Supplementary Material: An Auditable Geometry Contract for Multi-View Cardiac MRI Analysis

Ekaterina Antipushina, Daniil Sukhorukov, Elizaveta Cherkasskaya, Ruslan Kalimullin, Ilya Makarov, and Yury Koush

The view abbreviations are short-axis (SAX), two-chamber (2CH), four-chamber (4CH), and right-atrium-segmentation (RAS); LV and RV denote left- and right-ventricular cavities.

1 Exact Geometry Rewrite

For a Neuroimaging Informatics Technology Initiative (NIfTI) image–label pair with voxel array A, original affine M, and in-plane spacings (s_x, s_y), the converter executes the following operations under a stable case identifier:

1. load A without resampling and retain the untouched input path (hence M);
2. construct $M' = \mathrm{diag}(s_x, s_y, 999, 1)$;
3. save the same array as `Nifti1Image(A, M')` for image and label; after saving, the active `sform` equals M' with code 2 and `qform` code is 0;
4. train/infer with two-dimensional (2D) nnU-Net; invert the per-view label map; and
5. write the prediction array with the original affine/header from the input case.

Thus the training rewrite is not treated as invertible; restoration comes from the original file. No interpolation or voxel-array permutation occurs. The deployment contract requires three conditions before this rewrite is enabled: metadata must identify axis 2 as a non-spatial frame stack; its stored spacing must cause the planner to select another axis; and the prediction shape must match the untouched input before header restoration. If any condition fails, the adapter bypasses the rewrite. The challenge experiments used a uniform converter for all views; Section 2.2 documents why only cine changes the selected 2D plane.

2 Claim-to-Control Matrix

Table 1 makes explicit what was intervened on, what was held fixed, and what observation could have falsified each major claim. In particular, the late gadolinium enhancement (LGE) control tests specificity: if the header substitution affected a correctly planned stack, its plans, tensors or same-model predictions would differ.

Table 1. Controls supporting the principal claims.

Claim	Intervention and fixed factors	Falsification result / observation
Cine mechanism	Header only; array, labels, fold, trainer, budget and scorer fixed	All 3 views, 10 structures and score terms improve
LGE specificity	Untouched vs. rewritten header; same converged models	4/4 plans and all 400 tensors are exact; maximum Dice difference 1.1×10^{-5}
Scar-loss choice	Matching-fold refinement; budget, ensemble and threshold fixed	Neither alternative improves public and out-of-fold evidence; Dice plus cross-entropy is retained
Biomarker logic	Replace predictions by reference masks; retain metadata and estimator	All 15 ejection-fraction and 7 mass targets match within 0.01

Table 2. Per-structure Dice before and after geometry correction.

View	Structure	Original	Rewrite	Difference
Cine 2CH	LV cavity	0.900	0.922	+0.022
	Myocardium	0.807	0.861	+0.054
Cine 4CH	LV cavity	0.921	0.942	+0.021
	Myocardium	0.801	0.855	+0.054
	RV cavity	0.856	0.906	+0.050
	Right atrium	0.821	0.905	+0.084
	Left atrium	0.893	0.929	+0.036
Cine SAX	Myocardium	0.885	0.909	+0.024
	LV cavity	0.929	0.937	+0.008
	RV cavity	0.837	0.887	+0.049

2.1 Per-structure cine ablation

Table 2 expands the fold-0 geometry ablation from the main paper. Means are computed over the public validation patients. Corrected-minus-original differences are positive for every structure.

The mean-Dice correction wins in 15/15 patients for cine 2CH, 15/15 for 4CH, and 14/15 for SAX. Patient-bootstrap 95% intervals for corrected-minus-original mean Dice are $[0.027, 0.051]$, $[0.033, 0.068]$ and $[0.017, 0.038]$, respectively (20,000 percentile replicates; fixed bootstrap seed 20260731). Exact two-sided sign tests give $p = 6.10 \times 10^{-5}$, 6.10×10^{-5} and 9.77×10^{-4}, respectively; all remain significant after Bonferroni correction for the three views (largest adjusted $p = 0.00293$).

2.2 Dataset-wide header audit

Table 3 tests the planner-axis premise on every released train/validation/test image. For all 450 cine cases the concatenated frame axis has the smallest spacing and is never selected as through-plane. For all 299 LGE cases the genuine through-plane axis already has the largest spacing. We therefore built a no-rewrite control from each untouched LGE header. Across the 200 training cases, every preprocessed image and label tensor was exactly equal to its rewritten counterpart (400/400 files; maximum absolute difference zero), and the complete nnU-Net 2D configuration was identical for all four views. The temporary substitution therefore does not reach the LGE network input. This does not imply that arbitrary changes to a genuine 3D header are safe; a production adapter should still bypass the rewrite unless metadata identify a non-spatial concatenated axis.

Table 3. Raw-header audit over all released splits. Case counts are ordered train/validation/test. "Selected" counts cases in which axis 2 is already the largest-spacing (through-plane) axis.

Sequence	View	Cases by split	Selected	In-plane / axis-2 range (mm)
Cine	2CH	105/15/30	0/150	0.853–1.666 / 0.067–0.162
Cine	4CH	105/15/30	0/150	0.815–2.137 / 0.067–0.216
Cine	SAX	105/15/30	0/150	0.953–2.158 / 0.164–0.398
LGE	2CH	40/7/13	60/60	0.472–1.216 / 4.000–8.000
LGE	4CH	40/7/12	59/59	0.459–1.377 / 5.000–6.000
LGE	SAX	40/7/13	60/60	0.457–1.289 / 7.000–11.500
LGE	RAS	80/14/26	120/120	1.090–1.555 / 9.778

Table 4 completes the control by applying the same trained five-fold models to public-validation inputs prepared from rewritten and untouched headers. Dice is exactly equal for 2CH, 4CH and RAS. The SAX difference is 1.12×10^{-5} in favour of the untouched header and is immaterial at the reported three-decimal precision. Together with tensor/configuration identity, this verifies absence of practically relevant LGE harm for the released dataset; it is not a licence to alter arbitrary genuine 3D data.

3 Selected Inference Configuration

All outputs average the five fold softmax predictions. Mirroring-based test-time augmentation (TTA) is view specific. Table 5 records the view configuration chosen on the public development set for the revised results.

For LGE 2CH, no-TTA versus TTA Dice is 0.7113 versus 0.7059; the patient-bootstrap 95% interval for the paired gain is $[0.00004, 0.01342]$, and no-TTA

Table 4. LGE no-rewrite control. "Exact tensors" includes each training image and label; both paths use the same trained model.

View	Exact tensors	Rewrite Dice	Original-header Dice
LGE 2CH	80/80	0.711318	0.711318
LGE 4CH	80/80	0.736840	0.736840
LGE SAX	80/80	0.681665	0.681676
LGE RAS	160/160	0.874341	0.874341

Table 5. Selected per-view training and inference settings.

View	Epochs	Mirroring TTA	Additional rule
Cine 2CH/4CH/SAX	250	yes	none
LGE 2CH	250	no	none
LGE 4CH	1000	no	none
LGE SAX	250	no	$p(\text{scar}) \geq 0.9996$
LGE RAS	1000	yes	none

wins 6/7 cases. For LGE 4CH the values are 0.7407 versus 0.7398, with interval $[-0.00018, 0.00182]$ (5/7 wins), so the effect is negligible. For thresholded LGE SAX, no-TTA versus TTA improves mean/scar Dice from 0.7158/0.5254 to 0.7182/0.5495; the scar-mass Pearson correlation coefficient (PCC; vPCC in the challenge notation) and relative absolute error (RAE) change from 0.8439/0.4848 to 0.8509/0.4791; no-TTA is therefore selected.

3.1 Official composite score

The public score combines segmentation and biomarker terms as

$$\text{Task1} = 0.7\big(0.4\,\text{DSC} + 0.3/(1{+}\text{HD}) + 0.3/(1{+}\text{ASD})\big) + 0.3\max(0, \text{PCC}),$$
$$\text{Task2} = 0.7\,\text{DSC} + 0.3\big(0.5\max(0, \text{vPCC}) + 0.5/(1{+}\text{RAE})\big).$$

The selected system obtains Task 1 0.716, Task 2 0.764 and overall 0.740 on public validation. These values describe the development configuration and are not external-test estimates.

4 Isolated Ensemble Experiment

Table 6 isolates ensembling from TTA: both arms use the same 250-epoch models and mirroring, and differ only in fold 0 versus five-fold softmax averaging. At task level, cine Dice changes from 0.9052 to 0.9065 and LGE Dice from 0.7368 to 0.7507. For cine, a patient-clustered bootstrap over the 15 patients (three views per resample) gives $\Delta = +0.0012$ with a 95% confidence interval (CI)

Table 6. TTA-matched ensemble ablation on public validation. Intervals are paired patient-bootstrap 95% CIs for view-level mean Dice (20,000 replicates; seed 20260808). The final column reports ensemble-minus-fold-0 Dice for each structure.

View	Fold 0	Five folds	Δ [95% CI]	Per-structure ΔDice
Cine 2CH	0.8932	0.8937	+0.0004 [−0.0032, +0.0038]	LV_c −0.0010; Myo +0.0018
Cine 4CH	0.9092	0.9102	+0.0011 [−0.0001, +0.0023]	LV_c +0.0011; Myo +0.0004; RV_c +0.0016; RA +0.0009; LA +0.0014
Cine SAX	0.9133	0.9156	+0.0022 [−0.0006, +0.0050]	Myo +0.0009; LV_c +0.0006; RV_c +0.0052
LGE 2CH	0.7053	0.7059	+0.0006 [−0.0085, +0.0095]	LV_c −0.0022; Myo +0.0029; Scar +0.0013
LGE 4CH	0.7299	0.7372	+0.0073 [−0.0040, +0.0176]	LV_c +0.0134; Myo +0.0155; Scar −0.0065; RV_c +0.0067
LGE SAX	0.6449	0.6853	+0.0404 [−0.0049, +0.1139]	LV_c +0.0077; Myo +0.0110; Scar +0.1387; RV_c +0.0040
LGE RAS	0.8671	0.8743	+0.0072 [+0.0021, +0.0139]	RA +0.0072

of [−0.0006, +0.0029] (11/15 patient wins). Thus the isolated cine overlap effect is small and uncertain; the challenge-faithful maximum Hausdorff-distance comparison shows a clearer boundary-outlier reduction, from 8.31 to 7.99 mm.

The decomposition prevents two overclaims. First, ensembling is not the source of a statistically established cine Dice gain when TTA is held fixed; its principal cine benefit is boundary robustness. Second, the LGE gain is heterogeneous: it is dominated by SAX scar, while 2CH is essentially unchanged and 4CH scar decreases slightly. The public LGE intervals are wide at $n = 7$ (RAS: $n = 14$), so these are mechanism-oriented development-set results, not independent evidence of generalisation. Five-fold inference costs approximately five times a single fold.

5 Scar-Focused Experiments

The LGE SAX experiments in Table 7 use the official class-then-view aggregation. Longer training and foreground oversampling do not explain the final gain; the gain comes from conservative confidence filtering of argmax scar predictions.

With the selected no-TTA inference, leave-one-case-out threshold selection chooses 0.9996 in 6/7 folds and 0.998 in the remaining fold, yielding mean/scar Dice 0.7149/0.5364 and mass vPCC/RAE 0.8041/0.6679.

We additionally refined each converged Dice plus cross-entropy (Dice+CE) fold for 25 epochs at learning rate 0.001 using either Tversky+CE ($\alpha = 0.3, \beta = 0.7$) or Dice+focal ($\gamma = 2$). Each public result is a five-fold ensemble; each

Table 7. Targeted LGE SAX experiments on public validation.

Variant	Mean Dice	Scar Dice
250 epochs, argmax	0.6853	0.3918
1000 epochs, ordinary	0.6755	–
1000 epochs, foreground oversampling 0.55	0.6782	–
250 epochs, confidence threshold 0.9996	0.7158	0.5254

training case in the out-of-fold (OOF) result is predicted only by its held-out fold. Table 8 uses no mirroring TTA and applies the same pre-specified scar threshold 0.9996 to every loss, isolating the loss effect.

Table 8. Controlled LGE SAX loss experiment. Each entry is mean/scar Dice. OOF uses all 40 training cases and is an internal patient-independent check, not external validation.

Loss	Public raw	Public threshold	OOF raw	OOF threshold
Dice+CE	0.6817/0.3922	**0.7182/0.5495**	**0.6823/0.3761**	0.6929/0.4233
Tversky+CE	0.6794/0.3963	0.6805/0.4075	0.6761/0.3554	**0.6968/0.4438**
Dice+focal	0.6791/0.3943	0.7135/0.5417	0.6689/0.3263	0.6927/0.4266

Tversky improves thresholded OOF scar Dice by 0.0206, but this does not transfer to public validation: public thresholded mean Dice falls from 0.7182 to 0.6805 and scar Dice from 0.5495 to 0.4075. Focal is also below the baseline publicly and reduces raw OOF mean Dice. We therefore retain Dice+CE rather than selecting an asymmetric loss from one favourable internal endpoint. For completeness, the Tversky-minus-baseline thresholded OOF mean-Dice difference is $+0.0039$ (95% bootstrap CI $[-0.0004, +0.0087]$), whereas its public difference is -0.0378 ($[-0.1113, +0.0070]$). Focal significantly reduces raw OOF mean Dice by -0.0134 ($[-0.0323, -0.0004]$).

6 Per-Case Scar-Mass Results

Table 9 gives all seven public-validation pairs after the selected SAX confidence rule. Two cases have zero reference scar. Across the cohort: Pearson $r = 0.851$ (Fisher-z 95% CI $[0.272, 0.978]$), RAE $= 0.479$ (patient-bootstrap 95% CI $[0.160, 0.784]$), mean absolute error (MAE) $= 15.2$ g (bootstrap 95% CI $[1.4, 32.7]$), bias $= -15.2$ g (bootstrap 95% CI $[-32.7, -1.3]$), and Bland–Altman limits of agreement $[-61.4, 31.0]$ g. Bootstrap intervals use 100,000 patient resamples and seed 20260804.

The leave-one-case-out selected masses yield $r = 0.804$ and RAE $= 0.668$. This is more stable than the TTA configuration, but the seven-case cohort remains too small for a clinical claim.

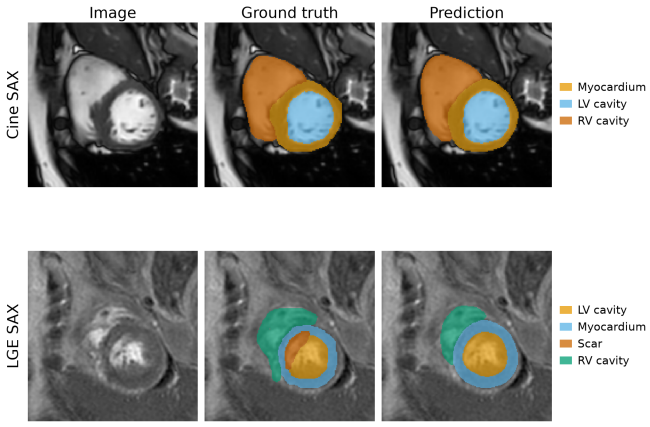

Fig. 1. Representative validation predictions. Chambers and myocardium are delineated accurately in cine (top) and LGE (bottom), whereas LGE scar is undersegmented.

Table 9. Reference and predicted scar mass for every LGE SAX validation case.

Case	Reference (g)	Prediction (g)	Error (g)
LGE_SAX_041	0.00	0.00	0.00
LGE_SAX_042	10.58	1.38	-9.20
LGE_SAX_043	0.00	0.00	0.00
LGE_SAX_044	54.99	21.29	-33.70
LGE_SAX_045	5.67	5.76	+0.10
LGE_SAX_046	19.51	16.83	-2.68
LGE_SAX_047	80.44	19.38	-61.06

7 Left-Ventricular Ejection-Fraction Smoother Selection

The cyclic seven-frame mean is selected in all 15 leave-one-public-validation-case-out fits. On the larger set of 105 out-of-fold training predictions it is also the modal winner under bootstrap perturbation and yields leave-one-out Pearson $r = 0.906$. On public validation it raises Pearson correlation from 0.891 to 0.920, but changes MAE from 4.4 to 5.9 ejection-fraction points and bias from $+0.2$ to -4.8. It is therefore a challenge-correlation choice, not an agreement improvement.