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
Native-Resolution Multi-Sequence Cardiac MRI Segmentation for Accurate LVEF and Scar Quantification

Gagneet Singh[1], Adinath Dukre[2], and Imran Razzak[1,2]

[1] MedOS, Abu Dhabi, UAE
[2] MBZUAI, Abu Dhabi, UAE

S1 Architectural variants and test-time augmentation

Table 1 lists every architectural variant evaluated, the failure mode each targets, the scoring protocol used, and its significance against the stated reference. Canonical (^{c}) and strict (^{s}) protocols are as defined in Sec. 2.1 of the main paper. Variants marked *tied* are not separated from their reference at $p < 0.05$ after Holm–Bonferroni correction; a tie is reported as a tie rather than as a small positive result.

Table 1. Architectural variants, both tasks. Δ is against the stated baseline.

Task	Variant	Targets	Baseline	Score	Δ	Significance
LGE	Best from-scratch ensemble (+TTA)	capacity	E1, 0.685	0.740^{c}	+0.055	t=5.32
LGE	**SMP (ImageNet ResNet-34)**	transfer	E1, 0.685	**0.778**^{c}	+0.093	t=4.10
LGE	CASA (contrast-aware scar attention)	scar contrast	E1, 0.685	≈E1	~ 0	tied
LGE	E5 (anatomy-gated head)	anatomical consistency	E1, 0.685	≈E1	~ 0	tied
LGE	Focal–Tversky loss (recall-biased)	scar recall	E1, 0.685	≈E1	~ 0	tied
LGE	UNet++ decoder	fine-structure localization	SMP	mildly +	< splices	not kept
LGE	Frozen MedSAM-style encoder	transfer	SMP	degraded	–	rejected
LGE	Oracle-box ROI-crop cascade	localization	SMP	mildly +	< splices	not kept
LGE	corrector3D (standalone)	through-plane consistency	0.7135^{s}	0.6665^{s}	−0.047	rejected
LGE	**corrector3D scar splice**	through-plane (scar only)	SMP+PP, 0.749	**0.765**^{s}	+0.016	deterministic
Cine	**Hybrid (2D, native-res)**	resolution ceiling	3D baseline, 0.856	**0.899**	+0.043	t=12.78
Cine	Y3D (frame-pair fusion)	temporal context	Hybrid, 0.899	0.899	0.000	t=−0.26 (tied)

Table 2 reports the per-view effect of flip test-time augmentation on the LGE validation set, evaluated as the mean Dice over ground-truth-present classes. The gain is positive in all four views but is concentrated in 4CH and SAX; 2CH and RAS are essentially unchanged. This qualifies the pooled TTA improvement quoted in the main paper: the intervention is consistent in *sign* across views, not uniform in magnitude.

S2 Mass-calibration fit and leave-one-out stability

The log-space mass calibration $\text{mass}_{\text{gt}} \approx \exp\big(A \log(1 + \text{mass}_{\text{pred}}) + B\big) - 1$ is fitted on the 30 SAX training volumes carrying non-zero ground-truth scar (of

Table 2. Per-view effect of flip TTA (LGE validation, mean Dice over GT-present classes).

View	without TTA	with TTA	Δ
2CH	0.7017	0.7024	+0.0007
4CH	0.7800	0.7927	+0.0127
SAX	0.7306	0.7541	+0.0236
RAS	0.8476	0.8502	+0.0027
mean			+0.0099

40 SAX training volumes in total). No validation case enters the fit. Table 3 reports the full-data coefficients, their range under leave-one-out refitting, and held-out mean absolute percentage error (MAPE) obtained by calibrating each held-out case with coefficients fitted without it.

Table 3. Calibration coefficients and leave-one-out behaviour ($n = 30$ training cases).

Quantity	A	B
Full-data fit	0.8667	0.4427
Leave-one-out range	[0.852, 0.909]	[0.273, 0.482]
Held-out MAPE over the 30 training cases		
Uncalibrated (raw voxel-count mass)		0.2719
Log-space calibration (leave-one-out)		**0.2425**
Single linear scale factor (leave-one-out)		0.2471

The coefficients move little across folds, and the log-space map is the only variant that improves on the uncalibrated mass under leave-one-out; the single global scale factor recovers less of the error and is rejected. The map is nevertheless a two-parameter fit from a single cohort and would require refitting under a different contrast or scanner protocol.

S3 Bootstrap confidence intervals

All intervals are 95% percentile bootstrap over patients, 10,000 resamples, resampling whole patients rather than slices. Cine Dice uses $n = 15$ per view; LVEF uses $n = 15$; scar-mass relative absolute error (RAE) is computed over the 5 of 7 validation SAX cases carrying non-zero ground-truth scar, following the challenge definition $\mathrm{RAE} = \mathrm{mean}\,|m_{\mathrm{pred}} - m_{\mathrm{gt}}|/m_{\mathrm{gt}}$.

Two readings matter. First, the Cine comparisons are decisive in opposite directions: the Hybrid advantage over the fixed-cube 3D baseline excludes zero in every view, whereas the Hybrid–Y3D difference straddles zero, so the tie reported in the main paper is a statistical tie rather than an underpowered comparison.

Table 4. Headline quantities with 95% bootstrap confidence intervals.

Quantity	Setting	Value 95% CI
LVEF PCC	Hybrid vs. GT, n=15	0.9724 [0.9265, 0.9907]
LVEF MAE (pp)	Hybrid vs. GT, n=15	3.107 [2.083, 4.361]
Cine Dice 2CH	Hybrid, n=15	0.8848 [0.8736, 0.8952]
Cine Dice 4CH	Hybrid, n=15	0.9064 [0.8947, 0.9166]
Cine Dice SAX	Hybrid, n=15	0.9055 [0.8901, 0.9181]
Cine Dice pooled	Hybrid, n=45	0.8989 [0.8909, 0.9062]
Δ Hybrid − 3D baseline	pooled, paired	+0.0426 [+0.0361, +0.0490]
Δ Hybrid − Y3D	pooled, paired	−0.0004 [−0.0038, +0.0030]
Scar-mass RAE, before calibration	n=5	0.4384 [0.2459, 0.7296]
Scar-mass RAE, after calibration	n=5	0.3707 [0.1778, 0.6471]
Δ RAE (paired reduction)	n=5	+0.0676 [−0.0259, +0.1611]

Second, the scar-mass improvement is a point-estimate gain whose paired interval includes zero at $n = 5$; three of the five evaluable cases improve. The evidence supporting the calibration stage is therefore the leave-one-out result on 30 training cases (Table 3), not the five validation cases, and we report it as such.