SAME DICE, DIFFERENT MISTAKES: LABEL COARSENING RELOCATES SEGMENTATION ERRORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Segmentation metrics such as Dice and IoU report how much a model gets wrong, but not what it gets wrong: they discard the identity of the pixels that receive false positives. That is harmless only if false positives are interchangeable; they are not. A multiple sclerosis (MS) lesion detector that flags a benign white-matter hyperintensity makes a different, clinically consequential mistake than one that flags ordinary tissue; a car detector that fires on a truck differs from one that fires on road. Training pipelines erase this distinction by collapsing every non-target structure into one background class. Prior work has shown that keeping such structures separate can raise Dice. What has not been shown is that coarsening changes the composition of a model's errors even when their quantity and Dice are held fixed: when coarse- and fine-labeled models make the same number of mistakes, do they make the same kinds of mistakes? We show they do not, and name the phenomenon semantic error relocation (SER). We introduce Semantic False-Positive Concentration (SFPC), the fraction of target false positives that fall on a pre-specified hidden negative subclass, and a protocol that compares models only after matching their total false-positive burden on validation data, then under an exact equal-error-budget audit in which any SFPC difference is a relocation of fixed error mass. On MS3SEG, in a pre-registered confirmation across three 3D architectures and three seeds, binary supervision places 20.96% of its false positives on normal white-matter hyperintensities versus 13.81% under four-class supervision (+7.15 percentage points, 95% CI 6.04–8.31, 39/40 patients), while total false positives are indistinguishable (1280 vs. 1283 voxels) and binary Dice is slightly higher (0.626 vs. 0.617). This is about one additional benign region falsely flagged per patient, a 266$\times$ enrichment over the subclass's area. An independent Cityscapes replication reproduces the effect: +7.24 points more car false positives land on trucks and buses (95% CI 5.09–9.59) at statistically identical Dice (0.865 vs. 0.867). Controlled follow-ups rule out prediction-head size as the cause and trace the shift to the training objective the taxonomy induces: reweighting the hidden class alone reproduces part of the effect, with any further contribution from explicit class identity left open. The practical conclusion is direct: whenever fine labels exist at evaluation time, even if deployment needs only a binary mask, reporting SFPC alongside overlap metrics exposes a difference in what errors mean that Dice cannot see.

1 INTRODUCTION

A segmentation score usually answers one question: *how much did the model get wrong?* Dice, intersection over union (IoU), precision, and sensitivity summarize overlap or error counts, but they discard the identity of the negative pixels receiving false-positive (FP) predictions. This is benign only if all false positives are interchangeable. They often are not. A multiple sclerosis (MS) lesion false positive on a non-target/normal white-matter hyperintensity (nWMH) is semantically different from one on ordinary background; a car prediction on a truck or bus is different from one on road or sky.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Training pipelines nevertheless collapse such structures routinely. Binary lesion segmentation maps every non-lesion structure to "background." A target-only Cityscapes task can similarly merge buses, trucks, vegetation, pedestrians, and road. This coarsening is convenient, but it changes the supervision supplied to a finite model. Recent work has shown that richer background labels can improve segmentation accuracy. Most directly, BackSplit demonstrates that subdividing heterogeneous biomedical background can improve lesion Dice and optimization (Saluja et al., 2026). Partial-label and overlapping-taxonomy methods likewise model the fact that observed labels can merge finer semantic states (Shi et al., 2021; Cour et al., 2011; Bevandić et al., 2022). Those results establish that label taxonomy can affect performance. We ask a different question:

> *If coarse and fine models make approximately the same amount of target error, do they make the same semantic mistakes?*

We find that the answer is no. Coarsening the negative taxonomy systematically reallocates residual target false positives toward a semantically related hidden subclass. We call this **semantic error relocation (SER)**. Our central measure, **Semantic False-Positive Concentration (SFPC)**, asks what fraction of all target false positives land on a pre-specified hidden negative subclass. Crucially, we do not compare SFPC at arbitrary thresholds. We first match operating points using validation data only, with total false-positive burden as the primary match. We then perform a stronger audit that gives the two models exactly the same per-example FP budget. Under that audit, a change in SFPC is literally a redistribution of fixed error mass.

Figure 1 previews the result. In the frozen MS3SEG confirmation panel, binary supervision produces slightly higher target Dice than fine supervision (0.6263 vs. 0.6172), yet allocates 20.96% rather than 13.81% of its false positives to nWMH. Mean total FP counts are essentially identical. In an independent Cityscapes experiment, car Dice is similar (0.8648 vs. 0.8665), while coarse supervision increases the fraction of car false positives landing on trucks/buses from 18.59% to 25.82%. The effect is nearly the same magnitude across these very different domains.

The hidden subclasses are not merely large regions that absorb errors by area. In the MS3SEG confirmation set, nWMH occupies only 0.079% of negative voxels on average, yet receives 20.96% of Binary false positives—about $266\times$ its area-only prevalence (Fine: $175\times$). In Cityscapes, trucks/buses occupy 2.60% of valid negative pixels on average but receive 25.82% of Coarse car false positives—about $9.94\times$ their area-only prevalence (Fine: $7.15\times$). We call these *area-only nulls*, not chance levels, because segmentation errors are not spatially uniform.

Our study is designed around the main alternative explanations. The MS3SEG comparison is paired on the *same held-out patient*, so anatomy, scanner/acquisition, lesion burden, and hidden-class prevalence cannot confound the supervision-regime contrast. Seeds are aggregated within patient before inference rather than treated as independent samples. Thresholds are selected only on validation patients and then frozen on test patients. We additionally stratify by nWMH and lesion burden, test three architecture families, and use three seeds. Conventional argmax metrics on folds 3–4 had been seen during development; however, the matched-error endpoint, matching rule, all-three-seed aggregation, exact-FP budgets, and statistical tests were frozen before any folds 3–4 matched-error endpoint result was inspected. Finally, we replicate the phenomenon independently on Cityscapes and add parameter-matched/reweighted controls that test whether the effect is merely a prediction-head parameter-count difference or partly a consequence of hidden-pixel loss weight.

Contributions.

1. We identify **semantic error relocation**: training-label coarsening can change the semantic destination of segmentation errors even when target overlap and total error burden are stable.
2. We formulate **SFPC** and a matched-error evaluation protocol, including an exact equal-FP-budget audit for which the SFPC gap has a direct fixed-error-mass interpretation.
3. We provide a **frozen endpoint-confirmation study** on three-dimensional (3D) MS magnetic resonance imaging (MRI) and an **independent cross-domain replication** on Cityscapes, spanning five architecture families and showing a similar approximately seven-percentage-point relocation effect.
4. We stress-test interpretation with **practical and mechanistic controls**: component-level clinical consequences, area-only enrichment baselines, parameter matching, and a hidden-negative

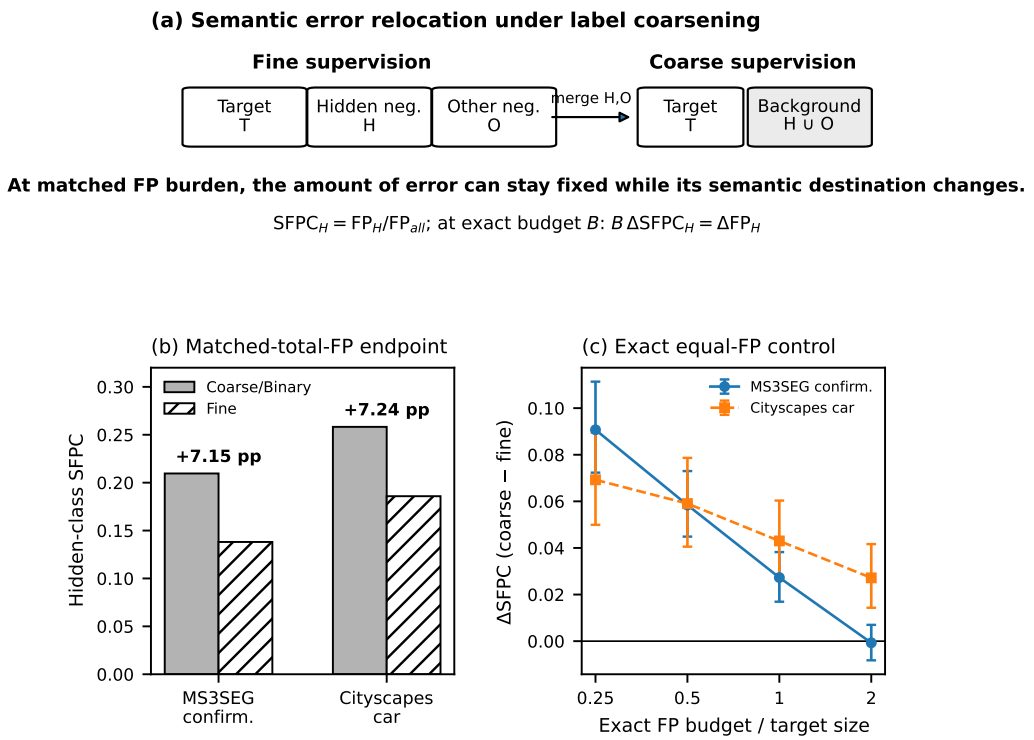


Figure 1: **Label coarsening relocates segmentation errors.** (a) Fine supervision distinguishes a target T, hidden semantic negative H, and other negatives O; coarse supervision merges $H \cup O$. SFPC measures the share of target false positives landing on H. Under an exact FP budget B, $B\Delta\text{SFPC}_H = \Delta FP_H$, so the SFPC gap is directly a relocation of fixed error mass. (b) Frozen MS3SEG confirmation and independent Cityscapes replication show approximately seven percentage points (pp) more hidden-class FP concentration under coarse supervision. (c) The gap is strongest at low/moderate error budgets; it vanishes at the largest MS3SEG budget but remains positive in Cityscapes. Error bars are paired-bootstrap 95% confidence intervals (CIs).

reweighting control showing that the phenomenon is not explained by the prediction-head parameter-count difference and is partly mediated by the training objective induced by the taxonomy.

2 Related Work and Scope of Novelty

Background subdivision and label granularity. BackSplit is the closest recent work (Saluja et al., 2026). It argues that biomedical "background" is heterogeneous, shows that auxiliary background subdivisions can improve lesion segmentation, and analyzes the gain through expected Fisher information. Our work does *not* claim that background heterogeneity or the benefit of finer labels is new. Instead, we condition on approximately or exactly matched error quantity and study the *composition of the remaining errors*. The primary Cityscapes result is deliberately chosen to make this distinction sharp: target Dice does not improve, yet semantic error composition changes substantially. Work on hierarchical segmentation (Li et al., 2022), label granularity (Cole et al., 2022), partial labels (Cour et al., 2011), marginal losses for merged labels (Shi et al., 2021), overlapping label spaces (Bevandić et al., 2022), and background shift in incremental segmentation (Cermelli et al., 2020) establishes complementary ways in which taxonomy matters.

Beyond aggregate overlap. Hidden stratification shows that aggregate metrics can conceal meaningful subgroup failures (Oakden-Rayner et al., 2020). Fine-grained segmentation-error analyses and the Critical Error Rate distinguish types or semantic severity of errors beyond IoU (Bernhard

et al., 2024; Zhou and Beyerer, 2023); Metrics Reloaded emphasizes that evaluation should match the domain question (Maier-Hein et al., 2024). Recent work on semantic label flips under correlation shift likewise shows that overlap can hide identity errors (Achara et al., 2026). That setting perturbs label–context correlation and studies foreground-identity swaps; ours instead holds the held-out cases fixed, intervenes on the *training taxonomy*, and measures how target false-positive mass is allocated among negative subclasses. Thus, we do not claim that semantic-aware error analysis in general is new; the novelty is the label-coarsening $\rightarrow$ error-relocation relationship and its matched-error identification.

3 Semantic Error Relocation

3.1 Setup and SFPC

Let the fine ground-truth label be $Z \in \{T, H, O\}$, where T is the deployment target, H is a pre-specified semantically related negative subclass, and O collects all remaining valid negatives. Fine supervision preserves this distinction. Coarse supervision applies

$$C(Z) = \begin{cases} T, & Z = T, \\ B, & Z \in \{H, O\}, \end{cases} \tag{1}$$

while the deployment output remains target vs. non-target in both regimes. For a target prediction $\hat{T}$ at a fixed operating point, define

$$\mathrm{SFPC}_H = \frac{FP_H}{FP_{\mathrm{all}}} = \frac{|\{i : Z_i = H,\ \hat{Z}_i = T\}|}{|\{i : Z_i \neq T,\ \hat{Z}_i = T\}|}. \tag{2}$$

SFPC is an error-composition quantity, not a replacement for Dice or false-positive rate (FPR). It is intentionally conditional on making a false-positive target prediction: *where did that error land?*

3.2 Why the effect is not algebraically inevitable

A simple population observation makes the empirical result less trivial. Coarsening only the *negative* labels leaves the Bayes target posterior unchanged:

$$P(C(Z) = T \mid X) = P(Z = T \mid X). \tag{3}$$

Thus an unrestricted, perfectly optimized predictor of the target marginal need not change merely because H and O are merged. SER is therefore not a definitional consequence of the label map.

At the same time, practical fine and coarse objectives are *not* identical: preserving H can change class competition, loss weights, and gradients on hidden-negative pixels. Our claim is therefore deliberately about the *supervision regime induced by label taxonomy*, not about semantic class names acting independently of the loss. A Cityscapes follow-up explicitly probes prediction-head parameter count, hidden-pixel cross-entropy (CE) weighting, and explicit hidden-class identity (Section 4.2). It shows that extra hidden-negative loss weight explains a measurable part of SER, while the additional role of explicit class identity is not statistically resolved. This narrows the mechanism without weakening the empirical phenomenon.

3.3 Matching error quantity before comparing error composition

Threshold choice is an immediate confound: a more aggressive model may have higher SFPC simply because it predicts more positives. Our frozen primary protocol selects thresholds *only on validation cases*. For each architecture/seed/fold, we first find each regime's Dice-optimal validation threshold on a 0.001 grid. We then set the target total-FP burden to the midpoint of the two regimes' validation FP counts and independently choose the coarse and fine thresholds whose validation FP counts best match that target. Ties are broken by higher validation Dice, then proximity to 0.5, then the lower threshold. These thresholds are frozen before test evaluation. Matched sensitivity, predicted volume, and Dice are pre-specified secondary operating-point controls.

We also use a stronger **exact equal-FP-budget audit**. For each held-out case, among ground-truth negative locations we rank the target posterior and select exactly B highest-scoring negatives, where

$B \in \{0.25, 0.5, 1, 2\}\times$ the ground-truth target size. This is an evaluation diagnostic, not a deployable threshold rule. Both regimes then have exactly B false positives by construction, and

$$B\left(\mathrm{SFPC}_H^C - \mathrm{SFPC}_H^F\right) = FP_H^C - FP_H^F. \tag{4}$$

The SFPC gap is therefore the fraction of a *fixed* error budget relocated onto H.

3.4 Statistical unit and testing

The inferential unit is the held-out patient (MS3SEG) or image (Cityscapes), never a seed, voxel, or model run. Seeds are averaged within architecture and case, then architectures are averaged within case for the primary test. We report paired mean differences, 10,000-resample paired bootstrap 95% CIs, and two-sided Wilcoxon signed-rank tests. Architecture-specific results are secondary. The primary endpoint is pre-specified and unadjusted; secondary families use Holm correction where reported.

4 Experiments

4.1 MS3SEG: controlled 3D medical experiment

MS3SEG contains 100 MS patients with axial T2-weighted fluid-attenuated inversion recovery (FLAIR) MRI and separate masks for ventricles, nWMH, and abnormal/MS-related white-matter hyperintensities (abWMH) (Bawil et al., 2026). The target is abWMH and the hidden negative subclass is nWMH. We compare Binary supervision (abWMH vs. all non-abWMH) with Fine supervision (background/ventricle/nWMH/abWMH). Images, patients, augmentation/sampling policy, optimization budget, and backbone are otherwise shared. The three backbones are 3D U-Net (Ronneberger et al., 2015), SegResNet (a residual segmentation network) (Myronenko, 2018), and SwinUNETR (a shifted-window transformer U-Net) (Hatamizadeh et al., 2022), trained from scratch with three seeds (90 core checkpoints total).

We use a five-fold patient-disjoint out-of-fold (OOF) design. Each fold has 64 training, 16 validation, and 20 test patients; no threshold or hyperparameter is selected on test. This paired hypothesis-testing protocol deliberately differs from the dataset's canonical 80/20 benchmark split and is not presented as a leaderboard estimate. FLAIR is canonicalized, robustly intensity-normalized, and resampled to $1 \times 1 \times 3$ mm. Native fine-mask overlaps are excluded from supervision/evaluation before a single nearest-neighbor resampling of the categorical label map. Training uses $96 \times 96 \times 32$ patches and a shared label-agnostic hyperintense hard-negative sampler. Full implementation details are in Appendix A.

Discovery and frozen endpoint confirmation. The relocation hypothesis emerged from exploratory analyses on folds 0–2. We therefore separate discovery from confirmation. A 60-patient folds 0–2/base-seed panel is labeled *discovery*. Conventional argmax metrics on folds 3–4 had also been inspected during development; however, the matched-error endpoint (SFPC at matched total FP), threshold-selection rule, seed/architecture aggregation, exact-FP budgets, burden-strata cut points, bootstrap, and paired test were frozen before any folds 3–4 matched-error endpoint result was inspected. The formal endpoint-confirmation panel uses folds 3–4, all three seeds, and all three architectures ($n = 40$ patients). The 100-patient, all-seed OOF analysis is reported as a robustness estimate, not as a second independent confirmation.

This paired design directly addresses patient-level confounding: each patient's scanner/acquisition, anatomy, abWMH burden, and nWMH prevalence are identical for its binary and fine predictions. We additionally stratify the frozen endpoint by discovery-derived nWMH and abWMH burden tertiles.

4.2 Cityscapes: independent cross-domain replication

Cityscapes provides fine pixel labels for urban scenes (Cordts et al., 2016). Before inspecting results we fixed **car** as the primary target and **truck+bus** as the hidden subclass. A person→rider contrast was also pre-specified as secondary and is reported in Appendix F. Coarse models predict target vs. all other valid classes. Fine models preserve other/target/hidden subclasses. The 2,975 official training images are split by city into 2,583 training and 392 internal-validation images; the 500

official validation images remain untouched until final testing. Eligibility (at least 32 target and 32 hidden-subclass pixels) is prediction-independent, yielding 137 primary car images.

We use DeepLabV3 with a 50-layer residual network (ResNet50) (Chen et al., 2017) and Lite Reduced Atrous Spatial Pyramid Pooling (LRASPP) with MobileNetV3 (Howard et al., 2019), ImageNet-initialized, with three seeds. Evaluation and inferential aggregation mirror MS3SEG. Appendix E documents all training details and the engineering-only DataLoader fix required by a final batch of size one; validation/test data and the frozen scientific configuration were unchanged.

Mechanistic stress test. To address the objection that the Fine result is simply a trivial consequence of a different loss, we ran a follow-up control on the pre-specified Cityscapes car task using both architectures and all three seeds. Parameter-matched Coarse (COARSE-PM) uses the Fine-sized four-output head but retains the binary target loss; H-WEIGHTED BINARY keeps truck/bus pixels as binary background while raising their CE weight from 0.25 to 1.0; H-UNION gives truck+bus an explicit competing class while keeping the same four-output capacity and the same nominal pixel weights. The primary diagnostic compares H-Weighted Binary with H-Union at an exact $1\times$ FP budget; matched-total-FP results are supporting. This follow-up was specified after the original cross-domain replication and is interpreted as mechanism evidence, not as a second confirmation of the primary phenomenon.

5 RESULTS

5.1 MS3SEG: COARSE LABELS RELOCATE ERRORS DESPITE SLIGHTLY HIGHER DICE

The discovery panel showed a large positive SFPC gap: 0.1920 for Binary vs. 0.1227 for Fine, $\Delta = +6.94$ pp (95% CI 5.87–8.06, $p = 3.13 \times 10^{-11}$; 58/60 patients). We then evaluated the frozen endpoint on folds 3–4.

The confirmation endpoint reproduces the effect almost exactly. Binary SFPC is 0.2096 versus 0.1381 for Fine, $\Delta = +7.15$ pp (95% CI 6.04–8.31, $p = 3.64 \times 10^{-12}$; paired rank-biserial 0.998; 39/40 patients positive). Crucially, mean total FP counts are 1280.3 vs. 1282.6 voxels ($\Delta = -2.3$, 95% CI $[-54.3, 52.8]$, $p = 0.536$). The result is therefore not explained by Binary simply making more false positives.

Binary target Dice is slightly higher: 0.6263 vs. 0.6172 ($\Delta = +0.0092$, 95% CI 0.0041–0.0144). Yet its errors are more concentrated on the clinically adjacent nWMH subclass. This difference has a concrete component-level consequence: at matched total FP, Binary predictions falsely detect **5.09 nWMH components per patient** versus 4.03 for Fine, an increase of **1.06 components/patient** (95% CI 0.76–1.39, $p = 3.79 \times 10^{-7}$; 35/40 patients positive). Among the 21 confirmatory patients eligible for the stringent large/non-periventricular nWMH sensitivity analysis, robust nWMH confounder false-positive rate (CFPR) is 35.95% vs. 20.40% ($\Delta = +15.55$ pp, $p = 0.0032$). A component is counted as falsely detected when at least 10% of its voxels are predicted as target; components smaller than five voxels are excluded. These quantities make the shift more tangible than a ratio alone.

The full five-fold/all-three-seed OOF robustness estimate is 0.1991 vs. 0.1307, $\Delta = +6.84$ pp (95% CI 6.14–7.55, $p = 5.43 \times 10^{-18}$; 97/100 patients). All five folds are positive (6.28–7.73 pp) and all three architectures are positive (6.56–7.35 pp). The gap also remains positive within every frozen nWMH-burden and abWMH-burden stratum.

5.2 EXACT FP BUDGETS ISOLATE COMPOSITION FROM QUANTITY

The exact-budget audit removes threshold-transfer mismatch entirely. In frozen MS3SEG confirmation, Binary−Fine SFPC is +9.07 pp at $0.25\times$ target size, +5.84 pp at $0.5\times$, and +2.73 pp at $1\times$, all with CIs excluding zero. At $2\times$, the gap is -0.07 pp (95% CI $[-0.82, 0.70]$, $p = 0.452$): relocation vanishes once the error budget becomes very large. Full OOF shows the same shape. The phenomenon is therefore strongest in the high-score/low-error tail rather than a universal ordering at arbitrarily permissive operating points.

Table 1: **Primary matched-total-FP results.** SFPC is the fraction of target false positives falling on the hidden semantic subclass. CIs and p values are paired at the patient/image level after seed and architecture aggregation. MS3 abbreviates MS3SEG. The 100-patient row is a robustness estimate because it includes discovery and confirmation folds.

Setting	Coarse SFPC	Fine SFPC	Δ pp	95% CI pp	p
MS3 discovery ($n = 60$)	.1920	.1227	+6.94	[5.87, 8.06]	3.1×10^{-11}
MS3 frozen confirm. ($n = 40$)	.2096	.1381	**+7.15**	**[6.04, 8.31]**	3.6×10^{-12}
MS3 full OOF ($n = 100$)	.1991	.1307	+6.84	[6.14, 7.55]	5.4×10^{-18}
City car→truck+bus ($n = 137$)	.2582	.1859	**+7.24**	**[5.09, 9.59]**	2.0×10^{-10}



Figure 2: **Where the errors go in a held-out Cityscapes scene.** Shown is the deterministic median-positive ΔSFPC example among eligible images for a fixed DeepLabV3/seed under the exact $1\times$ FP audit. Red denotes car ground truth (GT), gold denotes truck/bus, and false positives on other background are magenta. The Coarse model allocates visibly more of the fixed FP budget to truck/bus than the Fine model. Selection is based on SFPC, not visual appearance; Appendix H also shows a median non-positive example.

This budget dependence is scientifically useful. Fine supervision appears to move semantically adjacent negatives out of the most target-like portion of the score distribution. As the allowed error budget grows, easy "other" background increasingly enters both regimes and dilutes the compositional difference.

5.3 Cityscapes: same Dice, same phenomenon

The independent primary replication is deliberately cleaner with respect to target performance. At matched total FP, Coarse car SFPC is 0.2582 and Fine SFPC is 0.1859: $\Delta = +7.24$ pp (95% CI 5.09–9.59, $p = 2.03 \times 10^{-10}$). Both architectures independently agree (DeepLabV3 +6.45 pp; LRASPP +8.02 pp). Meanwhile car Dice is 0.8648 vs. 0.8665 ($\Delta = -0.0017$, $p = 0.554$), and aggregate total FP is 2294.7 vs. 2055.1 pixels ($p = 0.703$). Thus conventional target overlap views the systems as essentially tied while their error semantics differ markedly.

Validation-selected matching transfers imperfectly for one architecture: DeepLabV3 has 19.7% test FP mismatch (LRASPP 4.2%). The exact-budget audit is therefore the stronger control. It remains positive at *every* pre-specified budget: +6.92, +5.90, +4.30, and +2.72 pp at $0.25, 0.5, 1$, and $2\times$, respectively (all $p \leq 0.005$). Unlike MS3SEG, the Cityscapes primary effect persists even at the largest budget. Trucks/buses comprise only 2.60% of valid negative pixels on average, so Coarse and Fine SFPC correspond to $9.94\times$ and $7.15\times$ enrichment over this area-only baseline.

The pre-specified person→rider secondary contrast is also positive (+7.04 pp at matched total FP), but because Coarse person Dice is materially lower, we use it only as directional corroboration (Appendix F), not as evidence for the "same Dice" claim.

5.4 Mechanistic stress test: head-size difference does not explain SER; loss weighting explains part

At an exact $1\times$ FP budget, expanding Coarse to a Fine-sized four-output head has no detectable effect: Coarse SFPC is 0.2177 vs. 0.2219 for Coarse-PM ($\Delta = -0.42$ pp, 95% CI $[-1.18, 0.34]$, $p = 0.171$). Giving truck/bus pixels extra CE weight while keeping them in binary background reduces SFPC from 0.2219 to 0.2071 ($\Delta = +1.48$ pp, 95% CI 0.70–2.29, $p = 8.47 \times 10^{-6}$; Holm-

adjusted $p = 5.93 \times 10^{-5}$). At the validation-matched total-FP operating point, the corresponding reduction is +3.14 pp ($p = 1.34 \times 10^{-6}$; global Holm-adjusted $p = 1.07 \times 10^{-5}$).

Giving the hidden pixels an explicit competing class produces a lower *mean* SFPC than H-Weighted Binary (0.1853 vs. 0.2071 at exact $1\times$ FP), but the pre-specified paired test does not support a reliable additional class-identity effect ($p = 0.694$). The matched-total-FP comparison is nominally positive (+2.61 pp, $p = 0.0356$) but does not survive the global Holm correction ($p = 0.107$). We therefore draw a narrow conclusion: **SER is not explained by the prediction-head parameter-count difference, and the loss weighting induced by finer supervision accounts for a significant portion of the shift; the extra contribution of explicit class identity remains unresolved.** This directly answers the strongest alternative explanation without attributing a mechanism that the data do not establish.

5.5 Consistency without pseudoreplication

The effect is not a seed-count artifact. In MS3SEG, seeds are averaged inside each patient before primary inference, and 97/100 full-OOF patients have positive architecture-averaged differences. In Cityscapes, 96/137 (70%) primary images have positive differences. The latter misses our pre-specified 80% directional-consistency target, so SER should be interpreted as a strong *population-level* shift rather than a deterministic property of every scene. Both City architectures nevertheless show independently positive mean effects, and the exact-budget result remains positive at all four primary budgets.

6 Interpretation and Implications

SER is a property of supervision-induced error allocation. Equation 3 rules out a simplistic claim that merging negative labels must alter the population target posterior. The follow-up Cityscapes control, however, also shows why the finite-network result is not mysterious: the taxonomy changes the training objective. Parameter matching alone does not move SFPC, whereas increasing the loss weight on hidden negatives while retaining a binary target removes a significant portion of the gap. We therefore interpret SER as an empirical property of how a finite dense predictor allocates errors under the supervision induced by a label taxonomy. The current evidence does *not* establish that explicit semantic class identity, by itself, is the mechanism.

Why this is different from "fine labels improve Dice." BackSplit convincingly establishes a performance benefit from background subdivision in biomedical lesion segmentation (Saluja et al., 2026). SER is complementary: the primary object is *which negatives absorb residual errors after controlling how many errors occur*. The MS3SEG result is revealing because the coarse model has slightly higher Dice; the Cityscapes primary result is revealing because no statistically detectable Dice difference was found. In both cases, a roughly seven-point semantic shift remains hidden from target overlap.

Why the difference matters. The MS3SEG shift is not merely a change in a normalized ratio: at matched total error it corresponds to about one additional non-target nWMH component falsely detected per patient, and it persists under a stringent large/non-periventricular nWMH subset. More generally, different semantic destinations can have different downstream costs even when Dice is unchanged. The correct application-specific cost model will vary, but aggregate overlap alone cannot reveal this dimension.

Evaluation implication. If fine labels exist at evaluation time, even when deployment requires only a binary target mask, target false positives should be decomposed by semantic destination. SFPC is simple by design and should accompany, not replace, conventional metrics. When operating points differ, matched-FP or exact-budget auditing is necessary; otherwise one can mistake aggressiveness for semantic relocation. This follows the broader principle that evaluation metrics must reflect the failure mode of interest (Maier-Hein et al., 2024).

Annotation and objective implication. Auxiliary labels can matter even when they are not deployment outputs, but the Cityscapes mechanism control cautions against attributing the benefit

solely to "semantic understanding": part of the effect comes from where the loss places weight. This suggests a practical design question for future work—which high-cost negative subclasses should receive explicit or reweighted supervision—rather than assuming that exhaustive fine taxonomy is always necessary.

7 LIMITATIONS

Our evidence spans two datasets and five architecture families, but not every segmentation domain. MS3SEG contains 100 patients from one acquisition setting, and additional medical datasets with explicit benign/pathologic white-matter-hyperintensity distinctions would strengthen clinical generality. The MS3SEG dataset paper reports U-Net Dice of 0.7469 for binary lesion segmentation and 0.6686 for multi-class abWMH segmentation under its benchmark configuration (Bawil et al., 2026); our roughly 0.62 architecture-averaged values are lower and are *not* intended as leaderboard results. Our protocol instead fixes a paired 3D patch setting, $1 \times 1 \times 3$ mm resampling, a 100-epoch training budget, and five-fold OOF hypothesis testing to isolate supervision-regime effects; it is not the dataset's canonical leaderboard protocol. These absolute Dice values should therefore not be interpreted as the attainable performance ceiling on MS3SEG.

SFPC depends on a meaningful pre-specified hidden subclass and on its prevalence in the evaluation distribution; it is an error-composition estimand, not a prevalence-invariant intrinsic property. The area-only enrichment analysis helps contextualize this dependence but is not a literal random-error baseline. In MS3SEG, small/periventricular nWMH can be annotation-ambiguous; component-level size/distance sensitivity analyses preserve the conventional confounder-FP direction (Appendix D), but richer adjudication would be preferable. We did not model demographic effect modification.

In Cityscapes, one architecture misses the $< 10\%$ threshold-transfer mismatch criterion, and only 70% of eligible primary images have positive individual differences; we report both failures explicitly. Exact equal-FP budgets remove the former concern but not the latter. The mechanism decomposition is also limited: hidden-pixel reweighting explains part of SER, but the additional effect of explicit class identity is not statistically resolved. Finally, relocation is budget-dependent: the MS3SEG gap becomes null at the largest $2\times$ FP budget. We therefore claim a strong low-to-moderate-error phenomenon, not a universal ordering at every threshold. We also do not claim that semantic-aware error analysis or FP concentration ratios are unprecedented; prior work already decomposes segmentation errors (Zhou and Beyerer, 2023; Bernhard et al., 2024). Our contribution is the controlled link between *training-label coarsening* and *matched error composition*.

8 CONCLUSION

Label taxonomy is not merely bookkeeping. Across 3D MS MRI and natural-image street scenes, coarsening semantically distinct negatives moves a larger share of target false positives onto related hidden subclasses. This shift survives frozen endpoint confirmation, architecture/seed variation, matched operating points, and exact equal-FP controls, and it can coexist with slightly better Dice (MS3SEG) or similar Dice (Cityscapes). A follow-up control further shows that the effect is not explained by the small prediction-head parameter-count difference and is partly mediated by hidden-negative loss weighting. Dense predictors should therefore be audited not only for how many mistakes they make, but for where those mistakes go.

Reproducibility Statement

Appendices A–I provide the preprocessing/training settings, split logic, frozen endpoint definitions, statistical aggregation, robustness checks, and provenance notes. An anonymized supplementary archive accompanies the submission and contains the analysis/training code used for the reported experiments, split manifests, frozen configuration files, and the per-patient/per-image aggregate tables needed to recompute the manuscript statistics without model retraining. The final MS3SEG audit contains 90/90 usable core checkpoints; the original Cityscapes replication contains 24/24 usable checkpoints. Full configuration hashes are retained in the supplementary files rather than the main text. The archive distinguishes the original test-fold manifest from a deterministic reconstruction of training/validation membership.

Ethics Statement

MS3SEG is an existing de-identified public medical-imaging dataset, and Cityscapes is a public computer-vision benchmark. This work studies model evaluation and supervision rather than clinical deployment, and we make no diagnostic or treatment claims. Semantic-error auditing may expose clinically meaningful failure modes hidden by aggregate overlap, but it does not substitute for external, prospective, or subgroup-specific clinical validation.

AI Use Statement

AI tools were used to assist code development and debugging, literature retrieval, and manuscript drafting/editing. The authors reviewed the final manuscript and take responsibility for all claims and content.

References

Akshit Achara, Yovin Yathathugoda, Nick Byrne, Michela Antonelli, Esther Puyol-Antón, Alexander Hammers, and Andrew P. King. Right regions, wrong labels: Semantic label flips in segmentation under correlation shift. In *International Conference on Learning Representations (ICLR) Workshop on Catch, Adapt and Operate*, 2026.

Mahdi Bashiri Bawil, Mousa Shamsi, Aydin Ghalehasadi, Ali Fahmi Jafargholkhanloo, and Abolhassan Shakeri Bavil. A multiple sclerosis MRI dataset with tri-mask annotations for lesion segmentation. *Scientific Data*, 13:867, 2026.

Maximilian Bernhard, Roberto Amoroso, Yannic Kindermann, Lorenzo Baraldi, Rita Cucchiara, Volker Tresp, and Matthias Schubert. What's outside the intersection? Fine-grained error analysis for semantic segmentation beyond IoU. In *Winter Conference on Applications of Computer Vision (WACV)*, pp. 968–977, 2024.

Petra Bevandić, Marin Oršić, Ivan Grubišić, Josip Šarić, and Siniša Šegvić. Multi-domain semantic segmentation with overlapping labels. In *Winter Conference on Applications of Computer Vision (WACV)*, pp. 2615–2624, 2022.

Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulò, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9233–9242, 2020.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv:1706.05587*, 2017.

Elijah Cole, Kimberly Wilber, Grant Van Horn, Xuan Yang, Marco Fornoni, Pietro Perona, Serge Belongie, Andrew Howard, and Oisin Mac Aodha. On label granularity and object localization. In *European Conference on Computer Vision (ECCV)*, 2022.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban

scene understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, 2016.

Timothée Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 12:1501–1536, 2011.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *Brain Lesion (BrainLes) Workshop at the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2022.

Andrew Howard, Mark Sandler, Grace Chu, et al. Searching for MobileNetV3. In *International Conference on Computer Vision (ICCV)*, pp. 1314–1324, 2019.

Liulei Li, Tianfei Zhou, Wenguan Wang, Jianwu Li, and Yi Yang. Deep hierarchical semantic segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1246–1257, 2022.

Lena Maier-Hein, Annika Reinke, Patrick Godau, et al. Metrics reloaded: Recommendations for image analysis validation. *Nature Methods*, 21(2):195–212, 2024.

Andriy Myronenko. 3D MRI brain tumor segmentation using autoencoder regularization. In *Brain Lesion (BrainLes) Workshop at the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Association for Computing Machinery (ACM) Conference on Health, Inference, and Learning*, pp. 151–159, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pp. 234–241, 2015.

Rachit Saluja, Asli Cihangir, Ruining Deng, Johannes C. Paetzold, Fengbei Liu, and Mert R. Sabuncu. BackSplit: The importance of sub-dividing the background in biomedical lesion segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8492–8502, 2026.

Gonglei Shi, Li Xiao, Yang Chen, and S. Kevin Zhou. Marginal loss and exclusion loss for partially supervised multi-organ segmentation. *Medical Image Analysis*, 70:101979, 2021.

Jingxing Zhou and Jürgen Beyerer. Category differences matter: A broad analysis of inter-category error in semantic segmentation. In *CVPR Workshops*, pp. 3870–3880, 2023.

A Detailed MS3SEG Protocol

Data and preprocessing. We use the 100 preprocessed axial FLAIR volumes and tri-mask annotations supplied by MS3SEG (Bawil et al., 2026). Each volume is converted to closest canonical orientation. Intensities are centered by the nonzero-voxel median and scaled by the interquartile range (IQR) divided by 1.349, clipped to $[-5, 5]$, with background reset to zero. Volumes are resampled to $1 \times 1 \times 3$ mm. To avoid creating artificial fine-label overlap by independently resampling binary masks, source masks are first aligned to the native/canonical FLAIR grid; voxels with native fine-mask overlap are marked invalid, a single mutually exclusive categorical map is formed, and that map is resampled once with nearest-neighbor interpolation. Invalid voxels contribute neither training loss nor evaluation.

Splits and optimization. A frozen five-fold assignment gives 64 train, 16 validation, and 20 test patients per fold. Core seeds are 20270910, 20271011, and 20271112. Patches are $96 \times 96 \times 32$; 256 training patches are sampled per epoch. Shared sampling probabilities are 0.45 around abWMH, 0.30 around label-agnostic hyperintense non-lesion tissue, and the remainder elsewhere. This hard-negative sampler is shared by Binary and Fine models and does not use nWMH identity to privilege the Fine regime. Adam with decoupled weight decay (AdamW) uses learning rate 2×10^{-4}, weight decay 10^{-5}, gradient clipping at 12, and up to 100 epochs; validation occurs every 10 epochs. Batch size is 8 for U-Net and SegResNet and 4 for SwinUNETR. Inference uses the Medical Open Network for Artificial Intelligence (MONAI) sliding-window implementation with 0.25 overlap and brain floating-point 16-bit (BF16) precision on the A100 run. Binary training optimizes target/non-target Dice plus balanced negative log-likelihood (NLL); Fine training optimizes masked four-class cross-entropy plus foreground Dice. The backbone is otherwise identical for a given architecture; only the task-required output channels differ.

Nondeterminism. The original 3D training was speed-oriented (explicit seeds, Compute Unified Device Architecture (CUDA) Deep Neural Network library (cuDNN) benchmarking and non-bitwise-deterministic CUDA kernels). We therefore interpret seeds as repeated training realizations, not exact bitwise replicas. Seventeen checkpoint files lost during an earlier storage incident were retrained under their original run identities/configurations; the final audit verifies 90/90 usable core checkpoints. The statistical unit remains patient, so repeated seeds are aggregated rather than counted as independent evidence.

B Frozen Analysis Protocol and Statistical Details

The frozen MS3SEG matched-error configuration and full Secure Hash Algorithm 256-bit (SHA-256) method hash are provided in the supplementary files. The primary metric is nWMH SFPC at matched total FP. Candidate thresholds span 0.000–0.999 in steps of 0.001. Each regime first obtains a Dice-optimal validation threshold. For matched total FP, the target is the midpoint of the two regimes' mean validation FP counts at their own Dice-optimal thresholds. Candidate matched thresholds minimize absolute target error; ties favor higher validation Dice, then threshold closest to 0.5, then the lower threshold. Test labels never enter threshold selection.

The pre-specified secondary matches are sensitivity, predicted foreground volume, and Dice. Exact-FP budgets are 0.25, 0.5, 1.0, and 2.0 times ground-truth abWMH volume. Seeds are averaged within patient and architecture; primary inference averages architectures within patient. We use a two-sided Wilcoxon signed-rank test and a paired patient bootstrap of the mean difference with 10,000 resamples. Discovery-derived tertile cut points are 2318.7/3104.0 nWMH voxels and 1749.0/4617.7 abWMH voxels. The formal frozen endpoint is folds 3–4/all seeds/all architectures; full OOF is robustness because it includes discovery folds. For the secondary falsely detected component count, integer count differences summed across the nine runs preserve exact ties before the signed-rank test; the paired mean bootstrap uses seed 20270910.

Table 2: MS3SEG architecture and fold consistency for the full 100-patient, all-seed OOF robustness panel. Every architecture and every fold has positive Coarse−Fine SFPC.

Subset	ΔSFPC (pp)	Positive cases	Note
U-Net	+6.61	96/100	architecture-specific
SegResNet	+7.35	98/100	architecture-specific
SwinUNETR	+6.56	97/100	architecture-specific
Fold 0	+6.28	19/20	balanced 3 seeds
Fold 1	+6.72	20/20	balanced 3 seeds
Fold 2	+6.91	19/20	balanced 3 seeds
Fold 3	+6.58	20/20	balanced 3 seeds
Fold 4	+7.73	19/20	balanced 3 seeds

C Additional MS3SEG Robustness

At the frozen folds 3–4 confirmation endpoint, alternative operating-point matches all preserve a large positive SFPC difference: matched sensitivity +7.40 pp, matched predicted volume +7.31 pp, matched total FP +7.15 pp, matched Dice +6.48 pp, and individual Dice-optimal thresholds +6.77 pp. Thus the conclusion does not depend on one specific threshold-matching criterion.

The full-OOF matched-total-FP panel also shows conventional confounder metrics moving in the same direction: nWMH CFPR increases by 2.73 pp ($p = 4.6 \times 10^{-17}$), component-level confusion increases by 4.54 pp ($p = 2.4 \times 10^{-16}$), and the dimensionless nWMH FP enrichment ratio increases by 98.4 on average ($p = 6.5 \times 10^{-18}$). These are supporting analyses; SFPC remains primary because it directly conditions on total target FP burden.

Table 3: Exact equal-FP-budget SFPC gaps (Coarse/Binary minus Fine). The FP count is identical between regimes within each case by construction.

Dataset/contrast	Budget	ΔSFPC pp	95% CI pp	p
MS3 frozen confirm.	0.25×	+9.07	[7.23,11.14]	5.5×10^{-12}
	0.50×	+5.84	[4.49,7.30]	9.1×10^{-12}
	1.00×	+2.73	[1.69,3.82]	2.9×10^{-6}
	2.00×	−0.07	[−0.82,0.70]	.452
City car→truck+bus	0.25×	+6.92	[4.99,8.95]	7.6×10^{-8}
	0.50×	+5.90	[4.05,7.87]	1.7×10^{-5}
	1.00×	+4.30	[2.69,6.03]	7.7×10^{-4}
	2.00×	+2.72	[1.43,4.16]	.0050

D Burden Stratification and Label-Ambiguity Checks

Because Binary and Fine are evaluated on the same patient, burden is not a between-regime confound. Nevertheless, we tested effect modification using cut points frozen from discovery folds. In the confirmatory panel, Coarse−Fine SFPC remains positive in all nWMH-burden strata (low +6.17 pp, middle +6.46 pp, high +8.33 pp) and all abWMH-burden strata (low +7.54 pp, middle +7.68 pp, high +6.38 pp). Full OOF has the same qualitative pattern.

MS3SEG documents ambiguity in some small/periventricular hyperintensities. The analysis therefore also computes conventional confounder-FP sensitivity surfaces that progressively exclude nWMH components by equivalent axial diameter and distance from ventricles, with a stringent supporting cell of diameter ≥ 5 mm and ventricle distance ≥ 3 mm. In the full OOF panel, 43 patients remain eligible for this stringent analysis; robust CFPR is higher under Binary by 13.79 pp ($p = 5.5 \times 10^{-5}$) and robust component confusion by 20.59 pp ($p = 2.9 \times 10^{-5}$). These checks support, but do not replace, the primary all-annotated-nWMH SFPC estimand.

E Detailed Cityscapes Protocol

The frozen Cityscapes configuration and full SHA-256 method hash are provided in the supplementary files. For car→truck+bus, Coarse has target vs. all valid others; Fine has other/car/truck/bus. For person→rider, Fine has other/person/rider. The internal validation cities are Aachen, Jena, and Krefeld (392 images), leaving 2,583 official-train images for optimization; the 500 official validation images are final test. Evaluation resolution is 512×1024; training crops are 384×768 with target-focused cropping probability 0.7. Eligibility requires at least 32 target and 32 hidden-class pixels and is determined from ground truth before predictions.

Both architectures use ImageNet backbones, AdamW (learning rate 3×10^{-4}, weight decay 10^{-4}), batch size 2, 16-bit floating-point (FP16) precision, a maximum of 20 epochs, validation every two epochs, and early stopping after four non-improving validation rounds. Target Dice loss weight is 1 and background CE weight is 0.25. The local run used two graphics processing units (GPUs) and distributed independent runs across them without changing per-run batch size or optimization. One engineering fix set `drop_last=True` for the training loader because the odd-sized 2,583-image split otherwise produced a batch of one that crashes DeepLabV3's training-mode batch normalization (BatchNorm). This omits one randomly shuffled training image per epoch; validation and test loaders are unchanged, and no scientific hyperparameter or frozen analysis rule changed.

The original cross-domain experiment has 24/24 usable checkpoints across the two pre-specified pairs, two architectures, two regimes, and three seeds. For the primary car task, DeepLabV3 has 39,633,986 trainable parameters in Coarse and 39,634,500 in Fine; LRASPP has 3,218,308 vs. 3,218,648. The mechanism follow-up therefore uses a four-output Coarse-PM head with exactly the same parameter count as H-Weighted Binary, H-Union, and Fine (39,634,500 for DeepLabV3; 3,218,648 for LRASPP), showing that head-size matching alone does not reproduce the shift.

F Cityscapes Secondary Results and Pre-specified Checklist

The person→rider secondary pair gives matched-total-FP SFPC 0.2324 vs. 0.1620, $\Delta = +7.04$ pp (95% CI 5.40–8.72, $p = 9.98 \times 10^{-16}$). DeepLabV3 and LRASPP are both positive (+7.51 and +6.56 pp). Exact budgets are +6.98, +5.16, +3.36, and +2.00 pp from 0.25–2×, with the 2× point not significant ($p = .161$). Because Coarse person Dice is lower (0.5153 vs. 0.5609), this is only a directional replication of relocation, not a same-performance control.

The Cityscapes robustness checklist passed six of eight pre-specified criteria. It passed: positive primary mean, bootstrap CI excluding zero, paired $p < .05$, positive direction in both architectures, positive exact-budget effect at every primary budget, and positive secondary direction. It failed: at least 80% of primary images positive (observed 96/137 = 70%) and $< 10\%$ realized matched-FP mismatch in every primary architecture (DeepLabV3 19.7%; LRASPP 4.2%). Both failures are stated in the main paper; the exact equal-FP audit directly removes the second concern.

G Cityscapes Mechanism-Control Details

Table 4 summarizes the follow-up control sequence. Coarse-PM matches Fine's four-output head but uses only its first two logits for the binary target objective. H-Weighted Binary keeps truck/bus pixels in binary background but gives target and hidden pixels unit CE weight while ordinary background retains weight 0.25. H-Union assigns truck+bus to an explicit third class with the same four-output parameterization; Fine additionally separates truck and bus. All controls retain the same target Dice term.

Table 4: Cityscapes mechanism controls for car→truck+bus. Positive Δ means the first regime has higher SFPC. CIs are paired-bootstrap CIs for the mean difference; p values are two-sided Wilcoxon signed-rank tests. The pre-specified H-Weighted→H-Union exact-1× contrast does not support a reliable class-identity effect.

Endpoint	Contrast	Δ SFPC pp	95% CI pp	p
Exact 1×	Coarse − Coarse-PM	−0.42	[−1.18,0.34]	.171
Exact 1×	Coarse-PM − H-Weighted	+1.48	[0.70,2.29]	8.5×10^{-6}
Exact 1×	H-Weighted − H-Union	+2.18	[0.80,3.69]	.694
Exact 1×	H-Union − Fine	+1.06	[0.22,1.92]	.0206
Matched FP	Coarse-PM − H-Weighted	+3.14	[1.85,4.47]	1.3×10^{-6}
Matched FP	H-Weighted − H-Union	+2.61	[0.83,4.50]	.0356

Holm correction over non-primary mechanism tests leaves Coarse-PM−H-Weighted significant (global adjusted $p = 5.9 \times 10^{-5}$ exact; 1.1×10^{-5} matched), but not H-Weighted−H-Union at matched FP (adjusted $p = .107$) or H-Union−Fine at exact 1× (adjusted $p = .082$). The H-Weighted−H-Union exact-1× comparison is the pre-specified primary mechanism contrast and is therefore unadjusted. Its bootstrap CI for the paired *mean* excludes zero while the Wilcoxon test is non-significant because the signed-rank distribution is not consistently positive (52 positive, 68 negative, 17 ties); these statistics target different aspects of the paired distribution.

H Qualitative Error Localization

Figure 3 visualizes the original MS3SEG Binary/Fine comparison using a fixed U-Net/seed in the confirmatory folds. The displayed patient is selected algorithmically as the median positive SFPC-difference case among eligible patients, rather than by visual inspection. Figure 4 shows a Cityscapes counterexample selected as the median non-positive case under the same fixed architecture/seed used for Figure 2; this illustrates that SER is a population shift, not a deterministic per-image rule.

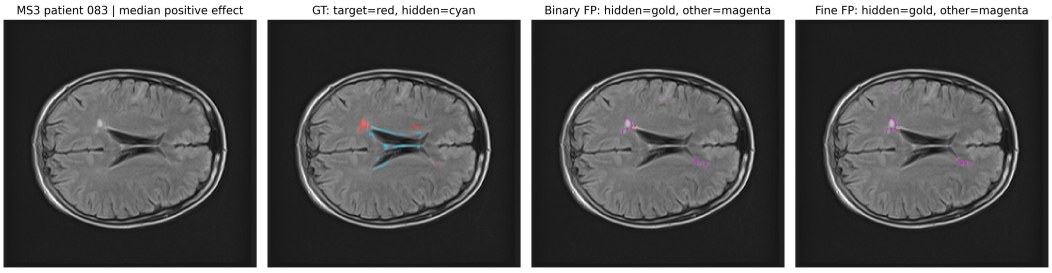


Figure 3: **MS3SEG qualitative localization from the original Binary/Fine experiment.** Red denotes abWMH target, cyan nWMH hidden negatives, gold false positives on nWMH, and magenta false positives on other background. The patient was selected by a deterministic median-positive SFPC criterion within a fixed model/seed, not by visual appearance.

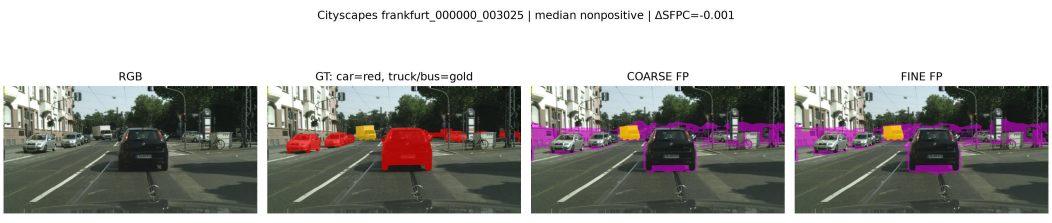


Figure 4: **Cityscapes non-positive counterexample.** This deterministically selected median non-positive image shows essentially no Coarse–Fine SFPC difference, illustrating the reported 70% rather than 100% per-image directional consistency.

I Checkpoint and Analysis Provenance

The final MS3SEG checkpoint audit contains 90/90 usable core runs: 5 folds $\times$ 3 architectures $\times$ 3 seeds $\times$ 2 supervision regimes. Earlier storage exhaustion caused 17 checkpoint files to be missing despite complete training histories. Those exact run identities were retrained using the original configuration and fold assignments, while 73 existing usable checkpoints were protected. Because the original speed-oriented CUDA setup was not bitwise deterministic, recovered checkpoints are replacement realizations under the same seeded configuration rather than byte-identical reconstructions.

For confirmation provenance, conventional argmax metrics on folds 3–4 had been observed during development. The matched-error endpoint (SFPC at matched total FP), matching procedure, exact-FP budgets, all-three-seed aggregation, burden strata, bootstrap, and paired test were then frozen before any folds 3–4 matched-error endpoint result was inspected. During checkpoint recovery, partial supporting panels later became visible, but the frozen formal endpoint and analysis rules were not changed. The final analysis reuses the exact preprocessing voxel counts and frozen fold assignments and verifies the stored method hash before execution.

The original Cityscapes package contains 24/24 usable checkpoints and the frozen split/configuration files. The follow-up mechanism-control package verifies the original split and reproduces the original Coarse/Fine result before evaluating new controls. All primary statistics reported in the manuscript are recomputable from the supplied aggregate per-patient/per-image comma-separated-value (CSV) files; neither model checkpoints nor test-time threshold retuning are required to reproduce the tables.