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 GAT-26: Release-Path Auditing and Confirmation-Gated Inference Selection for Cross-Tumor Brain Tumor Segmentation

Nathan Chen

Kang Chiao International School, Xiugang Campus, New Taipei City, Taiwan
naifenchen52@gmail.com

Scope. This supplement contains *aggregate* evidence only. It includes no patient image, no label, no prediction, no case identifier, no split membership and no per-case metric. It supports the main paper; it does not replace its core evidence.

1 Two Distinct Statistics: The Architecture Rank R and the Audit Utility U_τ

The paper uses two scalar summaries that must not be conflated, and neither is the challenge's own ranking procedure.

Architecture screen, R (fractional rank). The fold-0 ResEnc-M versus ResEnc-L screen ranks the two candidates component by component and averages those ranks, so R is a mean fractional rank over the six components and *lower R is better*. It is a *rank* statistic: it discards the size of each difference and retains only its direction. Two different tie conventions apply here and must not be confused. *Within R*, an exact component-level metric tie assigns both models the average of the tied positions, so a tied component contributes the same amount to each model's rank and can favour neither. *Separately*, the frozen selection rule resolves a tied or unmet advancement criterion in favour of ResEnc-M, the cheaper baseline plan; that rule governs the decision, not the rank arithmetic. The screen was applied once, to fold 0, under DSC/HD95.

The advancement rule in full. ResEnc-L could advance from fold 0 to fold-1 confirmation only if the executable rank-gain threshold on $R(\mathrm{M}) - R(\mathrm{L})$ was met — nominally $1/6$, one component of six, stored as the constant `MEANINGFUL_RANK_GAIN` of the committed policy — the paired-bootstrap 95% percentile interval for $\Delta R = R(\mathrm{L}) - R(\mathrm{M})$ lay entirely below zero, and every frozen auxiliary gate was supplied and passed, a gate with no input supplied counting as a failure. Note the two directions: the rank gain $R(\mathrm{M}) - R(\mathrm{L})$ is positive when ResEnc-L is ahead, whereas ΔR is negative when ResEnc-L is ahead. Passing on fold 0 did not select or prefer ResEnc-L; it only triggered fold-1 confirmation, and expansion required the same rule to hold on fold 1. In every other case ResEnc-M was retained.

What the executable comparison actually is. The gate is the binary64 comparison `rank_gain >= MEANINGFUL_RANK_GAIN`, not an exact real-number test of $R(\mathrm{M}) - R(\mathrm{L}) \geq 1/6$, and the two differ at the boundary. The stored threshold is the double nearest 1/6, 0.16666666666666666; a configuration that is exactly 1/6 in rational arithmetic — ResEnc-L ahead on three components, tied on one, behind on two — is computed as a difference of averaged ranks and evaluates to about 0.16666666666666652, some 1.4×10^{-16} short, so that nominal boundary case is rejected. We disclose this because the manuscript quotes the nominal value. It did not affect the recorded architecture decision, whose fold-0 rank gain was -0.333 against a required $+1/6$ and which failed the other conditions too. We did not modify the frozen policy code, and no result changed.

Audit C utility, U_τ (raw metric mean). Let S_τ be the subjects for which all six DSC/NSD components are finite under *both* policies, and $\bar{m}_{c,\tau}(P; S_\tau)$ the mean of component c over S_τ. Then

$$U_\tau(P; S_\tau) = \frac{1}{6} \sum_c \bar{m}_{c,\tau}(P; S_\tau), \qquad \Delta U_\tau = U_\tau(P_{\mathrm{cand}}; S_\tau) - U_\tau(P_{\mathrm{base}}; S_\tau),$$

where c ranges over the six DSC/NSD components, ET/TC/WT $\times$ DSC/NSD$_\tau$. U_τ retains the magnitude of every difference and is *not a rank statistic*: no ranking or tie-breaking enters it. R and U_τ are different objects and are never combined. Because both policies share S_τ, ΔU_τ is identically the arithmetic mean of the six component deltas; Table S1 verifies this against the committed record for every analysis reported.

Tolerance. The organizers confirmed that the final challenge ranking uses DSC and NSD, excludes HD95, and computes final-ranking NSD at $\tau = 1$. Throughout, $\tau = 1$ is the official-ranking-aligned analysis and $\tau = 0.5$ — Panoptica's default, and what our earlier runs used — is a prespecified sensitivity analysis reported at the reviewers' request. It does not carry equal official standing.

Table S1. The utility identity, re-derived from the committed component evidence. For every analysis reported, the recorded ΔU_τ equals the arithmetic mean of its six component deltas. A fractional-rank definition would instead yield exactly +1.000 in all six rows, since the candidate improved all six components everywhere; it does not describe these data.

Subset	τ	Recorded ΔU_τ	Mean of six deltas	|difference|
Development (folds 0–2)	1.0	+0.002532324	+0.002532324	$< 10^{-12}$
Development (folds 0–2)	0.5	+0.004167779	+0.004167779	$< 10^{-12}$
Policy-selection holdout (folds 3–4)	1.0	+0.000725887	+0.000725887	$< 10^{-12}$
Policy-selection holdout (folds 3–4)	0.5	+0.002473393	+0.002473393	$< 10^{-12}$
Pooled, supportive only (all folds)	1.0	+0.001809212	+0.001809212	$< 10^{-12}$
Pooled, supportive only (all folds)	0.5	+0.003489520	+0.003489520	$< 10^{-12}$

2 Per-Component Means and Deltas, Audit C (M8 Versus C0)

Each table below carries its own bootstrap summary and per-fold values inside the same float, so no cross-referencing between paragraphs is required. "Fraction of paired bootstrap resamples with positive ΔU_τ" is exactly that – a resampling frequency, not a probability and not a posterior.

Table S2. Development (folds 0–2), $\tau = 1.0$ (official-ranking-aligned). Common subject support $|S_\tau| = 806$. DSC does not depend on τ, so the DSC rows repeat between tolerances.

Component	Baseline C0	Candidate M8	Δ
ET DSC	0.862666	0.864668	+0.002002
TC DSC	0.916734	0.918590	+0.001856
WT DSC	0.927779	0.929696	+0.001918
ET NSD	0.893911	0.896031	+0.002120
TC NSD	0.876474	0.880153	+0.003678
WT NSD	0.865187	0.868807	+0.003620
U_τ / ΔU_τ	0.890458	0.892991	+0.002532

Fraction of paired bootstrap resamples with positive ΔU_τ: 1.0. 95% percentile interval $[+0.001284, +0.004548]$. Per-fold ΔU_τ: fold 0 +0.001235, fold 1 +0.003874, fold 2 +0.002497.

Table S3. Development (folds 0–2), $\tau = 0.5$ (sensitivity analysis). Common subject support $|S_\tau| = 806$. DSC does not depend on τ, so the DSC rows repeat between tolerances.

Component	Baseline C0	Candidate M8	Δ
ET DSC	0.862666	0.864668	+0.002002
TC DSC	0.916734	0.918590	+0.001856
WT DSC	0.927779	0.929696	+0.001918
ET NSD	0.650468	0.655845	+0.005377
TC NSD	0.616659	0.623215	+0.006556
WT NSD	0.568016	0.575313	+0.007297
U_τ / ΔU_τ	0.757054	0.761221	+0.004168

Bootstrap: *not present in the committed record for this tolerance.* No value is computed after the fact, and none is implied.

Table S4. Policy-selection holdout (folds 3–4), $\tau = 1.0$ (official-ranking-aligned). Common subject support $|S_\tau| = 538$. DSC does not depend on τ, so the DSC rows repeat between tolerances.

Component	Baseline C0	Candidate M8	Δ
ET DSC	0.863684	0.864145	+0.000461
TC DSC	0.912545	0.913423	+0.000878
WT DSC	0.925596	0.925685	+0.000089
ET NSD	0.890061	0.890349	+0.000288
TC NSD	0.872497	0.873684	+0.001187
WT NSD	0.860108	0.861561	+0.001452
U_τ / ΔU_τ	0.887415	0.888141	+0.000726

Fraction of paired bootstrap resamples with positive ΔU_τ: 0.7342. 95% percentile interval $[-0.002531, +0.002633]$. Per-fold ΔU_τ: fold 3 +0.002323, fold 4 −0.000883.

Table S5. Policy-selection holdout (folds 3–4), $\tau = 0.5$ (sensitivity analysis). Common subject support $|S_\tau| = 538$. DSC does not depend on τ, so the DSC rows repeat between tolerances.

Component	Baseline C0	Candidate M8	Δ
ET DSC	0.863684	0.864145	+0.000461
TC DSC	0.912545	0.913423	+0.000878
WT DSC	0.925596	0.925685	+0.000089
ET NSD	0.644482	0.647926	+0.003444
TC NSD	0.608462	0.612769	+0.004307
WT NSD	0.562636	0.568297	+0.005662
U_τ / ΔU_τ	0.752901	0.755374	+0.002473

Fraction of paired bootstrap resamples with positive ΔU_τ: 0.9609. 95% percentile interval $[-0.000313, +0.004167]$. Per-fold ΔU_τ: fold 3 +0.003882, fold 4 +0.001055.

Table S6. Pooled, supportive only (all folds), $\tau = 1.0$ (official-ranking-aligned). Common subject support $|S_\tau| = 1344$. DSC does not depend on τ, so the DSC rows repeat between tolerances.

Component	Baseline C0	Candidate M8	Δ
ET DSC	0.863074	0.864459	+0.001385
TC DSC	0.915057	0.916522	+0.001465
WT DSC	0.926905	0.928091	+0.001186
ET NSD	0.892370	0.893757	+0.001387
TC NSD	0.874882	0.877563	+0.002681
WT NSD	0.863154	0.865906	+0.002752
U_τ / ΔU_τ	0.889240	0.891050	+0.001809

Fraction of paired bootstrap resamples with positive ΔU_τ: 0.9875. 95% percentile interval $[+0.000256, +0.003361]$. Per-fold ΔU_τ: fold 0 +0.001235, fold 1 +0.003874, fold 2 +0.002497, fold 3 +0.002323, fold 4 −0.000883.

Table S7. Pooled, supportive only (all folds), $\tau = 0.5$ (sensitivity analysis). Common subject support $|S_\tau| = 1344$. DSC does not depend on τ, so the DSC rows repeat between tolerances.

Component	Baseline C0	Candidate M8	Δ
ET DSC	0.863074	0.864459	+0.001385
TC DSC	0.915057	0.916522	+0.001465
WT DSC	0.926905	0.928091	+0.001186
ET NSD	0.648072	0.652675	+0.004603
TC NSD	0.613378	0.619034	+0.005656
WT NSD	0.565862	0.572505	+0.006643
U_τ / ΔU_τ	0.755391	0.758881	+0.003490

Fraction of paired bootstrap resamples with positive ΔU_τ: 1.0. 95% percentile interval $[+0.002199, +0.004709]$. Per-fold ΔU_τ: fold 0 +0.002973, fold 1 +0.004721, fold 2 +0.004818, fold 3 +0.003882, fold 4 +0.001055.

3 Complete Decision-Check Matrices

Check names are the *legacy machine field names* recorded by the audit code, reproduced verbatim so the tables can be matched against the published JSON. They are identifiers, not claims: in particular `bootstrap_probability` names a gate on the *fraction of paired bootstrap resamples with a positive difference*, which is a resampling frequency and not a probability. Each check is reported exactly as the committed record holds it. The holdout matrix contains three checks whose only purpose is to make the audit fail closed — exact membership, zero evaluator errors and independent recomputation agreement — so that a missing, errored or membership-mismatched candidate evaluation is ineligible rather than silently compared on whatever subset survived.

Table S8. Development subset (folds 0–2): 18 checks — 17 of 18 passed.

Check (legacy machine field name)	Outcome
`dU_common_tau1`	pass
`dU_common_tau05`	pass
`official_deltas_positive`	pass
`all_components_nonnegative_tau1`	pass
`all_components_nonnegative_tau05`	pass
`bootstrap_probability`	pass
`bootstrap_lower_bound_above_zero`	pass
`enough_folds_nonnegative`	pass
`no_fold_collapses`	pass
`et_zero_dsc_not_increased`	pass
`total_zero_dsc_not_increased`	pass
`unique_subject_zero_dsc_not_increased`	pass
`lesion_fn_not_increased`	**FAIL**
`lesion_fp_within_limit`	pass
`no_denominator_only_gain`	pass
`exact_membership`	pass
`zero_evaluator_errors`	pass
`independent_recompute_agrees`	pass

4 Lesion-Level Evidence and Its Limits

Components are 26-connected. A reference component counts as missed if no predicted component of the same region overlaps it; the denominator therefore depends only on the reference and is policy-invariant, which the audit verified.

The counters below are matched components, not voxels, and they live in two different spaces. The first two are counted over *predicted* components, the third over *reference* components. The committed record labels the prediction-space

Table S9. Policy-selection holdout (folds 3–4): 23 checks — 15 of 23 passed.

Check (legacy machine field name)	Outcome
dU_common_tau1	**FAIL**
dU_common_tau05	pass
bootstrap_probability	**FAIL**
bootstrap_lower_bound_above_zero	**FAIL**
components_nonnegative_tau1	pass
components_nonnegative_tau05	pass
official_deltas_positive	pass
dsc_within_limit	pass
nsd_within_limit	pass
fold3_nonnegative	pass
fold4_nonnegative	**FAIL**
et_zero_dsc_not_increased	pass
total_zero_dsc_not_increased	**FAIL**
unique_subject_zero_dsc_not_increased	pass
n_ref_total_invariant	pass
miss_rate_point_within_margin	**FAIL**
miss_rate_upper_bound_within_margin	**FAIL**
no_region_miss_rate_regression	**FAIL**
fp_pred_within_limit	pass
no_denominator_only_gain	pass
exact_membership	pass
zero_evaluator_errors	pass
independent_recompute_agrees	pass

Table S10. Pooled, supportive only: 18 checks — 14 of 18 passed.

Check (legacy machine field name)	Outcome
dU_common_tau1	pass
dU_common_tau05	pass
bootstrap_probability	**FAIL**
bootstrap_lower_bound_above_zero	pass
components_nonnegative_tau1	pass
components_nonnegative_tau05	pass
et_zero_dsc_not_increased	pass
total_zero_dsc_not_increased	pass
unique_subject_zero_dsc_not_increased	pass
n_ref_total_invariant	pass
miss_rate_point_within_margin	**FAIL**
miss_rate_upper_bound_within_margin	**FAIL**
no_region_miss_rate_regression	**FAIL**
fp_pred_within_limit	pass
no_denominator_only_gain	pass
exact_membership	pass
zero_evaluator_errors	pass
independent_recompute_agrees	pass

overlap counter diagnostic only, and this supplement keeps that label: it is reported for transparency and never used to override an official metric or a safety gate. Prediction-space overlaps and reference-space misses are *not* complementary under component matching — one predicted component may overlap several reference components and one reference component may be covered by several predicted components — so their sum need not equal the reference-component total, and here it does not. The safety gate is computed as $\mathrm{FN}_{\mathrm{ref}}/N_{\mathrm{ref}}$, the missed fraction of reference components; it is not the legacy recall derived from the prediction-space overlap counter.

Table S11. Component-level counts and rates under 26-connectivity. Development counts are from the development subset; the non-inferiority analysis and its margins apply to the policy-selection holdout. The first row of each block is a prediction-space diagnostic counter and is not a reference-space true-positive count.

Quantity	Baseline C0	Candidate M8
Development subset		
Predicted components overlapping a reference	3990	4000
Predicted components with no reference overlap	1243	1087
Reference components with no predicted overlap	4103	4116
Policy-selection holdout		
Predicted components overlapping a reference	2698	2640
Predicted components with no reference overlap	864	781
Reference components with no predicted overlap	2726	2764
Reference components in total	5445	5445
Missed fraction of reference components	0.500643	0.507622

On the holdout the miss rate rose by +0.006979 against a margin of 0.0025, with a one-sided 95% upper bound of +0.012408 against 0.005. Per region the increases were ET +0.006286, TC +0.005848 and WT +0.008649, each above the per-region limit of 0.005. Predicted false-positive components fell, so the 0.05 relative-increase cap was not breached. The bootstrap used 540 subjects, seed 20260730 and 10000 resamples.

Limits of this aggregate interpretation. These are counts over a whole subset. They establish the *direction* of the change — fewer predicted false positives, more missed reference components — and nothing about which lesions changed status. We did not stratify by lesion size, location, region or cohort, and we did not test any mechanism, so the precision–sensitivity reading offered in the main paper is a hypothesis rather than a demonstrated cause. The absolute miss rate is high under this definition because every reference component counts equally regardless of size. The margins are *operational*: they were fixed in the protocol to be strict relative to observed between-fold variation, the committed protocol records no clinical derivation for them, and they are not externally validated clinical thresholds.

5 Architecture Diagnostic Under the Ranking Metric

The fold-0 ResEnc-M versus ResEnc-L screen was decided under DSC/HD95, on the values in Table S12. Because the challenge ranks DSC and NSD, the same fold-0 predictions were rescored under DSC/NSD at $\tau = 0.5$ as a diagnostic.

Table S12. The fold-0 screen as decided, under DSC and HD95. Means are over all 271 fold-0 cases.

Model	ET DSC	TC DSC	WT DSC	ET HD95	TC HD95	WT HD95
ResEnc-M (selected)	0.859	**0.914**	**0.934**	14.17	**5.96**	**3.89**
ResEnc-L	**0.861**	0.912	0.932	**11.19**	6.08	4.11

Why the two tables differ slightly in ET DSC. Table S12 averages over all 271 fold-0 cases, in which a case with an empty reference *and* an empty prediction scores 1. The rescoring below follows the official parser's skip-missing convention, which drops those cases from the denominator: one for ResEnc-M and two for ResEnc-L. Excluding them reproduces 0.8581 and 0.8602 exactly. Both figures are correct under their own stated convention, and no other component is affected at the precision reported.

Table S13. Fold-0 architecture diagnostic under the ranked component pair, $\tau = 0.5$. Bold marks the better value in each column.

Model	ET DSC	TC DSC	WT DSC	ET NSD	TC NSD	WT NSD
ResEnc-M (selected)	0.8581	**0.9140**	**0.9336**	0.6565	0.6213	0.5759
ResEnc-L	**0.8602**	0.9117	0.9322	**0.6679**	**0.6244**	**0.5762**

ResEnc-L is better at the point estimate on four of six DSC/NSD components, so the ordering *reverses* relative to the DSC/HD95 screen. On the rank statistic itself, with $\Delta R = R(\mathrm{L}) - R(\mathrm{M})$ so that negative values favour ResEnc-L, the observed value is $\Delta R = R(\mathrm{L}) - R(\mathrm{M}) = -0.333$ with a 95% percentile interval of $[-1.000, +0.667]$, which includes zero: no rank advantage is established, and the declared robustness requirement was not met. Under the frozen tie rule an unmet advancement criterion retains ResEnc-M.

Chronology, for the record. The fold-0 screen completed on 2026-07-23 and the organizers' clarification that the final ranking uses DSC/NSD at $\tau = 1$ and excludes HD95 was received on 2026-07-28, both dated in the committed decision log. That ordering is recorded as fact and is *not* offered as justification: the deployed model was still selected under a metric pair the challenge does not rank on, and that remains a limitation of this submission.

Limitations. The diagnostic is fold-0 only and was computed at $\tau = 0.5$ only; it was never repeated at the official-ranking tolerance $\tau = 1$. It scores each architecture's own end-of-training validation predictions, not the release path, so it is not comparable to the Audit C rows. The WT NSD difference is 0.0003, at the edge of what these data resolve. Fold-0 ResEnc-L weights already existed and are what the diagnostic scores; after the diagnostic no five-fold ResEnc-L expansion, architecture change or retraining campaign was undertaken. The architecture–metric mismatch is therefore recorded as an open limitation, not a resolved question.

6 Per-Component Evidence for the Checkpoint and Weight-Averaging Audits

Audit A screened the best-validation checkpoint, mirroring over the recorded axes, both combined, and enhancing-tumour connected-component cleanup at three predeclared volume thresholds. Audit B screened two predeclared per-fold weight averages of the final and best-validation checkpoints. The main paper reports their outcomes; the per-candidate components are here.

What these numbers are, exactly. All rows are the development subset ($n = 811$ subjects, folds 0–2), scored by BraTS-evaluation 0.0.8 GoAT. The policy-selection holdout was never opened for these audits, so no folds-3–4 value exists for any of them. Two measurement families appear below and are kept apart because they aggregate differently:

- the *decision* metrics — DSC and HD95 over all 811 subjects — which are the pair these audits were actually ranked on, and
- the *official* metrics — DSC and NSD — re-scored afterwards from the same preserved predictions once the organizers confirmed the ranking pair. These use the evaluator's skipna rule, so each policy carries its *own* denominators.

Short codes used in the tables. C1 is the best-validation checkpoint; C2 is mirroring over the recorded axes; C3 is C1 and C2 combined; ET_{10}, ET_{25} and ET_{50} are enhancing-tumour connected-component cleanup at the three predeclared volume thresholds, in mm^3; S1 and S2 are the per-fold weight averages $0.75\times$final $+ 0.25\times$best and $0.50\times$final $+ 0.50\times$best; C0 is the released baseline. Because the two families use different aggregation rules they disagree in the third decimal, and they are never combined into one column. The official-metric rows are *not* on the paired common support used for Audit C, so they are not line-comparable with the Audit C tables above. Every value here comes from nnU-Net's end-of-training validation lineage, which applies eightfold mirroring — not from the release path. $\tau = 1$ is the official-ranking-aligned tolerance; $\tau = 0.5$ is Panoptica's default and is reported as a prespecified sensitivity analysis only. Rendered values are rounded to four decimal places; the machine-readable record in `evidence/supplement_inputs.json` carries full float precision, and

its `provenance` block maps every value in this section back to the frozen record and key path it was read from, together with an explicit list of quantities that were never recorded and are therefore not shown.

Table S14. Audit A/B candidates, official-metric component means at $\tau = 1$ (official-ranking-aligned), development subset, $n = 811$. Denominators are per policy and are given in Table S18.

Policy	ET DSC	TC DSC	WT DSC	ET NSD	TC NSD	WT NSD
C0	0.8646	0.9172	0.9282	0.8960	0.8783	0.8664
C1	0.8620	0.9163	0.9265	0.8937	0.8757	0.8629
C2	0.8647	0.9172	0.9282	0.8960	0.8784	0.8664
C3	0.8641	0.9181	0.9280	0.8955	0.8789	0.8658
ET_{10}	0.8656	0.9172	0.9282	0.8966	0.8783	0.8664
ET_{25}	0.8665	0.9172	0.9282	0.8966	0.8783	0.8664
ET_{50}	0.8677	0.9172	0.9282	0.8971	0.8783	0.8664
S1	0.8632	0.9162	0.9266	0.8942	0.8755	0.8633
S2	0.8637	0.9169	0.9271	0.8951	0.8768	0.8643

Table S15. Audit A/B candidates, official-metric component means at $\tau = 0.5$ (sensitivity analysis), development subset, $n = 811$. Denominators are per policy and are given in Table S18.

Policy	ET DSC	TC DSC	WT DSC	ET NSD	TC NSD	WT NSD
C0	0.8646	0.9172	0.9282	0.6558	0.6215	0.5734
C1	0.8620	0.9163	0.9265	0.6491	0.6157	0.5659
C2	0.8647	0.9172	0.9282	0.6558	0.6215	0.5734
C3	0.8641	0.9181	0.9280	0.6549	0.6229	0.5725
ET_{10}	0.8656	0.9172	0.9282	0.6565	0.6215	0.5734
ET_{25}	0.8665	0.9172	0.9282	0.6570	0.6215	0.5734
ET_{50}	0.8677	0.9172	0.9282	0.6576	0.6215	0.5734
S1	0.8632	0.9162	0.9266	0.6518	0.6169	0.5677
S2	0.8637	0.9169	0.9271	0.6524	0.6180	0.5686

Outcome. No candidate advanced at either tolerance: `G75_RETAIN_BASELINE` for Audit A, `G76_RETAIN_C0` for Audit B, and the later official-metric re-scorings returned `None` and `G79V_RETAIN_C0_TAU1_SENSITIVITY`. The list of advancing candidates is empty and the release policy was not changed. Denominators differ between policies because enhancing-tumour cleanup removes small predicted components and can leave a subject with no finite ET score; that is precisely why an ET-only mean gain is not evidence of improvement, and why the audits' common-support rule exists.

Table S16. Candidate minus C0, official-metric component deltas. A positive value favours the candidate. Development subset, $n = 811$.

Policy	τ	ΔET DSC	ΔTC DSC	ΔWT DSC	ΔET NSD	ΔTC NSD	ΔWT NSD
C1	1	−0.0027	−0.0009	−0.0017	−0.0023	−0.0026	−0.0035
	0.5	−0.0027	−0.0009	−0.0017	−0.0067	−0.0057	−0.0075
C2	1	+0.0000	+0.0000	+0.0000	+0.0000	+0.0000	+0.0000
	0.5	+0.0000	+0.0000	+0.0000	+0.0000	+0.0000	+0.0000
C3	1	−0.0005	+0.0009	−0.0002	−0.0005	+0.0006	−0.0006
	0.5	−0.0005	+0.0009	−0.0002	−0.0009	+0.0014	−0.0009
ET_{10}	1	+0.0010	+0.0000	+0.0000	+0.0006	+0.0000	+0.0000
	0.5	+0.0010	+0.0000	+0.0000	+0.0007	+0.0000	+0.0000
ET_{25}	1	+0.0018	+0.0000	+0.0000	+0.0006	+0.0000	+0.0000
	0.5	+0.0018	+0.0000	+0.0000	+0.0012	+0.0000	+0.0000
ET_{50}	1	+0.0030	+0.0000	+0.0000	+0.0011	+0.0000	+0.0000
	0.5	+0.0030	+0.0000	+0.0000	+0.0018	+0.0000	+0.0000
S1	1	−0.0014	−0.0011	−0.0016	−0.0018	−0.0028	−0.0031
	0.5	−0.0014	−0.0011	−0.0016	−0.0040	−0.0045	−0.0057
S2	1	−0.0009	−0.0003	−0.0011	−0.0009	−0.0016	−0.0021
	0.5	−0.0009	−0.0003	−0.0011	−0.0034	−0.0035	−0.0048

Table S17. The metrics these audits actually decided on: DSC and HD95 over all $n = 811$ development subjects. HD95 is in millimetres and lower is better. This is a different aggregation from the official-metric tables above, so the DSC columns differ slightly.

Policy	ET DSC	TC DSC	WT DSC	ET HD95	TC HD95	WT HD95
C0	0.8655	0.9174	0.9282	13.26	5.73	5.86
C1	0.8627	0.9165	0.9265	13.79	5.86	5.87
C2	0.8655	0.9174	0.9282	13.26	5.73	5.86
C3	0.8648	0.9183	0.9280	13.65	5.92	5.61
ET_{10}	0.8666	0.9174	0.9282	13.74	5.73	5.86
ET_{25}	0.8678	0.9174	0.9282	13.71	5.73	5.86
ET_{50}	0.8693	0.9174	0.9282	13.75	5.73	5.86
S1	0.8641	0.9164	0.9266	13.72	6.26	6.49
S2	0.8646	0.9171	0.9271	13.42	5.96	5.94

Table S18. Per-policy denominators (finite scored subjects per component), the frozen advancement statistic against C0 at $\tau = 1$, and the advancement outcome. **Sign convention:** the rank gain is $R(\text{C0}) - R(\text{candidate})$, so a *positive* value favours the candidate. The interval is the paired subject-level percentile interval for that same quantity, in the *same* orientation as the point estimate. The frozen record stores the interval in the opposite orientation, as $R(\text{candidate}) - R(\text{C0})$; it is negated here (which reverses its endpoints) and retained unmodified in `evidence/supplement_inputs.json` under `bootstrap_delta_ci_candidate_minus_C0`. No candidate advanced.

Policy	n ET	n TC	n WT	Rank gain	95% interval	Advances
C0	806	809	811	—	—	—
C1	807	809	811	-1.000	$[-1.000, -0.333]$	no
C2	806	809	811	$+1.000$	$[-0.667, +1.000]$	no
C3	807	809	811	-0.333	$[-1.000, +1.000]$	no
ET_{10}	805	809	811	$+0.333$	$[-0.333, +0.333]$	no
ET_{25}	803	809	811	$+0.333$	$[-0.333, +0.333]$	no
ET_{50}	801	809	811	$+0.333$	$[-0.333, +0.333]$	no
S1	806	809	811	-1.000	$[-1.000, -0.667]$	no
S2	806	809	811	-1.000	$[-1.000, -0.333]$	no

On the two weight soups. Both lose to C0 on every component at both tolerances, and on all three HD95 diagnostics. The ordering between them is C0 > S2 > S1 throughout: the soup that mixes *more* of the best-validation checkpoint (0.50) is uniformly *better* than the one that mixes less (0.25). The degradation is therefore consistent in direction but *not* monotone in the amount mixed in, and the main paper says so.

The bounded 40-epoch fine-tuning screen. Three fine-tuning recipes were screened for 40 epochs at a learning rate of 0.0005 (seed 20260730) on folds 0–2, on a common support of $n = 806$ subjects, each initialised from a bit-identical copy of the corresponding C0 checkpoint. The screen tested exactly one point in the fine-tuning design space – 40 epochs at 5% of the original learning rate, from a converged model, on three folds. A bounded run of that size cannot speak to convergence, and it does not establish that fine-tuning cannot help: a longer schedule, a higher learning rate or a different intervention was never tested, because the preregistration made the 40-epoch screen the gate for spending more compute and the screen did not pass. The frozen record stores *deltas* against C0 only: absolute per-component means for the candidate arms were never computed and are not reconstructed here, and HD95 was excluded by the preregistration, so no HD95 value exists for any arm. Bootstrap, per-fold and tail statistics exist at $\tau = 1$ only.

At $\tau = 1$ the aggregate utility change was -0.000442 (T), $+0.000087$ (DG) and $+0.000237$ (TDG); at $\tau = 0.5$ it was -0.000746, -0.000262 and -0.000189. The $\tau = 1$ paired bootstrap reproduces those point estimates, with positive-resample fractions 0.0714, 0.5969 and 0.8192; every interval spans zero. All three arms failed the $\tau = 0.5$ utility gate and the total zero-DSC gate, so the terminal status

Table S19. Bounded 40-epoch fine-tuning screen: component deltas against C0 on the common support ($n = 806$). Positive favours the candidate. Aggregate utility changes are given below. The screen selected on $\min(\Delta U_{\tau=1}, \Delta U_{\tau=0.5})$; no arm was eligible.

Arm	τ	ΔET DSC	ΔTC DSC	ΔWT DSC	ΔET NSD	ΔTC NSD	ΔWT NSD
T	1	−0.00031	−0.00007	−0.00015	−0.00054	−0.00068	−0.00090
	0.5	−0.00031	−0.00007	−0.00015	−0.00076	−0.00088	−0.00231
DG	1	+0.00030	+0.00023	−0.00002	+0.00021	+0.00018	−0.00037
	0.5	+0.00030	+0.00023	−0.00002	−0.00027	−0.00015	−0.00165
TDG	1	+0.00026	+0.00036	+0.00018	+0.00024	+0.00034	+0.00004
	0.5	+0.00026	+0.00036	+0.00018	+0.00015	−0.00048	−0.00160

was `G82_RETAIN_C0_NO_SCREEN_ADVANCEMENT` and C0 was retained. Of the ten frozen eligibility gates, T passed 5/10, DG passed 7/10 and TDG passed 8/10.

The dense-overlap variant D25. D25 differed from C0 in exactly one parameter, the sliding-window tile step ($0.5 \rightarrow 0.25$). It was never run. The prerequisite baseline-reproduction gate failed because the twelve required baseline values are reproducible only by the eightfold mirroring lineage, while the specification defines C0 with mirroring disabled and prohibits test-time augmentation, so the targets cannot be reproduced by any run that complies with the specification. No D25 prediction, DSC, NSD, runtime or utility value exists. This is a specification/lineage mismatch, not evidence that D25 failed; D25 remains scientifically unresolved. An analytic tile-count study was nevertheless computed, without generating any prediction. It found that the step change multiplies tiles per volume by 2.25, 1.33 and 1.5 across three representative geometries.

7 Bounded Experiment Inventory

The search was bounded, and is listed rather than implied to be exhaustive. The list mixes candidates that were executed and scored with proposals that were screened out *before* execution; the outcome column says which is which.

Table S20: Complete bounded experiment inventory. Every candidate the main paper refers to appears here. The status column separates candidates that were executed and scored from proposals that were screened out before any execution. Nothing here was adopted; the released policy is unchanged.

Item	Status	Outcome
C1 best-validation checkpoint	Executed	Scored on the development subset. No robust improvement; the rank gain against C0 is negative. Stopped.
C2 mirroring over the recorded axes	Executed	Point-positive on the decision metrics but the interval included zero; superseded by Audit C.

Item	Status	Outcome
C3 best checkpoint + mirroring	Executed	The combined arm. No robust improvement; stopped on the development subset.
ET connected-component cleanup, $10\,\mathrm{mm}^3$	Executed	Raised ET DSC but changed ET support, trading overlap against boundary distance. Stopped.
ET connected-component cleanup, $25\,\mathrm{mm}^3$	Executed	As above, with fewer scored ET subjects. Stopped.
ET connected-component cleanup, $50\,\mathrm{mm}^3$	Executed	As above; the apparent mean gain comes from the changed ET denominator. Stopped.
S1 weight soup, 0.75 final + 0.25 best	Executed	All six components moved the wrong way at both tolerances, and it also failed the HD95 tail gate.
S2 weight soup, 0.50 final + 0.50 best	Executed	All six components moved the wrong way at both tolerances. Uniformly better than S1, so the degradation is not monotone in the amount mixed in.
M8 eightfold mirroring TTA	Executed	Improved all six components on the development subset but failed the frozen lesion false-negative veto and stopped; a separate folds-3–4 follow-up then failed the utility and miss-rate gates. C0 retained.
T tail-aware sampling	Executed (bounded screen)	40-epoch screen. Failed 5 of 10 gates, including both utility gates; ineligible.
DG appearance transform	Executed (bounded screen)	40-epoch screen. Failed 3 of 10 gates, including the $\tau = 0.5$ utility gate; ineligible.
TDG (T + DG)	Executed (bounded screen)	40-epoch screen. Failed 2 of 10 gates, including the $\tau = 0.5$ utility gate; ineligible.
D25 dense overlap, tile step 0.25	**Never executed**	Preregistered and considered, then blocked before any run by an invalid baseline specification: the required baseline values are reproducible only under the mirroring lineage, while the specification defines C0 with mirroring disabled and prohibits test-time augmentation. Zero D25 predictions were generated and no D25 result of any kind exists. This is not evidence that D25 fails.

8 Runtime, Container, and Repeatability

Full-cohort run. The exact release policy processed 451 of 451 official validation cases with exit code 0 and zero inference errors, at a peak framework-reserved VRAM of 2.48 GiB and a wall time of 3 h 35 m (28.6 s per case) on a single NVIDIA A40. Both figures describe that A40 execution. The 2.48 GiB is memory reserved by the tensor framework, not total device memory, and is neither an A10G measurement nor a statement of A10G headroom.

Repeatability. Across two independent 24-case runs, 18 of 2.14×10^8 output voxels differed. The runner is near-deterministic but not bit-exact, because deterministic-algorithm enforcement runs in warn-only mode. Every output remained label-valid, hierarchy-consistent and geometry-exact. This variation is orders of magnitude smaller than the release-path gap discussed in the main paper.

Container contract. Linux/AMD64, zero network, all dependencies and all five fold checkpoints baked in at build time. `/input` is read-only, one folder per case; `/output` is flat, one `.nii.gz` per case. Each output file name is the complete validated basename of its input case folder — no truncation to a fixed-width identifier, no hardcoded cohort prefix. Labels are restricted to $\{0, 1, 2, 3\}$ and geometry is preserved exactly. Modality discovery fails closed before any output is written. The runner refuses to start without five distinct fold checkpoints. The official contract supplies a fresh writable `/output`; no claim is made that the runner refuses to overwrite an existing file.

Superseded A10G qualification. An *earlier, superseded* pre-correction image was exercised on one NVIDIA A10G using synthetic four-modality volumes. That image was executed by the organizers on the hidden test set and failed before writing any prediction, because its runner required the input case-folder basename to end in exactly five digits. The synthetic fixtures used in that qualification had fixed-width folder names and therefore never exercised the variable-length condition that caused the failure. **The corrected image that was finally submitted has never been measured on an A10G.** No organizer execution log, no hidden-test result and no rank exists for it, and none is claimed. The historical record is retained in the public repository under a filename that marks it as superseded.

9 What This Supplement Does Not Contain

No patient image, segmentation label, model prediction, model weight, case identifier, fold membership or per-case metric appears here or in the public repository. No submission identifier, image tag or digest, registry path, cloud resource identifier or credential appears here. Every quantity above is an aggregate over a named subset, generated directly from the committed audit records rather than transcribed by hand.