Supplementary Material: Multi-Field MRI Synthesis with Paired–Synthetic–Paired Latent Diffusion

Changseon Park[1⋆] and Doyeon Kim[1]

Research Institute, Neurophet Inc., Seoul, Republic of Korea
changseon@neurophet.com, doyeon@neurophet.com

Scope. This supplement provides additional analyses requested during review, including stage-wise development results, same-run comparisons of diffusion sample counts, inference cost, and additional analyses of synthetic source generation and validation selection. All challenge scores are aggregate Synapse validation metrics.

S1 Curriculum Development Results

Table S1 gives Task 3 endpoints spanning paired, synthetic, and final-paired development, followed by a more closely matched two-branch comparison.

Table S1. Task 3 development results. P and S denote acquired-pair training and synthetic expansion, respectively. The first three rows are heterogeneous development endpoints; the final four compare two raw U-Net initializations before and after the same 150k paired fine-tuning recipe, using different training random seeds.

Training endpoint	Submission	SSIM ↑	nRMSE ↓	LPIPS ↓
P→S (global 300k)→P 60k	9768383	0.886254	0.240271	0.089782
P→S (global 452k)→P 18k, no SSIM3	9771604	0.885412	0.231492	0.096731
P→S (global 452k)→P 54k, SSIM3	9771606	0.888801	0.241210	0.093247
Recipe-matched: P 200k, raw	9776808	0.877757	0.257505	0.096576
Recipe-matched: P 200k→P 150k	9775938	0.881271	0.246483	0.094001
Recipe-matched: P→S global 452k, raw	9776810	0.817785	0.301357	0.134626
Recipe-matched: P→S global 452k→P 150k	9775939	0.882501	0.246300	0.094393

The two fine-tuned rows use the same paired data and fine-tuning configuration, including latent scale, batch size, loss, learning rate, and 150k budget, but their U-Net initializations and training random seeds differ. Both weight lineages begin with acquired-pair training to 200k; one starts fine-tuning there, whereas the other first continues on synthetic data to global 452k.

⋆ Corresponding author: changseon@neurophet.com

The raw paired–synthetic endpoint is weaker than the paired-only endpoint by 0.059972 SSIM (0.817785 versus 0.877757), with nRMSE higher by 0.043852 and LPIPS higher by 0.038050, consistent with a synthetic-to-acquired domain gap. Under the same 150k-step fine-tuning recipe, its scores change by +0.064716 SSIM, −0.055057 nRMSE, and −0.040233 LPIPS, versus +0.003514, −0.011022, and −0.002575 for the paired-only branch. Thus paired recalibration recovers acquired-domain performance after synthetic expansion. The larger absolute recovery is descriptive rather than a causal estimate of initialization quality: the synthetic branch starts much lower, the training seeds differ, and only one run is available per branch.

At 150k, the paired–synthetic branch scores 0.001230 higher in SSIM and 0.000183 lower in nRMSE but 0.000392 higher in LPIPS. The SSIM difference is only 5.8% of the reported $K = 1$-to-$K = 6$ difference of 0.021165. These results do not establish a net benefit of synthetic expansion over paired-only training.

Table S2. Task 3 SSIM across paired fine-tuning steps. All entries use one-sample 50-step DDIM sampling. The fine-tuning configuration is shared, but the U-Net initializations and training random seeds differ. Step 0 evaluates the raw weights that `--init-unet` loads rather than the EMA shadow.

Initialization	0	108k	122k	150k
Paired-only 200k	0.877757	0.887485	0.886067	0.881271
Paired–synthetic 452k	0.817785	0.883664	0.886178	0.882501

Table S2 adds two cautions. The paired-only branch's 0.006214 spread across the shared fine-tuning steps exceeds each between-branch gap: its SSIM falls from 0.887485 at 108k to 0.881271 at 150k. The ordering is also unstable: paired-only leads by 0.003821 at 108k, the two branches are within 0.000111 at 122k, and paired–synthetic leads by 0.001230 at 150k. Each branch achieves its highest evaluated SSIM before 150k, so the 150k endpoint is one slice of a non-monotone trajectory rather than the best evaluated checkpoint of either branch. The large step-0-to-150k recovery of the synthetic branch remains descriptive because run-to-run variance is unavailable.

Two limits on this comparison remain. First, the initialization budgets are unequal at 452k versus 200k updates, introducing an additional training-duration and compute confound. We did not train a compute-matched paired-only branch, so we cannot separate the contribution of the synthetic data from that of the additional 252k updates. Second, each branch is a single training run evaluated with one sample, so run-to-run variance is not estimated. The comparison is therefore recipe-matched fine-tuning from each curriculum's own raw endpoint rather than a common-budget ablation.

S2 Effect and Cost of Multi-Sample Averaging

Number of samples. Table S3 expands the one-versus-six comparison with three-sample results. All $K = 1, 3, 6$ results were generated from saved outputs of the same 54k SSIM3 development run using 50-step DDIM sampling. Multi-sample predictions are voxel-wise arithmetic means of independently seeded outputs. For Tasks 1 and 2, segmentation is recomputed from the averaged image rather than averaged across seeds.

Table S3. One-, three-, and six-sample validation results. Panel (a) reports image metrics; panel (b) reports robust-mode segmentation metrics for Tasks 1–2, which Task 3 does not evaluate.

(a) Image metrics

Task	K	Seeds	Submission	SSIM ↑	nRMSE ↓	LPIPS ↓
1	1	1	9772744	0.898731	0.299373	0.088809
1	3	1–3	9772748	0.906110	0.300440	0.097640
1	6	1–6	9773385	0.911074	0.293686	0.102099
2	1	1	9772746	0.867612	0.226074	0.108193
2	3	1–3	9772749	0.887013	0.223927	0.119184
2	6	1–6	9773386	0.894016	0.216012	0.129384
3	1	0	9771606	0.888801	0.241210	0.093247
3	3	1–3	9773249	0.904248	0.234594	0.099322
3	6	1–6	9773387	0.909966	0.229181	0.105667

(b) Robust-mode SynthSeg metrics

Task	K	Seeds	Submission	Dice ↑	Volume ↑
1	1	1	9772744	0.832921	0.856148
1	3	1–3	9772748	0.837506	0.857896
1	6	1–6	9773385	0.839325	0.861351
2	1	1	9772746	0.773677	0.784530
2	3	1–3	9772749	0.796717	0.800286
2	6	1–6	9773386	0.808357	0.832733

SSIM improves monotonically from $K = 1$ to 3 to 6 in every task, whereas LPIPS worsens monotonically. nRMSE improves monotonically for Tasks 2 and 3. For Task 1, the three-sample nRMSE is slightly worse than one sample before the six-sample result improves over both. These trends support the interpretation that averaging suppresses stochastic high-frequency variation and stabilizes structure rather than uniformly improving perceptual fidelity.

Task 1 and the matched Task 2 rows use robust-mode SynthSeg. The selected Task 2 result used default-mode SynthSeg on the identical six-sample averaged image (submission 9773396) and obtained Dice 0.839531 and volume 0.841659. Image metrics are unchanged by segmentation mode. The Task 3 one-sample

baseline uses seed 0 and therefore lies outside its three- and six-sample seed sets; this prevents a strictly nested Task 3 comparison.

Inference time and memory. Table S4 summarizes 1,512 timing records from the validation inference runs used to construct the six-sample submissions. Timing excludes input loading, output writing, voxel averaging, packaging, and SynthSeg. Each seed is processed independently; therefore serial $K = 6$ GPU time is approximately six times $K = 1$. Peak CUDA memory is reported per single-seed process. Concurrent multi-seed sampling was not measured.

Table S4. Validation-slab inference cost. One output is one $364 \times 436 \times 96$ slab. Records are pooled over six seed runs. Median and mean are measured at $K = 1$; the $K = 6$ estimate is six times the measured mean. Memory is the measured single-seed process peak.

Task	Records	Median (s)	Mean (s)	$K = 6$ est. (s)	Alloc. (GiB)	Reserv. (GiB)
1	216	1.675	1.982	11.894	14.896	22.424
2	216	1.648	1.957	11.742	14.896	22.424
3	1,080	1.660	2.067	12.400	14.896	22.424

These measurements cover the cropped validation path, on which each output is a $364 \times 436 \times 96$ slab rather than a whole head, and should not be read as full-volume deployment latency. Occasional host-load stalls make the means higher than the medians.

DDIM steps versus averaging. In an auxiliary Task 3 comparison using a different development checkpoint, a 50-step single sample reached SSIM 0.885412. Increasing sampling to 200 steps produced SSIM 0.882103, nRMSE 0.235365, and LPIPS 0.096433 (submission 9772733). A three-sample 200-step average reached SSIM 0.900740 and nRMSE 0.227189, while LPIPS increased to 0.102427 (submission 9772736). This comparison supports averaging, rather than DDIM step count alone, as the source of the observed structural-metric gain, but remains auxiliary because it used a different development checkpoint from the selected run.

S3 Synthetic Source Model and Source-Field Token

Synthetic sources were generated with Gaussian-randomized tissue intensities, blur, clipping, a multiplicative bias field, additive Gaussian noise, gamma transformation, and randomized anisotropic downsampling and upsampling. We used this broad degradation distribution as domain randomization, with the aim of robustness to heterogeneous or unseen acquisition conditions, rather than as a calibrated simulator of the five challenge fields. Field-calibrated SNR may be better matched to the challenge domains, but we did not compare these two

objectives. The noise term is neither Rician nor noncentral-chi and is not conditioned on measured field-specific SNR. We therefore do not claim that the chosen distribution is superior, and did not perform a matched Gaussian-versus-Rician comparison. Final foreground normalization sets voxels outside the label mask to zero, so the synthetic sources also omit background magnitude-noise-floor behavior.

During synthetic expansion, renders without a physical source field were assigned source-field index 0, which is also used for acquired 0.1 T inputs. Thus, the same embedding represents both synthetic degradation and acquired 0.1 T appearance, potentially conflating the two domains. We did not evaluate a separate synthetic/unknown-source token and consider this, together with a matched Gaussian-versus-Rician or noncentral-chi renderer comparison, an important future control.

S4 Validation and Qualitative Limitations

Multiple leaderboard submissions were used to compare checkpoints and inference settings across the three tasks. This adaptive use of the validation set can make the selected scores optimistic and should not be interpreted as an unbiased estimate of generalization. Because the paired cohort contained only three subjects, a conventional subject-level internal validation split would leave only two subjects for paired training. An unbiased estimate requires a held-out test set and a prespecified selection rule.

The target-free validation images in the main paper demonstrate only field-conditioned appearance. They cannot establish voxelwise fidelity or exclude hallucination; the absence of target-referenced held-out qualitative evaluation remains a limitation.