22] K. Lin and R. Heckel. Robustness of Deep Learning for Accelerated MRI: Benefits of Diverse Training Data. In *International Conference on Machine Learning*, 2024.

[23] M. Lyu, L. Mei, S. Huang, S. Liu, Y. Li, K. Yang, Y. Liu, Y. Dong, L. Dong, and E. X. Wu. M4Raw: A multi-contrast, multi-repetition, multi-channel MRI k-space dataset for low-field MRI research. *Scientific Data*, 2023.

[24] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008.

[25] M. Mardani, J. Song, J. Kautz, and A. Vahdat. A Variational Perspective on Solving Inverse Problems with Diffusion Models. In *International Conference on Learning Representations*, 2023.

[26] P. Mayilvahanan, T. Wiedemer, E. Rusak, M. Bethge, and W. Brendel. Does CLIP's generalization performance mainly stem from high train-test similarity? In *International Conference on Learning Representations*, 2024.

[27] L. Mendes Pereira. Data for "Pulmonary ventilation and extra-vascular lung water quantification with free-running 3D-UTE MRI at 3T". https://doi.org/10.7910/DVN/LZSR8O, 2020.

[28] M. J. Muckley, B. Riemenschneider, A. Radmanesh, S. Kim, G. Jeong, J. Ko, Y. Jun, H. Shin, D. Hwang, M. Mostapha, S. Arberet, D. Nickel, Z. Ramzi, P. Ciuciu, J.-L. Starck, J. Teuwen, D. Karkalousos, C. Zhang, A. Sriram, Z. Huang, N. Yakubova, Y. W. Lui, and F. Knoll. Results of the 2020 fastMRI Challenge for Machine Learning MR Image Reconstruction. *IEEE Transactions on Medical Imaging*, 2021.

[29] T. Nguyen, G. Ilharco, M. Wortsman, S. Oh, and L. Schmidt. Quality Not Quantity: On the Interaction between Dataset Design and Robustness of CLIP. In *Advances in Neural Information Processing Systems*, 2022.

[30] F. Ong, X. Zhu, J. Y. Cheng, K. M. Johnson, P. E. Z. Larson, S. S. Vasanawala, and M. Lustig. Extreme MRI: Large-scale volumetric dynamic imaging from continuous non-gated acquisitions. *Magnetic Resonance in Medicine*, 2020.

[31] K. Pawar, Z. Chen, J. Zhang, N. J. Shah, and G. F. Egan. Application of compressed sensing using chirp encoded 3D GRE and MPRAGE sequences. *International Journal of Imaging Systems and Technology*, 2020.

[32] G. Penedo, H. Kydlíček, L. B. Allal, A. Lozhkov, M. Mitchell, C. Raffel, L. Von Werra, and T. Wolf. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. In *Advances in Neural Information Processing Systems*, 2024.

[33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, 2021.

[34] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2015.

[35] C. Schuhmann, R. Beaumont, R. Vencu, C. W. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. R. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, 2022.

[36] E. Solomon, P. M. Johnson, Z. Tan, R. Tibrewala, Y. W. Lui, F. Knoll, L. Moy, S. G. Kim, and L. Heacock. FastMRI Breast: A Publicly Available Radial k-Space Dataset of Breast Dynamic Contrast-enhanced MRI. *Radiology: Artificial Intelligence*, 2025.

[37] J. Song, C. Meng, and S. Ermon. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*, 2020.

[38] R. Souza, O. Lucena, J. Garrafa, D. Gobbi, M. Saluzzi, S. Appenzeller, L. Rittner, R. Frayne, and R. Lotufo. An open, multi-vendor, multi-field-strength brain MR dataset and analysis of publicly available skull stripping methods agreement. *NeuroImage*, 2018.

[39] A. Sriram, J. Zbontar, T. Murrell, A. Defazio, C. L. Zitnick, N. Yakubova, F. Knoll, and P. Johnson. End-to-End Variational Networks for Accelerated MRI Reconstruction. In *Medical Image Computing and Computer Assisted Intervention*, 2020.

[40] G. Stein, J. C. Cresswell, R. Hosseinzadeh, Y. Sui, B. L. Ross, V. Villecroze, Z. Liu, A. L. Caterini, J. E. T. Taylor, and G. Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, 2024.

[41] S. Sundaram, S. Fu, L. Muttenthaler, N. Y. Tamir, L. Chai, S. Kornblith, T. Darrell, and P. Isola. When does perceptual alignment benefit vision representations? In *Advances in Neural Information Processing Systems*, 2024.

[42] Tamir, Jon, Uecker, Martin, and Lustig, Michael. BART: Computational Magnetic Resonance Imaging. https://mrirecon.github.io/bart/.

[43] R. Tibrewala, T. Dutt, A. Tong, L. Ginocchio, R. Lattanzi, M. B. Keerthivasan, S. H. Baete, S. Chopra, Y. W. Lui, D. K. Sodickson, H. Chandarana, and P. M. Johnson. FastMRI Prostate: A public, biparametric MRI dataset to advance machine learning for prostate cancer imaging. *Scientific Data*, 2024.

[44] C. Wang, J. Lyu, S. Wang, C. Qin, K. Guo, X. Zhang, X. Yu, Y. Li, F. Wang, J. Jin, Z. Shi, Z. Xu, Y. Tian, S. Hua, Z. Chen, M. Liu, M. Sun, X. Kuang, K. Wang, H. Wang, H. Li, Y. Chu, G. Yang, W. Bai, X. Zhuang, H. Wang, J. Qin, and X. Qu. CMRxRecon: A publicly available k-space dataset and benchmark to advance deep learning for cardiac MRI. *Scientific Data*, 2024.

[45] Q. Yang, D. Chen, Z. Tan, Q. Liu, Q. Chu, J. Bao, L. Yuan, G. Hua, and N. Yu. HQ-50K: A Large-scale, High-quality Dataset for Image Restoration. *arXiv:2306.05390*, 2023.

[46] T. Yu, T. Hilbert, G. F. Piredda, A. Joseph, G. Bonanno, S. Zenkhri, P. Omoumi, M. B. Cuadra, E. Canales Rodriguez, T. Kober, and J.-P. Thiran. Validation and Generalizability of Self-Supervised Image Reconstruction Methods for Undersampled MRI. *Machine Learning for Biomedical Imaging*, 2022.

[47] J. Zbontar, F. Knoll, A. Sriram, T. Murrell, Z. Huang, M. J. Muckley, A. Defazio, R. Stern, P. Johnson, M. Bruno, M. Parente, K. J. Geras, J. Katsnelson, H. Chandarana, Z. Zhang, M. Drozdzal, A. Romero, M. Rabbat, P. Vincent, N. Yakubova, J. Pinkerton, D. Wang, E. Owens, C. L. Zitnick, M. P. Recht, D. K. Sodickson, and Y. W. Lui. fastMRI: An Open Dataset and Benchmarks for Accelerated MRI. *arXiv:1811.08839*, 2019.

[48] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[49] D. Zhu, H. Ding, M. M. Zviman, H. Halperin, M. Schär, and D. A. Herzka. Accelerating whole-heart 3D T2 mapping: Impact of undersampling strategies and reconstruction techniques. *PLOS ONE*, 2021.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Claims are not overstated and consistently supported by the content throughout the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 4.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This is a purely empirical paper without theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Implementation details are provided in the main body and appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All datasets used in this work are publicly available. We provide the code used in this work.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Details are provided in the main body and appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See Figure 3 in Section 3.3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Information is provided in Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research complies fully with the NeurIPS Code of Ethics with no foreseeable risks of harm, bias, or misuse.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We believe no such risks are posed by this paper.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We reference all assets used in this work (e.g. the datasets in Table 1).

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide code for reproducing the results.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This work only uses imaging data that is already publicly available prior to this work.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work only uses imaging data that is already publicly available prior to this work.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This research does not involve LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Table 2: Performance measured in PSNR [dB] (↑) and LPIPS (↓) of different filtering methods at 4-fold acceleration, and 120k slices in the unfiltered dataset with 1% in-distribution data.

Filtering strategy	Dataset	fastMRI	fastMRI	In-	Out-of-	Mean over
	size	knee	brain	distribution	distribution	48 datasets
No filtering	120k	40.11	40.54	39.25	40.57	39.99
		0.155	0.103	0.115	0.108	0.111
Heuristic	80k	40.25	40.75	39.50	40.74	40.19
		0.152	0.101	0.112	0.106	0.109
Alignment	40k	40.38	40.96	39.74	40.88	40.38
		0.150	**0.100**	**0.111**	**0.104**	**0.107**
Heuristic→Alignment	40k	40.40	41.00	39.77	40.94	40.42
		0.152	**0.100**	0.112	**0.104**	**0.107**
Weighted Alignment	40k	**40.60**	**41.31**	**40.05**	**41.03**	**40.60**
		0.151	0.103	0.113	**0.104**	0.108

A Details on the experimental setup

Code. The code for this work can be found here: https://github.com/MLI-lab/data_filtering_for_accelerated_mri. The repository also contains the raw evaluation output data analyzed in this work.

Access to datasets. Due to licensing restrictions from the original dataset sources, we unfortunately cannot host the curated datasets ourselves. However, all datasets used in this work are publicly available from their respective source (see Table 1). We provide code to convert these datasets into a unified format used throughout this work.

Data conversion. Different sources store k-space data in different formats. We organize and save the data with the fastMRI convention, where each k-space volume has shape [number of slices, number of coils, ky, kx] and is stored in a HDF5 file. We split scans that originally included more dimensions, for example, due to multiple echoes (e.g. SKM-TEA [7]) or temporal frames as in cine MRI [44], along those dimensions and treat them as separate volumes. For 3D MRI scans, the k-space data is converted into three distinct volumes, each corresponding to a coronal, axial, or sagittal view. Storing the data from all sources in Table 1 after conversion requires 20TB of disk space.

Models. We rely on the end-to-end VarNet [39] implementation provided by the fastMRI repository. We consider VarNets with 80M parameters that have eight cascades where each reconstruction U-net has 36 channels in the first pooling layer and 4 pooling layers. The original VarNet implementation maps the predicted k-space to a root-sum-of-square reconstruction. However, since we evaluate on MVUE ground-truths, we perform a MVUE reconstruction with the predicted k-space.

Training. We train the VarNets until saturating performance is reached on the validation set. We use the Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$ and a batch size of two. The learning rate is warmed up linearly to 4e-4 using 1% of total training time and then linearly decayed to 1.6e-5. Training a model on an unfiltered dataset of 120k slices using a single NVIDIA L40 GPU and four workers takes around 90 hours and 43GiB in GPU memory. Using the same setup, training a model on an unfiltered dataset of 40k slices takes around 36 hours, on an unfiltered dataset of 360k slices around 170 hours, and on the entire data pool totaling 1.1M slices around 500 hours.

Evaluation. To compute the mean performance score, we compute the average reconstruction performance for each data distribution and then average these scores over all considered data distributions. Moreover, we use the sensitivity maps to compute a mask that better captures the region of interest. This mask is then applied to both the model output and the ground truth to exclude the background before computing a performance metric. This approach reduces variations in the metric caused by reconstruction errors in the background, which are not relevant for evaluation. Moreover, following Lin and Heckel [22], we normalize the reconstructions to have the same mean and variance as the reference image. This reduces metric fluctuations caused by minor, imperceptible differences in brightness and contrast that could otherwise disproportionately impact scores.

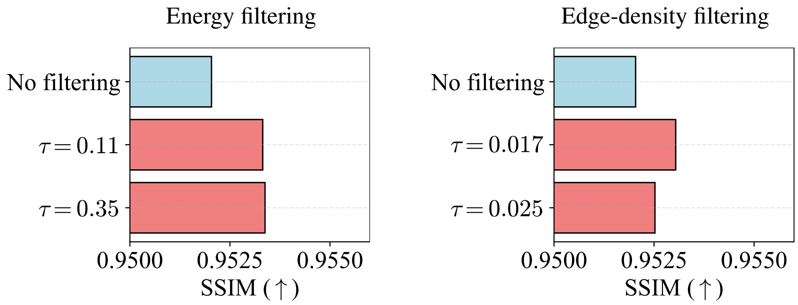

Figure 12: Ablation study for the energy (left) and edge-density (right) thresholds of our heuristic filtering methods. The results show that both heuristic filters provide improvement in SSIM over no filtering. Our chosen thresholds ($\tau = 0.11$ for energy and $\tau = 0.017$ for edge-density) yield the optimal performance.

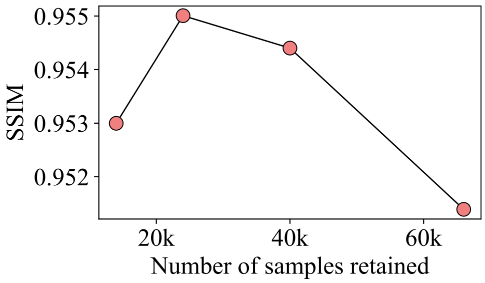

Figure 13: We ablate the choice of the number of nearest neighbors for alignment filtering on a random subset of 120k slices for 4-fold acceleration. Choosing the number of nearest neighbors such that between 20k and 40k samples are retained yields best performance.

Edge-density filtering. We use scikit-image's implementation of the Canny edge detector with the following configuration: `skimage.feature.canny(image, sigma=2, low_threshold=0.01, high_threshold=0.2)`.

Threshold of heuristic filtering. We perform an ablation study on the thresholds for our heuristic filtering methods, as shown in Figure 12. For the energy threshold, we compare the threshold $\tau = 0.11$ against a higher threshold of $\tau = 0.35$ and the no filtering baseline. Both thresholds achieve a similar SSIM and better than baseline; we selected $\tau = 0.11$ because $\tau = 0.35$ resulted in the removal of many slices that contained clear signals, which defeats the purpose of the energy filter. For the edge-density threshold, we test our choice of $\tau = 0.017$ against $\tau = 0.025$ and the no filtering baseline. The chosen value of $\tau = 0.017$ outperforms both the no filtering and the alternative threshold of $\tau = 0.025$. However, these heuristic filtering methods are less effective than the DreamSim-based alignment filtering presented in Section 3.3.

Number of nearest neighbors for alignment filtering. In the main body, we choose the number of nearest neighbor for alignment filtering such that 33% of the data is retained. We ablate this choice for 4-fold acceleration on the unfiltered dataset with 120k slices. Figure 13 shows that choosing the number of nearest neighbors such that between 20k and 40k (33%) samples are retained yields best performance. Based on this observation, we always retain 33% of the data when applying alignment filtering when the unfiltered dataset contains at least 120k slices. For unfiltered datasets with 40k slices the number of nearest neighbors is chosen such that 20k (50%) samples are retained as this choice yielded better results than retaining 33% of the data.

B Additional details and results for Section 3.3

Confidence intervals. We use bootstrapping to compute the confidence intervals in Figure 3. We first compute and store the SSIM difference obtained by the model trained on a filtered over the

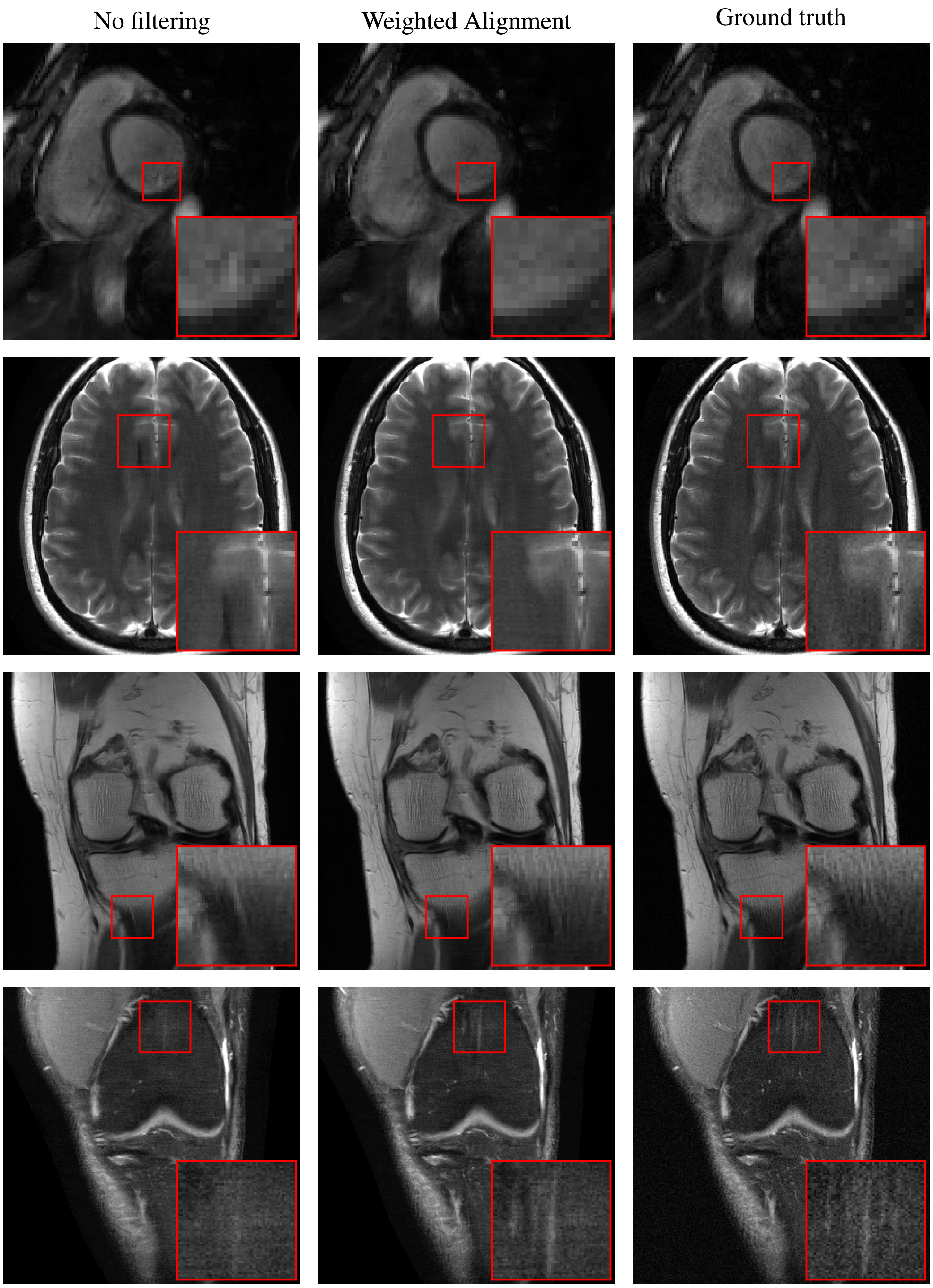

Figure 14: Reconstruction examples at 4-fold acceleration showing reduced artifacts and sharper details in the reconstructions obtained with weighted alignment filtering compared to those obtained with no filtering.

model trained on the unfiltered dataset for each individual reconstruction. From this set of SSIM differences, we sample scores with replacement until we obtain the size of the original test set. Then, we compute the mean SSIM difference by computing first the average SSIM difference for each data distribution and then average these over all considered data distributions. This process is repeated 10000 times which yields a distribution of mean SSIM differences. Finally, from this distribution, we take the 2.5 percentile as lower bound and the 97.5 percentile as upper bound for reporting the 95% confidence interval.

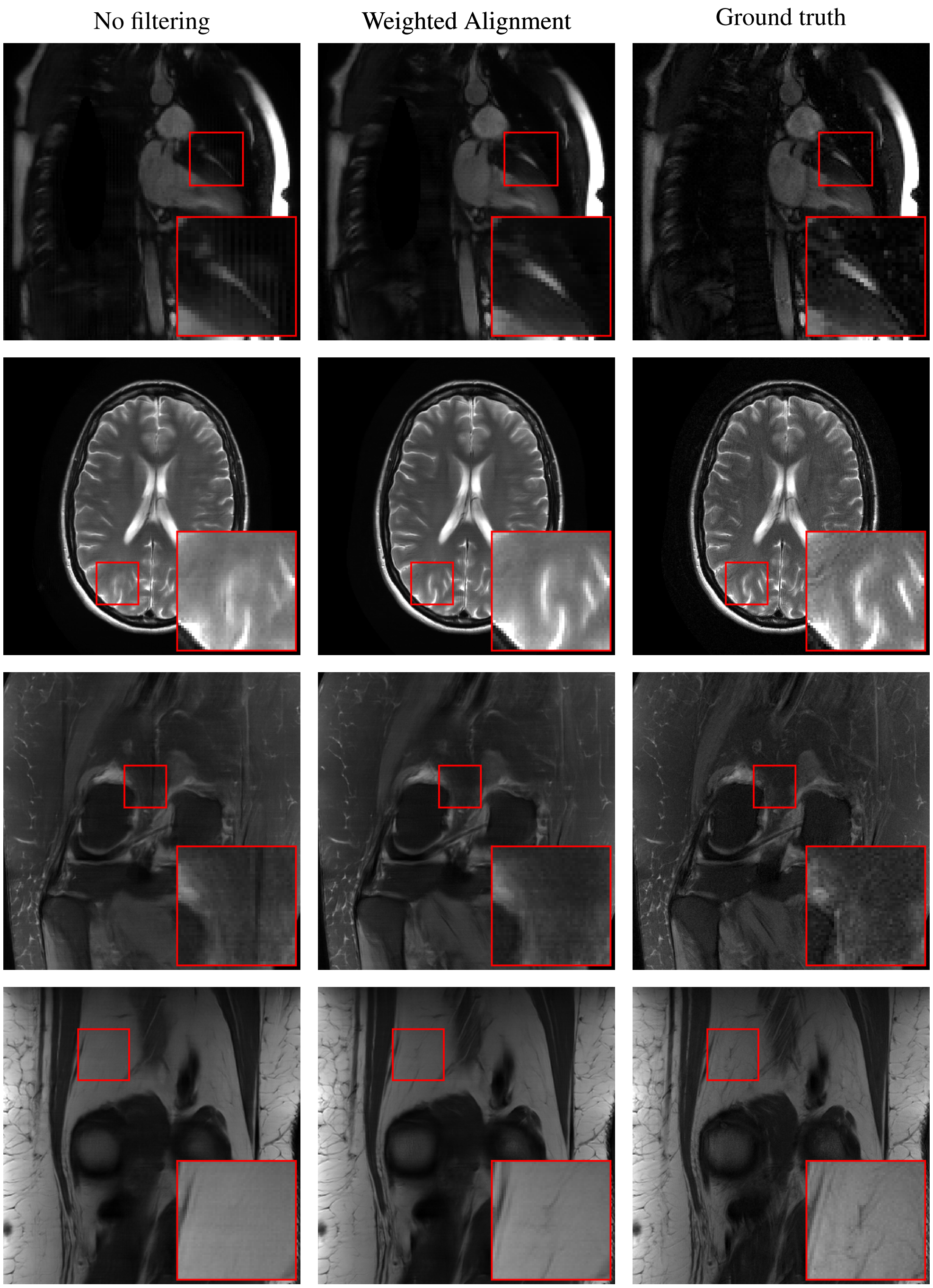

Figure 15: Reconstruction examples at 8-fold acceleration showing reduced artifacts and sharper details in the reconstructions obtained with weighted alignment filtering compared to those obtained with no filtering.

Additional metrics. In Section 3.3, we use SSIM as performance metric for comparing different filtering methods. Table 2 provides additional performance metrics: PSNR and LPIPS [48]. LPIPS is a metric based on features of a pretrained neural network. We include LPIPS because studies have shown that LPIPS correlates well with radiologist readings [1]. Interestingly, we observe a trade-off between weighted alignment filtering and alignment filtering: Weighted alignment filtering obtains higher PSNR but lower LPIPS compared to alignment filtering.

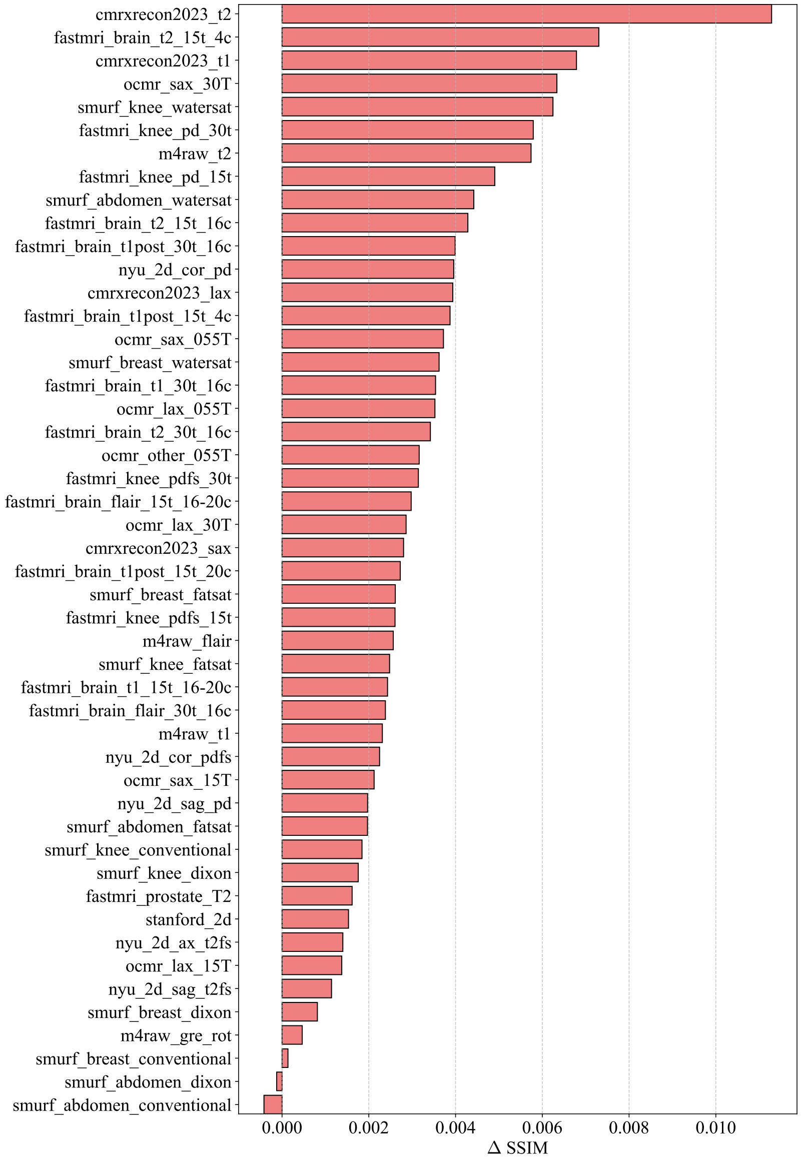

Figure 16: Weighted alignment filtering improves on 46 out of 48 sets for 4-fold accelerated MRI and a dataset size of 120k slices.

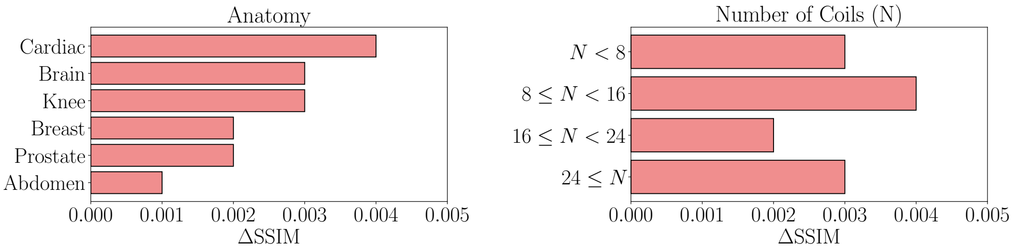

Figure 17: Breakdown of the average SSIM improvement from weighted alignment filtering across subgroups for anatomy, and number of coils. The experimental setup is the same as Figure 16: Weighted alignment filtering on a 120k-slice unfiltered dataset at 4-fold acceleration.

Additional reconstructions. Figure 14 (4-fold acceleration) and Figure 15 (8-fold acceleration) provide additional reconstruction examples for the models reported in Section 3.3, further demon-

Table 3: Filtering results for U-net and ViT trained for 4-fold acceleration, and 120k slices in the unfiltered dataset. The unfiltered dataset for training U-net contains 1% in-distribution data and for ViT 10%. Performance is measured in SSIM(↑), PSNR [dB](↑) and LPIPS(↓).

Model	Filtering strategy	Dataset size	fastMRI knee	fastMRI brain	In-distribution	Out-of-distribution	Mean over 48 datasets
U-net	No filtering	120k	0.905	0.935	0.909	**0.911**	0.910
120M			36.81	36.33	35.29	35.42	35.36
param.			0.235	0.170	0.185	0.186	0.186
	Weighted Alignment	40k	**0.911**	**0.941**	**0.916**	**0.911**	**0.913**
			37.58	**37.28**	**36.05**	**35.50**	**35.74**
			0.220	**0.162**	**0.177**	**0.184**	**0.181**
ViT	No filtering	120k	0.916	0.948	0.923	**0.918**	0.920
60M			37.81	37.52	36.36	35.88	36.09
param.			0.212	0.158	0.170	0.178	0.174
	Weighted Alignment	40k	**0.922**	**0.954**	**0.929**	**0.918**	**0.923**
			38.41	**38.54**	**37.15**	**35.97**	**36.49**
			0.202	**0.148**	**0.160**	**0.177**	**0.169**

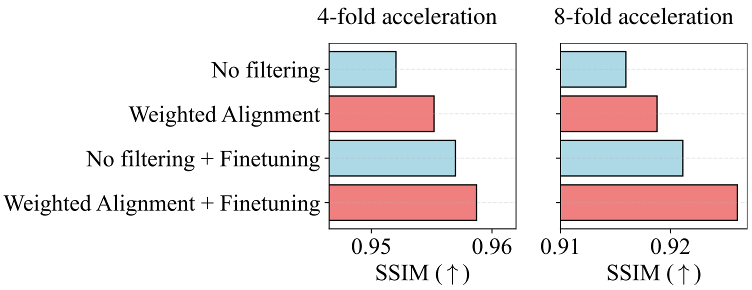

Figure 18: Weighted alignment filtering outperforms no filtering also after the models pretrained on either datasets are fine-tuned on the validation set (269 slices) used for alignment filtering. The unfiltered dataset size is 120k slices.

strating that models trained on the weighted alignment filtered datasets reduce artifacts and produce slightly sharper details compared to those trained on the unfiltered dataset.

Evaluation on each test set. In the main body, we report aggregated performance scores. Figure 16 provides for 4-fold accelerated MRI a detailed performance comparison between weighted alignment filtering and no filtering on all 48 test sets. Reported is the difference in SSIM between weighted alignment filtering and no filtering. Filtering yields improvements on 46 out of 48 sets. Dataset names follow the format: <dataset source >_<contrast >_<magnet strength >_<number of coils >.

Evaluation on different subgroups. We provide a detailed breakdown of the average SSIM improvement across different subgroups for anatomy, field strength, and number of coils. The results are shown in Figure 17, and it can be seen that all subgroups show improvements. We observe that the most gains were achieved in cardiac scans when evaluating different anatomies. For coil numbers, the highest gains are obtained between 8 and 16 coils, but a global trend cannot be concluded across the entire spectrum of coil numbers considered.

Other model architectures. Beside VarNet [39], which is an unrolled network relying on data consistency, we investigate a standard U-net trained for accelerated MRI [47] and a Vision Transformer (ViT) adjusted for accelerated MRI reconstruction [21]. Those two models do not rely on data consistency. While the overall performance is lower than that of VarNet, Table 3 shows that weighted alignment filtering improves performance over no filtering also for those models with gains up to 1dB in PSNR.

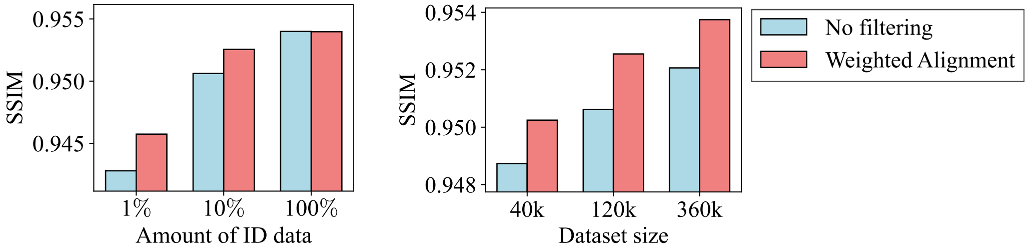

Figure 19: Scaling results (in-distribution performance) at 4-fold acceleration. **Left:** Performance as a function of the amount of in-distribution data in the unfiltered dataset. The unfiltered dataset size is fixed at 120k slices. Filtering improves performance when little in-distribution data is available. **Right:** Performance as a function of the amount total data in the unfiltered dataset. The unfiltered datasets contain 10% in-distribution data. Performance improvements are consistent over different data scales.

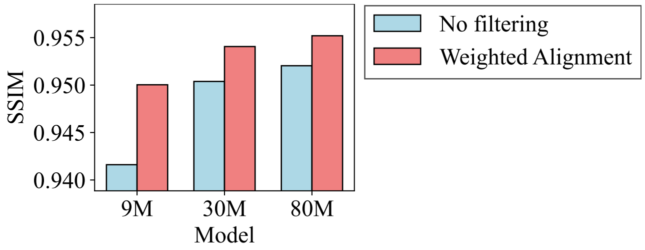

Figure 20: Filtering improves performance across models sizes. The unfiltered dataset contains 120k slices with 1% in-distribution data. The 30M parameter model is fastMRI's default VarNet configuration [39]. Filtering is particularly beneficial for the small VarNet with 9M parameters.

Fine-tuning. Alignment filtering relies on a validation set to retrieve images from a data pool that are similar to the evaluation data. We investigate how further fine-tuning on our validation set (269 slices) changes the performance difference between pretraining on the unfiltered dataset and pretraining on the weighted alignment filtered dataset. For each pretrained model, we perform grid search across number of fine-tuning epochs and learning rates and report the best performance obtained on our evaluation set. Figure 18 shows for 4-fold and 8-fold acceleration that weighted alignment filtering outperforms no filtering also after the models pretrained on either datasets are fine-tuned on the validation set.

C Additional details and results for Section 3.4

Scaling experiments. In Section 3.4, we report results on scaling-experiments for 8-fold acceleration where we investigate in-distribution performance as a function of in-distribution data proportion (Figure 7) and as a function of dataset size (Figure 8). Figure 19 reports the result on the same scaling-experiments but for 4-fold acceleration and using alignment filtering. Also here, filtering improves performance across dataset sizes and for low in-distribution data proportions.

Model size. Figure 20 shows that weighted alignment filtering improves performance across models sizes. The unfiltered dataset contains 120k slices with 1% in-distribution data. Filtering is particularly beneficial for the small VarNet with 9M parameters.

Retrieval Metric. Table 4 shows that DreamSim outperforms a pixel-based metric for filtering. The comparison is based on an unfiltered dataset of 40k slices for 8-fold accelerated MRI, where the pixel-level metric is the Euclidean distance. The pixel-based approach is ineffective on the cardiac test sets, resulting in performance degradation.

t-SNE visualization. To provide a visual interpretation of the filtering process, we applied t-SNE [24] to the DreamSim embeddings of the 120k slices dataset before and after alignment filtering,

Table 4: Comparison of DreamSim and a pixel-based metric (Euclidean distance) for alignment filtering on 40k slices with 8-fold acceleration. Numbers indicate the SSIM gain over no filtering. DreamSim shows consistent positive improvement in SSIM over no filtering. In contrast, the pixel-based approach performs poorly on the cardiac subset and yields no average improvement across all test sets.

Filtering strategy	Cardiac	non-Cardiac	All test sets
Pixel-based	–0.010	+0.003	0.0
DreamSim	+0.006	+0.006	+0.006

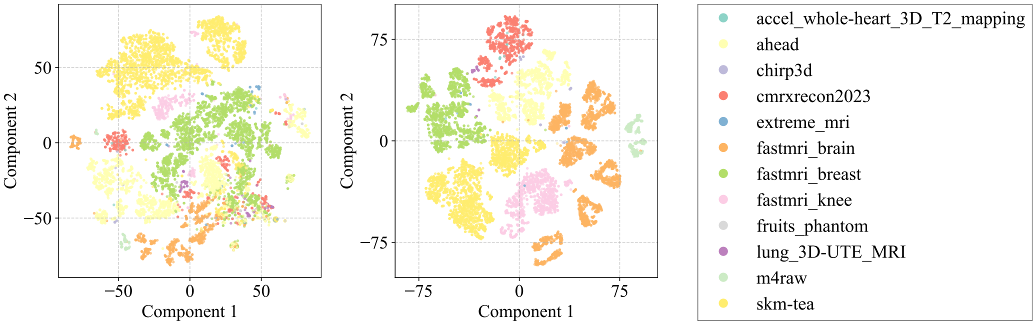

Figure 21: t-SNE visualization of DreamSim embeddings on the 120k-slice dataset, before (left) and after (right) applying alignment filtering. Each color corresponds to a different data source from the initial unfiltered data pool. The filtered dataset exhibits more distinct and separated clusters.

as shown in Figure 21. The unfiltered dataset (left) shows that embeddings from different data sources exhibit significant overlap. In contrast, the filtered dataset (right) displays considerably more distinct and well-separated clusters.

D Additional details for Section 3.6

In Section 3.6, we extend our results to diffusion model-based MRI reconstruction. This section provides background and implementation details for the diffusion models that we consider.

Background. A diffusion model $\boldsymbol{\epsilon}_\theta$ with parameters θ aims to learn a data distribution $p(\mathbf{x})$. Diffusion models consist of a forward process which gradually adds noise to images from the distribution $p(\mathbf{x})$, and a reverse process which aims to invert forward process. We adopt the denoising diffusion probabilistic models (DDPM) formulation [17], where each step of the forward process is Gaussian distributed $p(\mathbf{x}_{t+1}|\mathbf{x}_t) = \mathcal{N}(\sqrt{1-\beta_t}\mathbf{x}_t, \beta_t^2\mathbf{I})$ for time steps $t = 0, 1, \ldots, 1000$ and increasing noise levels β_t. The reverse diffusion process is modeled with Gaussian transition probabilities $p(\mathbf{x}_{t-1}|\mathbf{x}_t)$, and the mean of the Gaussian is learned with a neural network. The diffusion model $\boldsymbol{\epsilon}_\theta(\mathbf{x}; t)$ is trained using the residual denoising objective

$$\mathcal{L}(\theta, \mathbf{x}) = \mathbb{E}_{t\sim\mathcal{U}(0,T), \boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})}\left[\left\|\boldsymbol{\epsilon}_\theta(\sqrt{1-\sigma_t^2}\mathbf{x} + \sigma_t\boldsymbol{\epsilon}; t) - \boldsymbol{\epsilon}\right\|_2^2\right],$$

where $\sigma_t^2 = 1 - \Pi_{s=1}^t(1-\beta_s)$. Samples from the distribution $p(\mathbf{x}_0)$ can then be obtained by sampling $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ and successively applying the learned Gaussian transitions $p(\mathbf{x}_{t-1}|\mathbf{x}_t)$. Unlike end-to-end models, diffusion models do not require measurement data for training, but enforce data-consistency during reconstruction using the diffusion model as a pretrained image prior.

In our setup, the diffusion models learn the data distribution of the fully-sampled MVUE reconstructions. For the diffusion model, we choose the U-Net [34] architecture adopted from [8] with 80M parameters.

Table 5: Datasets used for accelerated 3D MRI setup. First two correspond to in-distribution datasets and the last two are used for out-of-distribution evaluation.

Dataset	Anatomy	View	Image contrast	Vendor	Magnet	Coils	Vol./Subj.
SKM-TEA [7]	knee	various	qDESS	GE	3T	8, 16	930/155
AHEAD [3]	brain	various	MP2RAGE-ME	Philips	7T	32	1.1k/77
Stanford-3D [9]	knee	various	CUBE (3D-FSE)	GE	3T	16	19/19
CC359 [38]	brain	various	GRE	GE	3T	32	165/165

Accelerated MRI is modeled as a linear inverse problem of the form $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{z}$, with linear forward operator $\mathbf{A}$ and additive Gaussian noise $\mathbf{z}$. We consider the following two approaches for reconstruction with diffusion models.

Posterior sampling. We consider decomposed diffusion sampling [6]. Assuming that $\mathbf{x}$ is drawn from the true data distribution of MVUE reconstructions, solving the inverse problem consists of sampling from the posterior $p(\mathbf{x}_0|\mathbf{y})$. Diffusion models enable posterior sampling by conditioning the reverse process $p(\mathbf{x}_{t-1}|\mathbf{x}_t, \hat{\mathbf{x}}_0(\mathbf{x}_t), \mathbf{y})$ on the measurements $\mathbf{y}$. In each step of the reverse sampling process, decomposed diffusion sampling updates the denoised estimate $\hat{\mathbf{x}}_0(\mathbf{x}_t)$ by minimizing $\frac{\gamma}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \frac{1}{2}\|\mathbf{x} - \hat{\mathbf{x}}_0(\mathbf{x}_t)\|_2^2$. This allows to control the influence of the diffusion prior via the estimate $\hat{\mathbf{x}}_0(\mathbf{x}_t)$. Moreover, we use denoising diffusion implicit model sampling to accelerate the sampling process [37].

Variational approach. We consider the approach proposed by Mardani et al. [25], which consists of solving $\min_q KL(q(\mathbf{x}_0|y), p(\mathbf{x}_0|\mathbf{y})))$, with a variational distribution $q = \mathcal{N}(\mu, \sigma^2)$. This motivates the following variational objective:

$$\hat{\mathbf{x}}(\mathbf{y}) = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{C}^N} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda \mathbb{E}_{t\sim\mathcal{U}(0,T'),\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})}\left[w(t)\left\|\boldsymbol{\epsilon}_\theta(\sqrt{1-\sigma_t^2}\mathbf{x} + \sigma_t\boldsymbol{\epsilon}; t) - \boldsymbol{\epsilon}\right\|_2^2\right],$$

where λ is a hyperparameter. We choose the time step dependent weighting factor $w(t)$ following [25]. The measurements $\mathbf{y}$ are scaled such that the reconstruction $\mathbf{x}$ has approximately unit variance. We minimize the objective using a first order gradient optimizer, and initialize with the zero-filled least-square reconstruction. Finally, we perform uniform time step sampling with upper bound $T' = 0.4 \cdot T$ (similar to [18]).

Choice of hyperparameters. The performance of diffusion models for reconstruction is strongly dependent on hyperparameter choices. For example, the performance of the variational approach critically depends on the choice of the regularization parameter λ. To more confidently attribute performance differences to variations in dataset design rather than suboptimal hyperparameter choices, we study diffusion models under best-case conditions: For both the posterior sampling and the variational approach for reconstruction with diffusion models, we tune hyperparameters for each sample in the test set individually with a grid search based on the ground-truth image.

Training. Similar to the end-to-end models we train the diffusion models until saturating reconstruction performance is reached on the validation set. We use the Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$, and a batch size of two. The learning rate is warmed up linearly to 4e-4 using 1% of total training time and then linearly decayed to 1.6e-5. Compute resources when using four workers are similar to the end-to-end experiments (Appendix A).

D.1 Results for 3D MRI reconstruction

In the following, we investigate filtering for diffusion model-based 3D MRI reconstruction. In previous sections, we used 3D MRI datasets only as auxiliary data sources for improving 2D MRI reconstruction performance. However, in this subsection we perform 3D reconstruction with 3D undersampling masks on a curated set of 3D volumes.

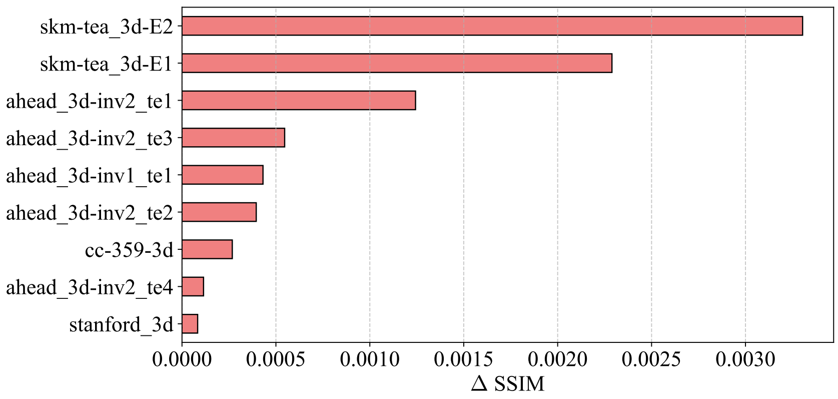

Figure 22: Detailed evaluation of weighted alignment filtering for 3D reconstruction performance at 36× acceleration. Similar to the 2D reconstruction results presented in Fig 16, we find that gains via filtering are larger on in-distribution samples (AHEAD and SKM-TEA) than on out-of-distribution samples (Stanford 3D and CC-359).

Variational 3D MRI reconstruction. We perform 3D reconstruction using 2D diffusion models trained on complex-valued MVUE reconstructions, and using a variational approach, where the diffusion model is applied to regularize randomly selected slices [18]. We minimize the following objective using gradient descent:

$$
\begin{aligned}
\hat{\mathbf{x}}(\mathbf{y}) = \underset{\mathbf{x}\in\mathbb{C}^N}{\text{argmin}} &\sum_{i=1}^{C} \|\mathbf{y}_i - \mathbf{M}\mathbf{F}_{\text{3D}}\mathbf{S}_i\mathbf{x}\|_2^2 \\
&+ \lambda \mathbb{E}_{\mathbf{s}\sim\text{2D-Slices}(\mathbf{x})}\left[\mathbb{E}_{t\sim\mathcal{U}(0,T'),\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})}\left[w(t)\left\|\boldsymbol{\epsilon}_\theta(\sqrt{1-\sigma_t^2}\mathbf{s}+\sigma_t\boldsymbol{\epsilon};t)-\boldsymbol{\epsilon}\right\|_2^2\right]\right].
\end{aligned}
$$

Here, $\mathbf{S}_i$ encodes the sensitivity map associated with the i-th receiver coil, $\mathbf{F}_{\text{3D}}$ is the 3D discrete Fourier transform, and $\mathbf{M}$ a 2D hybrid-cartesian Poisson undersampling mask in our experiments. Moreover, we employ a pre-trained 2D complex diffusion model $\boldsymbol{\epsilon}_\phi(\mathbf{x}_t, t)$. We follow the same instance-specific hyperparameter tuning method for λ as for 2D reconstruction, and approximate the expectation with respect to random slices by uniformly sampling 50 slices per anatomical view and gradient descent iteration.

We follow the same training setup as for 2D diffusion models, but scale the slices by the norm of the 3D volume during training.

Evaluation set. To evaluate the reconstruction performance on 3D MRI we curate a diverse set of 3D MRI volumes based on the datasets stated in Table 5. SKM-TEA and AHEAD contain two and five echoes, respectively. We split SKM-TEA into two subsets, one for each echo, and perform a similar split with the AHEAD dataset. During curation, we excluded many volumes with artifacts, such as wraparound or de-identification artifacts apparent in brain datasets. In total, our validation dataset consists of 30 in-distribution volumes and 9 out-of-distribution volumes.

Filtering. We perform weighted alignment filtering using a validation set similarly curated as the 3D evaluation dataset. We adapt the filtering method to 3D, by randomly selecting slices along all anatomical planes of the volumes, while excluding slices near the boundaries.

Results on datasets from the main body. We train a diffusion model on the unfiltered 120k slices dataset (same unfiltered dataset as in Section 3.3 for 4-fold accelerated 2D MRI), and train a diffusion model on the filtered dataset with 40k slices retained. Figure 22 provides a detailed evaluation for 36×-accelerated MRI. Similar to 2D MRI reconstruction, we observe that the benefit of weighted alignment filtering is larger on in-distribution data than on out-of-distribution datasets. However, on average, our filtering setup benefits 3D reconstruction performance only marginally (+0.001 SSIM) as shown in Figure 23 for 24-fold and 36-fold acceleration; therefore, we cannot conclude that filtering yields meaningful improvements.

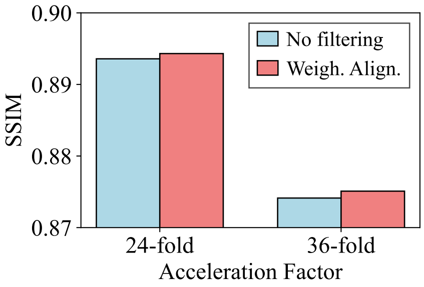

Figure 23: We train one 2D diffusion model on the 120k slices unfiltered dataset (same dataset as in Sec 3.3), and one model on the weighted alignment filtered dataset. We evaluate on a curated set of 3D volumes (see Table 5) for 24-fold acceleration and 36-fold acceleration. The gain obtained by weighted alignment filtering is 0.001 SSIM.

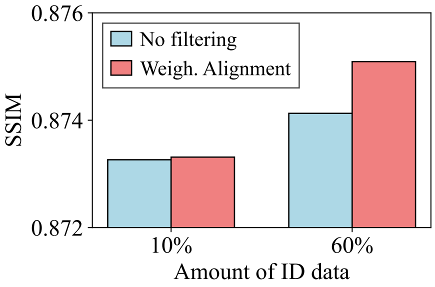

Figure 24: We perform filtering experiments with different fractions of in-distribution data (with respect to our 3D evaluation Table 5) in the unfiltered dataset with 120k slices in total. Different to the 2D MRI experiments, we do not observe improved performance of filtering when the fraction of in-distribution data is low.

Results for lower amounts of ID data. Note that the majority of the slices contained in the data pool (see Table 1) used in our work originates from 3D MRI and therefore the unfiltered dataset used in the previous paragraph contains around 60% in-distribution data (with respect to our 3D evaluation). However, in Section 3.4, we observed that filtering can provide a larger benefit when the fraction of in-distribution data in the unfiltered dataset is low. We investigate whether this holds true for our 3D MRI setup and create an unfiltered dataset with reduced in-distribution data (10%) by randomly sampling 10% of the data from 3D volumes with the reaming 90% from 2D volumes, totaling 120k slices. We perform weighted alignment filtering on this created unfiltered dataset, and present the reconstruction results with the correspondingly trained diffusion models in Figure 24. However, different to our results for 2D MRI, we do not find that filtering is more beneficial when the fraction of in-distribution data decreases.

E Licenses for datasets and software

- fastMRI datasets [47, 43, 36]: Custom agreement: https://fastmri.med.nyu.edu/
- CMRxRecon2023 [44]: CC-BY
- M4Raw [23, 23]: CC-BY 4.0
- SKM-TEA [7]: Custom agreement: https://stanfordaimi.azurewebsites.net/datasets/4aaeafb9-c6e6-4e3c-9188-3aaaf0e0a9e7
- AHEAD [3]: CC BY 4.0
- Lung 3D UTE [27]: CC0-1.0
- Chirp 3D [31]: CC-BY-4.0
- Extreme MRI [30]: CC-BY-4.0
- Fruits, Phantom [46]: CC-BY-4.0
- Heart T2-mapping [49]: CC0-1.0
- Stanford 2D [5]: CC BY-NC 4.0
- NYU data [15]: CC BY-NC 4.0
- SMURF [2]: CC0-1.0
- OCMR [4]: Custom agreement: https://www.ocmr.info/download/
- fastMRI code [47]: MIT License
- BART toolbox [42]: BSD-3-Clause license