This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material Cross-Field MRI Synthesis with an Unpaired Neural Schrödinger Bridge

Titouan Le Gourrierec[1], HyoSeok Lee[1], and Sung-Hong Park[1]

Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea
{titouanlegourrierec,hyord,sunghongpark}@kaist.ac.kr

S1 Per-contrast breakdown

Table S1. Per-contrast scores of the submitted system, with ranks over 36 teams. The ranks invert the absolute numbers: T2w holds the best raw score on all three intensity metrics and the worst rank on all five, while T2-FLAIR holds the worst raw scores and the best ranks.

Contrast	nRMSE ↓	LPIPS ↓	SSIM ↑	Dice ↑	VC ↑
T1w	0.2635 (4)	0.0641 (2)	0.9220 (4)	0.9127 (2)	0.9324 (2)
T2w	**0.2205** (6)	**0.0628** (6)	**0.9349** (5)	0.9125 (4)	0.9062 (6)
T2-FLAIR	0.2945 (3)	0.0759 (3)	0.9163 (3)	**0.8961** (1)	**0.9117** (1)
Mean	0.2595	0.0676	0.9244	0.9071	0.9168

S2 The pre-training budget

Referenced from Sec. 3.4. Stage-1 was re-run for every non-0.1 T cell on one uniform schedule; its epoch-5 and epoch-10 checkpoints were each deployed as a complete twelve-model submission.

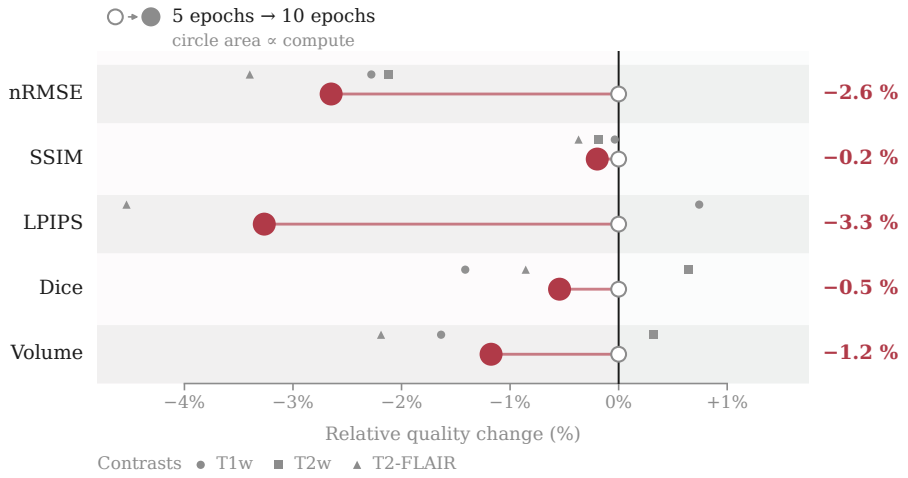

Fig. S1. Doubling the Stage-1 budget from 5 to 10 epochs, relative to the 5-epoch arm, scored on the challenge server. Filled circles are the mean over all twelve models, small markers the three contrasts. Every metric moves the same way, and T2-FLAIR degrades on all five, T1w on four, T2w on three.

Table S2. Stage-1 budget. Both arms read the *same* trajectory, so hub, warm-start chain, optimiser state and seed are fixed and only the budget varies; the three 0.1 T cells are identical in both, so nine of twelve carry the contrast and the per-model effect exceeds these means. The relative changes are those plotted in Fig. S1, and every one exceeds the run-to-run variability in the last row.

Stage-1 budget	nRMSE ↓	SSIM ↑	LPIPS ↓	Dice ↑	VC ↑
5 epochs	**0.2634**	**0.9228**	**0.0685**	**0.9031**	**0.9154**
10 epochs	0.2704	0.9210	0.0708	0.8982	0.9046
run-to-run variability	0.97%	0.02%	0.48%	0.07%	0.51%

S3 Complete ablation record

The main text reports the arms that support a claim; this is the full record, so that the selection is visible rather than implicit.

Table S3. Every arm scored on the challenge server, each a complete twelve-model deployment. Positive means the reference configuration is better than the arm. † = inference-only: these reuse one set of weights and a seeded rollout, so they carry no training variability and a 0.1% difference is exact; the rest are single runs carrying one draw of the variability in Sec. 3.1. All were measured on the sibling configuration of Sec. 3.3 and are evidence about the factor each varies.

Arm	nRMSE	SSIM	LPIPS	Dice	VC
Input channels					
1 channel (no context)	+8.3%	+1.4%	+13.4%	+3.1%	+5.8%
3 channels (through-plane only)	+1.0%	+0.4%	+4.7%	+3.1%	+3.0%
5 channels at every field	+1.9%	+0.1%	+2.3%	−0.7%	−1.2%
Inference					
No calibration †	+14.8%	+1.1%	+5.6%	0.00%	0.00%
Deterministic rollout, $\tau = 0$ †	+0.9%	+0.2%	+1.6%	+0.02%	+0.3%
Independent latent per slice †	−0.1%	−0.01%	−0.1%	−0.1%	−0.03%
Trimmed-mean calibration †	+1.1%	+0.1%	—	0.00%	0.00%
Fine-tuning					
$\lambda_{\text{SSIM}} = 0.5$	worse composite; not adopted				
Stop at epoch 20	−2.3%	+0.1%	+1.2%	+0.2%	+1.0%
Bridge and budget					
One-step rollout, $N = 1$	+3.4%	+0.2%	+2.2%	+0.4%	+0.9%
10 epochs of Stage-1 instead of 5	+2.6%	+0.2%	+3.2%	+0.5%	+1.2%

S4 Residual intensity scale

The calibration of Sec. 3.5 matches each prediction's brain mean to a per-contrast constant. This removes a global intensity offset, but does not in general minimise nRMSE, which is sensitive to the overall scale. For a fixed prediction $\hat{\boldsymbol{y}}$ and target $\boldsymbol{y}$,

$$\mathrm{nRMSE}(s) = \frac{\|s\hat{\boldsymbol{y}} - \boldsymbol{y}\|}{\|\boldsymbol{y}\|}$$

is minimised at

$$s^\star = \frac{\langle \boldsymbol{y}, \hat{\boldsymbol{y}} \rangle}{\langle \hat{\boldsymbol{y}}, \hat{\boldsymbol{y}} \rangle}.$$

Writing μ for the matched brain mean, σ for the corresponding standard deviations, and ρ for the prediction–target correlation,

$$s^\star = \frac{\mu^2 + \rho\,\sigma_{\boldsymbol{y}}\sigma_{\hat{\boldsymbol{y}}}}{\mu^2 + \sigma_{\hat{\boldsymbol{y}}}^2}, \qquad s^\star > 1 \iff \sigma_{\hat{\boldsymbol{y}}} < \rho\,\sigma_{\boldsymbol{y}}. \tag{1}$$

Thus, a scale above one indicates residual under-dispersion after mean calibration.

We estimated this residual scale using both held-out training data and challenge feedback. On the three available training subjects, we performed leave-one-subject-out calibration, and we also repeated the analysis by fitting the scale on a subset of slices and evaluating it on the remaining slices. These estimates consistently indicated a small residual global scale mismatch, with the largest correction on T2w. The final submitted scales were then fixed per contrast and applied uniformly at inference time.

This analysis is a post-processing diagnostic rather than a learned component of the synthesis model. It changes only the global intensity scale, not the spatial image content, and the held-out calibration experiments indicate that the effect is not specific to a single training subject or slice subset.

S5 Additional qualitative results

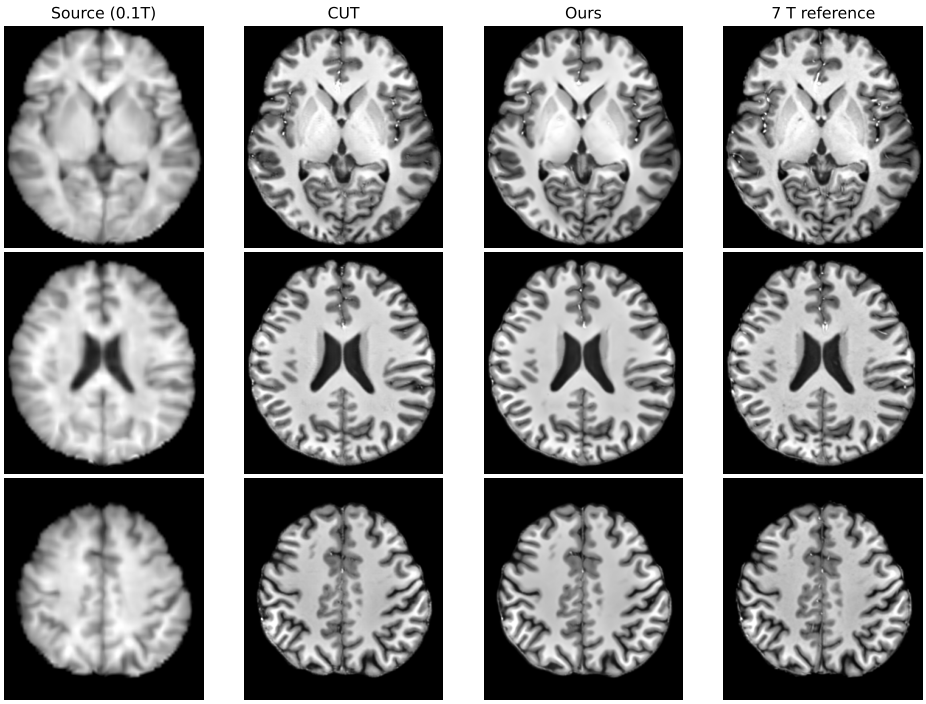

Fig. S2. T1w 0.1 T → 7 T: source, CUT, ours, and the acquired 7 T volume. Same training subject as the T2-FLAIR panel of the main text. We include the less flattering of the two: at T1w the baseline recovers cortical texture about as convincingly as our model, and the margin of Table S1 is not one the eye would find.