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.

METS-PRO R2: Lesion-Aware Positive Residual Refinement for Brain Metastasis Segmentation Supplementary File

Manthan Singla[1 [0009−0004−5898−6949]], Ajo Babu George[2 [0009−0005−3026−0959]], and Uma Ranjan[1 [0000−−0001−6258−4513]]

[1] Indian Institute of Technology Jammu
[2] DiceMed Pvt. Ltd.

1 Additional Qualitative results

An additional qualitative example as in Fig 1 shows improved enhancing tumour agreement with METS PRO R2 compared with C0, supporting the observed benefit while providing a separate visual case beyond the main manuscript.The ground truth ET region is small and spatially localized. C0 identifies the approximate lesion location but shows less precise overlap with the annotated enhancing component. In contrast, METS PRO R2 produces a prediction that more closely follows the ground truth ET boundary while retaining the surrounding anatomical context. The improvements are consistent with the lesion detection gains observed quantitatively.

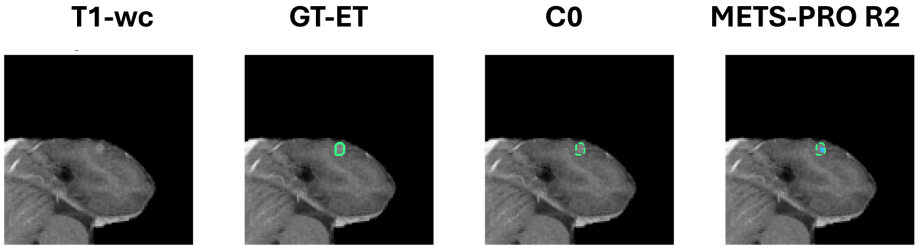


Fig. 1. ET Improvement C0 vs R2

2 Detailed Five-Fold Split and Patient-Level Audit

Tables 1 and 2 describe the data split by fold. The five-fold challenge split was defined at the scan level, with 1,036 training and 259 validation scans per fold. Because longitudinal scans from the same patient may occur across partitions, we audited patient identifiers within each fold. This identified 582 validation scans from 552 unique patients whose identifiers were absent from the corresponding

training partition. These scans were used to assess the robustness of the out-of-fold results to patient-level overlap; no additional patient-grouped model training was performed.

Table 1. Five-fold split and patient-level overlap audit.

Fold	Train	Val.	Shared patients	Independent scans	Overlap scans
0	1036	259	125	112	147
1	1036	259	118	115	144
2	1036	259	118	121	138
3	1036	259	117	115	144
4	1036	259	120	119	140
Total	–	1295	–	582	713

Table 2. Patient-independent subset by fold.

Fold	Independent scans	Unique patients
0	112	107
1	115	111
2	121	114
3	115	107
4	119	113
Total	582	552

3 Complete OOF results

The complete five-fold out-of-fold evaluation is provided in Tables 3 and 4. Table 3 reports lesion-level performance across ET, TC, WT, and RC using micro-F1, macro-F1, DSC, NSD, and TP/FP/FN counts, while Table 4 focuses on small-lesion detection performance

4 Five-Fold Matched Continuation Control

For each fold, C0 and R2 were initialized from the same 600-epoch baseline checkpoint and continued for 60 epochs. C0 used the baseline configuration, whereas R2 included the proposed refinement components. Thus, the two models differed only in the presence of the METS-PRO components. Table 5 reports the fold-wise results.

Table 3. Complete five-fold OOF results for all lesions.

Model	Region	Micro-F1	Macro-F1	DSC	NSD	TP/FP/FN
Baseline	ET	0.7056	0.6880	0.5677	0.6387	6390/1610/3722
C0	ET	0.7071	0.6875	0.5665	0.6365	6416/1619/3696
R2	ET	0.7171	0.6936	0.5680	0.6367	6597/1691/3515
Baseline	TC	0.7186	0.7034	0.5905	0.6491	6381/1469/3528
C0	TC	0.7209	0.7038	0.5900	0.6480	6405/1455/3504
R2	TC	0.7298	0.7088	0.5898	0.6472	6589/1560/3320
Baseline	WT	0.7044	0.6949	0.5526	0.5709	5825/1510/3379
C0	WT	0.7064	0.6978	0.5546	0.5728	5847/1503/3357
R2	WT	0.7148	0.6996	0.5555	0.5733	6013/1608/3191
Baseline	RC	0.1427	0.1408	0.3157	0.2526	150/1699/103
C0	RC	0.1415	0.1429	0.3307	0.2650	153/1757/100
R2	RC	0.1435	0.1368	0.3087	0.2477	151/1701/102

Table 4. Small-lesion results on the complete five-fold OOF cohort.

Model	Region	Micro-F1	Macro-F1	TP	FP	FN
Baseline	ET	0.4107	0.3344	1316	930	2846
C0	ET	0.4140	0.3388	1336	956	2826
R2	ET	0.4417	0.3590	1465	1006	2697
Baseline	TC	0.4194	0.3436	1253	823	2646
C0	TC	0.4248	0.3502	1273	822	2626
R2	TC	0.4528	0.3692	1402	891	2497
Baseline	WT	0.3683	0.2761	951	840	2422
C0	WT	0.3718	0.2800	966	857	2407
R2	WT	0.3953	0.3025	1060	930	2313
Baseline	RC	0.0112	0.0100	1	115	61
C0	RC	0.0255	0.0292	2	93	60
R2	RC	0.0111	0.0083	1	118	61

5 Patient-Independent Out-of-Fold Analysis

The OOF predictions were restricted to the 582 patient-independent scans from 552 unique patients. No additional training was performed. Table 6 reports the corresponding baseline, C0, and R2 results.

6 Paired Statistical Analysis

Statistical comparisons were performed using patient-level paired differences. For each patient, the corresponding baseline and R2 performance values were paired before resampling. This avoids treating multiple longitudinal scans from the same patient as independent statistical units.

For each comparison, 20,000 patient-level bootstrap replicates were used to obtain a 95% confidence interval for the paired difference. A two-sided paired permutation test with 20,000 permutations was used to assess the null hypothesis

Table 5. Fold-wise lesion-count micro-F1 for the baseline, matched continuation control (C0), and METS-PRO R2.

	ET			TC			WT			RC		
Fold	Base	C0	R2	Base	C0	R2	Base	C0	R2	Base	C0	R2
0	0.6826	0.6783	0.6890	0.6956	0.6938	0.7041	0.6798	0.6754	0.6831	0.1368	0.1155	0.1144
1	0.6714	0.6790	0.6943	0.6836	0.6935	0.7034	0.6876	0.6969	0.7052	0.1875	0.1990	0.2057
2	0.7535	0.7499	0.7614	0.7683	0.7639	0.7751	0.7446	0.7415	0.7503	0.1222	0.1250	0.1076
3	0.7070	0.7058	0.7145	0.7179	0.7171	0.7274	0.6947	0.6967	0.7065	0.1107	0.1139	0.1388
4	0.7068	0.7168	0.7189	0.7209	0.7307	0.7319	0.7106	0.7175	0.7236	0.1702	0.1761	0.1611
Mean Δ (R2–C0)	+0.00967			+0.00860			+0.00817			-0.00036		

Table 6. Patient-independent OOF evaluation on 582 scans from 552 unique patients.

Model	Region	Micro-F1	Macro-F1	Small-F1	DSC	NSD	TP/FP/FN
Baseline	ET	0.7095	0.7319	0.3835	0.5592	0.6386	2801/727/1567
C0	ET	0.7097	0.7340	0.3821	0.5583	0.6368	2805/732/1563
R2	ET	**0.7215**	0.7390	**0.4123**	0.5606	0.6387	2892/757/1476
Baseline	TC	0.7241	0.7523	0.4063	0.5941	0.6562	2800/644/1490
C0	TC	0.7256	0.7561	0.4079	0.5952	0.6566	2804/635/1486
R2	TC	**0.7344**	0.7581	**0.4338**	0.5916	0.6537	2890/690/1400
Baseline	WT	0.7102	0.7389	0.3658	0.5489	0.5605	2493/700/1335
C0	WT	0.7127	0.7460	0.3709	0.5577	0.5674	2496/680/1332
R2	WT	**0.7217**	0.7469	**0.3940**	0.5554	0.5658	2573/729/1255
Baseline	RC	0.0459	0.0455	0.0000	0.3425	0.2569	20/823/8
C0	RC	0.0441	0.0458	0.0000	0.3716	0.2806	20/859/8
R2	RC	0.0446	0.0412	0.0000	0.3478	0.2598	19/805/9

of zero paired improvement. For comparisons across the four regions, permutation p-values were adjusted using the Holm procedure. Random seeds were fixed and recorded for reproducibility.

7 Training Constraints associated with Hidden Validation

The hidden validation phase for the segmentation tasks (Tasks 1) enforces strict computational limitations to guarantee standardized evaluations.(Table 9) Submissions were required to operate within a Docker container restricted to 48 GiB of system memory and 16 GiB of shared memory. Total inference processing time was capped at 12 hours. Additionally, algorithms had fully allocated 16 vCPUs, 200 GB of storage, and an NVIDIA A10G GPU with 24GB of VRAM, supporting up to CUDA version 13.0.

Table 7. Patient-clustered paired statistical analysis of R2 versus the original baseline on the complete OOF cohort. Δ denotes R2 minus baseline micro-F1. Confidence intervals are 95% bootstrap intervals; p_{Holm} is the two-sided permutation p-value after Holm correction across the four regions.

Metric	Region	Patients	Δ	95% CI	p	p_{Holm}
All	ET	810	+0.01146	[0.00707,0.01573]	<0.0001	0.00020
All	TC	810	+0.01114	[0.00660,0.01553]	< 0.0001	0.00020
All	WT	810	+0.01037	[0.00551,0.01528]	<0.0001	0.00020
All	RC	810	+0.00075	[-0.00715,0.00870]	0.8689	0.8689
Small	ET	475	+0.03099	[0.02276,0.03905]	<0.0001	0.00020
Small	TC	452	+0.03343	[0.02483,0.04208]	<0.0001	0.00020
Small	WT	433	+0.02698	[0.01800,0.03589]	<0.0001	0.00020
Small	RC	28	-0.00019	[-0.00241,0.00141]	0.8381	0.8381

Table 8. Patient-clustered paired statistical analysis on the 582-scan patient-independent subset.

Metric	Region	Patients	Δ	95% CI	p	p_{Holm}
All	ET	552	+0.01199	[0.00491,0.01889]	0.0013	0.0052
All	TC	552	+.01036	[0.00321,0.01727]	0.0043	.0092
All	WT	552	+.01158	[0.00383,0.01915]	0.0031	0.0092
All	RC	552	-.00132	[-0.00826,0.00437]	0.7158	0.7158
Small	ET	266	+.02883	[0.01579,0.04272]	0.0001	0.0004
Small	TC	246	+0.02750	[0.01335,0.04163]	0.0004	0.0011
Small	WT	228	+0.02813	[0.01477,.04128]	0.0004	0.0011
Small	RC	3	0.00000	[0,0]	1.0000	1.0000

Table 9. Constraints for Hidden Validation (MICCAI BraTS 2026 Tasks 1)

Constraint Type	**Specification**
Compute Resources	
GPU	NVIDIA A10G (24GB VRAM)
vCPUs	16
Storage	200 GB
Max CUDA version	13.0
Docker Runtime Limits	
Max memory (`-memory`)	48 GiB
Shared memory (`-shm-size`)	16 GiB
Inference Time Limit	
Total time	12 hours