Experimental verification (R2, R3). Although we agree that additional experimental verification would be valuable, we believe our conclusions are warranted. Before submission, we performed initial experiments to test the stability of our conclusions and the RL method: more epochs with various learning rates, larger policy models, and adding subsampling location features to the policy model input. We observed nothing that changed our conclusions. We would characterise the primary claims of our paper as a proof-of-concept for policy gradient methods, as well as the insight that adaptivity is key, and that variance provides at least a partial explanation for the greedy/non-greedy performance gap. We agree that additional explanations for this gap may exist, but believe our primary claims as stated are sufficiently supported. We explored various other reward baselines: a running average baseline for each acquisition step, both per state and averaged over states, and a baseline for the non-greedy model that used full returns of parallel trajectories. These approaches all underperformed the reported baseline. We additionally explored an extension to our reported baseline that samples actions *without replacement* and computes gradients using the estimator in [22]. This performed on par with our reported method. We chose to train on SSIM reward because it typically corresponds to human evaluations of image quality more closely than PSNR [21]. Interestingly, evaluating our SSIM non-adaptive and adaptive oracles on PSNR gives scores of 27.21, 27.50 on the base horizon, and 25.59, 26.13 on the long horizon task respectively, whereas SSIM scores were clearly higher for the long horizon task, and indeed this is what one would expect for oracles. This suggests that SSIM and PSNR care about distinct features (notably, PSNR seems to favour more low-frequency columns), which complicates drawing conclusions from PSNR evaluations of SSIM optimised methods: verification of reconstruction quality by human experts may be necessary. To this end, we will include example reconstructions in the paper ready version. We furthermore plan to add experiments on the fastMRI brain data mentioned by **R2**. Additional experimental results for $\gamma \in [0, 1]$ would indeed be valuable to verify the observed SNR trend that underpins our conclusions on gradient variance, and we plan to include an analysis of this in the camera-ready version (**R3**). Finally, as per suggestion of **R1**, we will include an equispaced baseline in Table 1: initial results indeed show improvements over the random baseline, but our models still dominate in performance.

Scaling to larger images (R2). We ran initial experiments on larger 256×256 images, which indicated that our models are still able to learn performant policies. We used the raw k-space data for all our experiments, and chose to use the smaller image setting for our final experiments solely due to computational constraints.

Differences with [18, 44] (R2). The approach in [18] uses an RL based method in which the reconstruction and policy models can be decoupled. Unlike our approach however, MCTS based training does not naturally allow for training greedy models. This aspect is crucial to our further analysis, which indicates that greedy models may be favoured. Additionally, our approach enjoys a computational advantage due to the use of smaller models and converges more quickly. Finally, direct policy optimisation involves fewer design choices than computing an MCTS distribution. We will include pseudo-code of our training process that shows additional discrepancies, such as the omission of a replay buffer (**R3**). The approach in [44] requires joint training of the reconstruction network with an evaluator network that guides acquisition through a similarity score between ground truth and fantasised k-space. Joint training is crucial, as the reconstruction network must be incentivised to produce reconstructions that have correct k-space representation for evaluator based acquisition to perform well. This contrasts with our method, where joint training is optional, and our acquisition function is directly (reinforcement) learned using policy gradients on image-space input. This also poses a challenge for making a fair comparison (using the same reconstruction model): the reconstruction model in [44] is incentivised to care about features that are not necessarily relevant to our policy, and our reconstruction method is not necessarily incentivised to care about features that are crucial to their evaluator. We did a proxy comparison using our reconstruction model and replacing their evaluator score with the true spectral map score computed from ground truth images. Using ground truth test images makes this an oracle method - infeasible in practice - but provides an upper bound for the performance of [44] under our reconstruction model, as we now use true spectral map scores, rather than the estimate learned by the evaluator network. However, this oracle method performed far worse than our models, suggesting that the strategy in [44] indeed depends heavily on reconstruction model design choices that force consistency of k-space, as well as on joint training with the evaluator. We also note that there is no code available for [44], further complicating attempts at a fair comparison. We will include this discussion in our paper.

Equation (2) (R2, R3). Equation (2) indeed erroneously conflates m and M. We will include the fixed formulation:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}} \left[\eta(\boldsymbol{x}, A_\theta(S_M F \boldsymbol{x}))\right], \quad S_M = \left[k_1, k_2, ..., k_M\right]^\intercal, \quad k_m \sim \pi(\boldsymbol{y}_m).$$

Finally, we thank the reviewers for indicating where the paper could be clearer in notation and contains inconsistencies in the discussion of related works: these will be addressed. We will furthermore include the suggested references. To answer some final questions: **R2**: In equation (5) γ is set to 1. A factor $\gamma^{t-t'}$ should indeed be included inside the sum over t' in general: we will clear this up. **R2**: In the MDP formulation as presented the reward indeed depends on the ground truth image, and transitions are only deterministic when additionally conditioned on this. We - like the reviewer - do not expect this point to affect our conclusions, but will fix it in the final paper. **R3**: The denominator $\frac{1}{B(B-1)}$ follows from equation (13) in [4], where we are computing $\hat{\sigma}^2_{\hat{\mu}} = \frac{1}{B}\mathrm{Var}[\hat{g}]$, with $\mathrm{Var}[\hat{g}] \approx \frac{1}{B-1}\sum_i (g_i - \hat{\mu})^2$.