SAME IOU, DIFFERENT DECISIONS: SYNTHETIC MASK CORRUPTIONS MISLEAD CLASSIFIERS GUIDED BY SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Segmentation masks guide downstream classifiers through masking, cropping, auxiliary channels, and attention. Synthetic corruptions of ground-truth masks are a convenient substitute for repeatedly training segmenters, but their use assumes that matching intersection over union (IoU) also matches downstream effects. We test this assumption with a source-crossed audit that trains and evaluates classifiers on both synthetic masks and held-out segmenter predictions, while matching the two sources to the same IoU target for each image. On frozen PlantSeg test data, the two training sources respond differently to the same change in evaluation mask, yielding a $+6.28\,\mathrm{pp}$ difference-in-differences (95% CI over six seeds $[5.31, 7.26]$). A preregistered BUSI replication gives $+16.50\,\mathrm{pp}$ ($[13.93, 19.07]$), whereas a CUB extension using six seeds gives a smaller $+2.07\,\mathrm{pp}$ interaction ($[1.53, 2.62]$) in the opposite practical direction: synthetic evaluation fails to reveal a $+1.99\,\mathrm{pp}$ benefit from ROI cropping with predicted masks. On PlantSeg, synthetic validation consequently selects a masking method that incurs $+1.91\,\mathrm{pp}$ deployment regret ($[0.48, 3.34]$). Matching fragmentation $350\times$ more closely does not remove the discrepancy, and training masks obtained by cross fitting only partially mitigate it. Thus matching IoU for each image does not make synthetic and predicted masks interchangeable downstream. Their fidelity must be tested with a source-crossed downstream audit rather than inferred from mask similarity.

1 INTRODUCTION

Classification guided by segmentation is widely used when recognition depends on a localized structure. A predicted mask can crop a region of interest, suppress background, enter the classifier as an additional channel, or guide spatial attention. These mechanisms can improve classification under limited data or distribution shift (Mahbod et al., 2020; Rizhko et al., 2024). This creates a cascade in which errors in intermediate masks can propagate downstream, particularly when the classifier treats the mask as a deterministic input (Mehta et al., 2022).

Existing studies have examined this dependency through two complementary approaches. Models that jointly perform segmentation and classification demonstrate that segmentation cues can improve recognition (Xie et al., 2020; Rizhko et al., 2024). Studies using controlled perturbations instead erode, dilate, or randomly corrupt ground-truth masks, allowing researchers to vary segmentation quality without repeatedly training new segmenters (Zwanenburg et al., 2019; Bruhns et al., 2025). These approaches establish the value of controlled mask perturbations, but do not determine whether a synthetic mask matched in overlap faithfully substitutes for a deployed segmenter's prediction.

This limitation matters because a classifier consumes the spatial structure of a mask, not its IoU score. Masks with identical IoU can place false positives in different regions, fragment boundaries differently, or retain different amounts of content useful for class discrimination (Bernhard et al., 2024). An adaptive classifier may also exploit recurring patterns specific to its training inputs, including artifacts introduced by a synthetic corruption process (Geirhos et al., 2020). Our key insight is that mask fidelity depends on the interaction among mask source, training provenance, and downstream consumer, not on the mask alone. A synthetic proxy is faithful only if replacing deployment predictions preserves downstream behavior.

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.

Based on this insight, we introduce a *source-crossed downstream audit*. For each image, we construct a synthetic mask anchored to GT that matches the IoU of the corresponding held-out segmenter prediction. We then train one classifier on synthetic masks and another on predicted masks representative of deployment, and evaluate both classifiers on both mask sources. The resulting design with four cells separates the evaluation effect of changing the mask source from the classifier's adaptation to its training source. Their difference-in-differences measures the interaction between training and evaluation provenance. We also test model selection effects and whether OOF masks reduce provenance mismatch, following the principle of sample splitting in stacked generalization (Wolpert, 1992).

The audit reveals a lack of downstream interchangeability across several domains, but the direction of the discrepancy is not universal. On frozen PlantSeg test data, the source interaction is $+6.28$ pp (95% CI over six seeds $[5.31, 7.26]$). A preregistered BUSI replication yields $+16.50$ pp ($[13.93, 19.07]$) (Wei et al., 2026; Al-Dhabyani et al., 2020). In contrast, a CUB experiment with six seeds produces a smaller $+2.07$ pp interaction ($[1.53, 2.62]$) in the opposite practical direction. Synthetic evaluation fails to reveal a $+1.99$ pp benefit from ROI cropping with predicted masks, while the prespecified RIM-ONE DL primary analysis remains inconclusive (Wah et al., 2011; Batista et al., 2020). On PlantSeg, synthetic validation also selects a mask consumer that incurs $+1.91$ pp of deployment regret.

Our contributions are:

1. We formulate a source-crossed audit with four cells that evaluates synthetic mask fidelity through downstream behavior, distinguishing a mask's immediate evaluation effect from adaptation to its training source.
2. We establish a lack of downstream interchangeability across plant disease, breast ultrasound, and fine-grained bird recognition. Its direction varies across domains and mask consumers, showing that synthetic masks matched in IoU can either inflate or suppress the apparent value of segmentation guidance.
3. We connect proxy mismatch to model selection consequences and structural mechanisms. It persists under stricter IoU matching, alternative segmenters, metrics balanced across classes, and closer fragmentation matching; masks obtained by cross fitting provide only partial mitigation whose magnitude depends on architecture.

2 Related Work

Classification with segmentation guidance. Classifiers with segmentation guidance incorporate spatial information through masked images, ROI crops, additional input channels, attention, or jointly optimized segmentation and classification networks. Prior studies show that such guidance can improve recognition, particularly in settings with limited data, but its benefit depends on how the classifier consumes the mask (Mahbod et al., 2020; Rizhko et al., 2024). Recent systems apply these designs to ultrasound, chest radiography, dermoscopy, plant disease, and MRI (Cho et al., 2025; Kang et al., 2025; Wang et al., 2023; Yu & Son, 2020; Zhang et al., 2026). Coupled models further allow the two tasks to provide mutual supervision (Xie et al., 2020). This literature establishes that masks can be useful inputs, but it primarily optimizes classification performance under a chosen mask source. Our work instead asks whether the conclusion reached with synthetic masks remains valid when the classifier receives predictions from an actual segmenter.

Synthetic perturbations as reliability proxies. Controlled perturbations provide an efficient way to study model sensitivity without repeatedly collecting rare failures or fitting new upstream models. Erosion, dilation, and stochastic ROI perturbations have therefore been used to evaluate the robustness of radiomic features and other downstream measurements (Zwanenburg et al., 2019; Bruhns et al., 2025). At the input image level, however, evidence on synthetic proxies is mixed. Natural shift robustness may transfer poorly from synthetic corruptions (Taori et al., 2020), whereas recent work reports that carefully generated synthetic data can reproduce model level reliability rankings for semantic segmentation (Loiseau et al., 2024; Agnihotri et al., 2025). These studies evaluate synthetic images or fixed pretrained models. We inherit their controlled testing motivation but study a different setting in which the synthetic object is an intermediate mask and the downstream classifier adapts to

that mask during training. The unresolved question is whether matching IoU for each image preserves this adaptive downstream response.

Mask quality, provenance, and evaluation validity. Overlap metrics summarize segmentation accuracy but do not identify where errors occur or how they alter object extent, boundaries, and connected components. Recent fine-grained analyses show that models with similar IoU can exhibit substantially different error structures, which may matter to downstream applications (Bernhard et al., 2024). Related studies also find that manual and automatic masks need not be interchangeable for downstream analysis, and that deterministic intermediate predictions can propagate error through cascaded systems (Mahbod et al., 2020; Mehta et al., 2022). A second concern is provenance. Out-of-fold prediction and sample splitting prevent a downstream learner from consuming over-optimistic in-sample outputs (Wolpert, 1992), while learning with privileged information distinguishes supervision available during training from inputs available at deployment (Vapnik & Vashist, 2009). These ideas motivate honest intermediate predictions, but they do not directly test whether two mask sources support the same scientific or model selection conclusion. Our central distinction from all three lines of work is that we treat fidelity as a property of the complete interaction between mask and classifier and test it by crossing four combinations of training and evaluation provenance, rather than inferring it from mask similarity alone.

3 Method: Source-Crossed Downstream Audit

3.1 Overview and Design Intuition

A proxy matched in IoU is useful only if conclusions obtained with it survive replacement by masks from the deployed segmenter. This motivates an audit of the complete interaction between mask and classifier rather than the mask in isolation. As summarized in Figure 1, the audit (i) constructs a synthetic mask at the IoU of each predicted mask, (ii) trains classifiers separately on the two mask sources, (iii) crosses both trained classifiers with both sources at evaluation, and (iv) reports source interaction and decision consequences. The crossing is essential because a classifier can adapt to recurring spatial errors in its training masks even when their overlap scores are controlled.

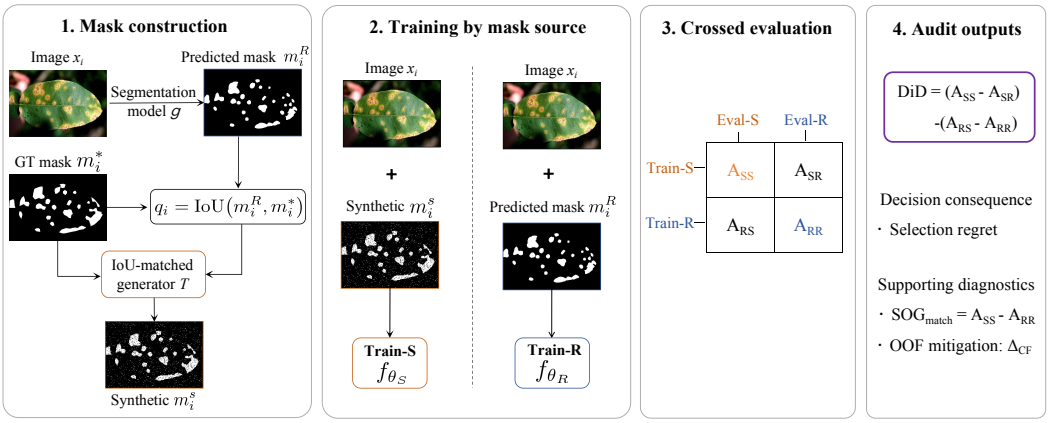


Figure 1: Overview of the source-crossed downstream audit. For each image, the synthetic mask matches the IoU target of the corresponding predicted mask but may retain a different spatial error structure. S denotes synthetic masks and R denotes predicted masks. The first index of $A_{\tau\nu}$ is the training source and the second is the evaluation source. The crossed cells isolate the interaction between mask provenance and classifier adaptation.

3.2 Problem Formulation and Source Construction

Let $\mathcal{D} = \{(x_i, m_i^*, y_i)\}_{i=1}^n$ contain an image, its ground-truth (GT) region mask, and its class label. A segmenter g produces a predicted mask $m_i^R = g(x_i)$, where R denotes provenance derived from a

segmenter rather than manual annotation. Its overlap for each image defines the synthetic target

$$q_i = \mathrm{IoU}(m_i^R, m_i^*), \qquad m_i^S = T(m_i^*, q_i; \omega_i), \tag{1}$$

where T is an erosion, dilation, or probabilistic mixing generator and ω_i is its sampled error field. Thus m_i^S targets q_i for the same image rather than matching only a mean over the dataset. This controls the overlap quality distribution while deliberately leaving error location, topology, boundary shape, and false-positive anchoring unconstrained. The absolute target residual has train median 0.0093 and 95th percentile 0.0218; on frozen test these values are 0.0100 and 0.0713.

We keep GT, synthetic, in-sample predicted, out-of-fold (OOF) predicted, and held-out deployment predictions conceptually distinct. The primary R training source is OOF: five SegFormer models each predict one fold after training on the other four, and the predictions are reassembled before any classifier is fitted. Synthetic training masks target the IoU of those same OOF predictions. Validation and test masks are generated from their fixed deployment predictions. This construction aligns prediction quality during training with deployment while preventing the downstream classifier from learning from in-sample segmenter outputs. For validation and test, R is the fixed deployment prediction and S is its synthetic counterpart matched in IoU.

3.3 Crossing four cells and audit estimands

For training source $\tau \in \{S, R\}$, let $f_{\theta_\tau}(x, m)$ be a classifier trained on source τ. All factors unrelated to masks, including images, labels, architecture, optimization budget, seed set, and checkpoint rule, are held fixed. For evaluation source $\nu \in \{S, R\}$, define

$$A_{\tau\nu} = M(\{(f_{\theta_\tau}(x_i, m_i^\nu), y_i)\}_{i=1}^n), \tag{2}$$

where M is accuracy unless stated otherwise. The four cells are therefore A_{SS}, A_{SR}, A_{RS}, and A_{RR}.

For a fixed trained classifier, the effect of the evaluation source is

$$\mathrm{SOG}_{\mathrm{eval}}(\tau) = A_{\tau S} - A_{\tau R}. \tag{3}$$

The source-matched contrast $\mathrm{SOG}_{\mathrm{match}} = A_{SS} - A_{RR}$ changes both training and evaluation provenance and is therefore descriptive rather than isolating. The primary interaction estimand is the difference-in-differences

$$\mathrm{DiD} = (A_{SS} - A_{SR}) - (A_{RS} - A_{RR}) = \mathrm{SOG}_{\mathrm{eval}}(S) - \mathrm{SOG}_{\mathrm{eval}}(R). \tag{4}$$

A departure from zero means that substituting the evaluation mask has a different effect after training on synthetic versus predicted masks. Its sign does not by itself declare either source superior. Conversely, an interval containing zero is treated as inconclusive rather than evidence of equivalence.

3.4 Decision and mechanism diagnostics

The audit also tests whether proxy mismatch changes model choice. For a fixed candidate set $\mathcal{H}$, synthetic validation selects $h_S = \arg\max_{h \in \mathcal{H}} M_{\mathrm{val},S}(h)$, whereas deployment mask validation selects $h_R = \arg\max_{h \in \mathcal{H}} M_{\mathrm{val},R}(h)$. Both arms score the same checkpoint for each method; only the validation mask source used to rank methods changes. Deployment regret is

$$\mathrm{Regret} = M_{\mathrm{test},R}(h_R) - M_{\mathrm{test},R}(h_S). \tag{5}$$

Validation chooses the methods, and frozen test is used only to measure the consequence. Appendix Q gives the complete selection pseudocode.

Two controls probe why provenance matters. A prediction-anchored generator mixes the predicted mask with GT at the same IoU target for each image. A separate ladder with five levels varies the error field correlation length $\sigma \in \{0, 1.5, 3, 6, 12\}$ while holding the IoU target and marginal error rates fixed, thereby changing fragmentation without changing overlap. Finally, for matched fusion and seed, cross-fitting mitigation is measured by $\Delta_{\mathrm{CF}} = M_{\mathrm{test},R}(\mathrm{OOF}) - M_{\mathrm{test},R}(\text{in-sample})$.

4 EXPERIMENTAL SETUP

4.1 EVALUATION DOMAINS

The audit covers four domains with distinct evidence roles. PlantSeg (Wei et al., 2026) supplies the primary analysis on the frozen test set; BUSI (Al-Dhabyani et al., 2020) is a preregistered replication limited to validation; CUB-200-2011 (Wah et al., 2011) is a natural image extension using six seeds; and RIM-ONE DL (Batista et al., 2020) is a preregistered boundary analysis with hospitals kept disjoint. Exact tasks, sample counts, splits, exclusions, licenses, and evidence status are reported in Appendix A; screening and protocols specific to each domain appear in Appendices E, F, and H.

4.2 MASK SOURCES AND COMPARISON ARMS

The primary PlantSeg mask source is SegFormer-B0 (Xie et al., 2021), initialized from `nvidia/segformer-b0-finetuned-ade-512-512` and fine-tuned on the training split. Two zero-shot sources test segmenter dependence: CLIPSeg (Lüddecke & Ecker, 2022) and Grounding DINO (Liu et al., 2024) paired with SAM 3 (Carion et al., 2025). Neither is fine-tuned on PlantSeg. Prompts, model revisions, thresholds, and handling of empty outputs appear in Appendix L.

The comparison arms are mask consumers rather than externally reported SOTA scores, because the audit requires identical images, masks, training budgets, and checkpoints. We compare an image-only reference with hard-zero concat, soft concat, mask channel, padded ROI crop, feature attention pooling, ICGD, and UASMG. The two concat arms replace background content, mask channel retains the original image, ROI changes the field of view, and feature attention acts on downsampled features. ICGD and UASMG are in-house diagnostic consumers that use a scalar gate conditioned on the image and a gate based on self confidence, respectively. They broaden the consumer family and are not claimed as the primary method contribution.

4.3 IMPLEMENTATION DETAILS

The primary classifier is ConvNeXt-Tiny pretrained on ImageNet (Liu et al., 2022). Backbone checks add ResNet50 pretrained on ImageNet (He et al., 2016) and ViT-Base (Dosovitskiy et al., 2021) without tuning for each backbone. All classifiers use AdamW with learning rate 10^{-4}, weight decay 0.05, 15 epochs, batch size 64, patience 5, and no scheduler. Inputs are resized to 224×224 and normalized with ImageNet statistics; the only augmentation is a joint horizontal flip with probability 0.5. Validation alone selects checkpoints. Experiments use one of four NVIDIA GeForce RTX 4090 GPUs with Python 3.12, PyTorch 2.9.1, and CUDA 12.8. Detailed memory use and runtimes appear in Appendix A.

4.4 METRICS, CONTROLS, AND STATISTICAL INFERENCE

Accuracy is the primary downstream metric for PlantSeg, BUSI, and CUB; balanced accuracy is the preregistered primary metric for RIM-ONE. Balanced accuracy and macro-F1 provide PlantSeg and CUB checks balanced across classes, while accuracy and AUROC are secondary for RIM-ONE. IoU is matched for each image and is not the downstream endpoint. The primary interaction is DiD; $\mathrm{SOG}_{\mathrm{eval}}$, $\mathrm{SOG}_{\mathrm{match}}$, deployment regret, and cross-fitting effects answer the source, decision, and mitigation questions defined in Section 3.

Within each paired contrast, split, data order, seeds, training budget, and checkpoint policy are fixed. Evaluation with two masks at the same checkpoint is kept separate from accounting for checkpoints selected on validation. The crossed design, selection experiment, fusion sweep, class balance analysis, and control with three arms use six seeds; the fragmentation ladder and prediction-anchored arm retain three and are labelled accordingly.

We report intervals over images, intervals over seeds, and hierarchical intervals, and base significance statements only on the latter two because they resample training runs. Intervals over images average the contrast for each image over seeds and resample 2,294 image IDs 20,000 times with checkpoints fixed. Intervals over seeds intervals and two-sided one-sample t-tests use paired seed effects as the analysis unit; hierarchical intervals resample seeds and then images. BUSI's small cohort used only for validation limits inference across domains. CUB additionally uses a hierarchical bootstrap over

seeds, classes, and images, with repeated class draws retaining their multiplicity, and RIM-ONE uses a bootstrap stratified by hospital and class. Backbone and domain checks use hard-zero concat, while a separate fusion sweep tests consumer dependence. A fixed pretrained embedding audit screens possible duplicates between training and validation or test, followed by exclusion sensitivity at cosine thresholds 0.99 and 0.95.

5 Results

5.1 Masks matched in IoU are not interchangeable downstream on PlantSeg

On the frozen PlantSeg test set, synthetic and predicted masks matched in IoU are not interchangeable downstream. Both training arms use the same IoU target derived from OOF predictions for each image and differ only in mask provenance. A hard-zero classifier trained on synthetic masks scores 0.7833 on synthetic evaluation masks and 0.7427 on deployment masks, whereas a classifier trained on OOF masks scores 0.7350 and 0.7573, respectively. The effect of evaluation source therefore changes from $\mathrm{SOG}_{\mathrm{eval}}(S) = +4.06\,\mathrm{pp}$ to $\mathrm{SOG}_{\mathrm{eval}}(R) = -2.22\,\mathrm{pp}$, yielding $\mathrm{DiD} = +6.28\,\mathrm{pp}$ (95% CI over seeds $[5.31, 7.26]$). This sign reversal supports adaptation to mask provenance rather than a difference in mask difficulty that is independent of source.

The source-matched contrast is $+2.61$ pp ($[1.93, 3.29]$), but it changes both training and evaluation provenance and cannot isolate the interaction. The hierarchical DiD interval, which resamples seeds and images, is $[5.03, 7.56]$ pp and agrees with the result over seeds. Full uncertainty accounting and construction checks are in Appendices A through R.

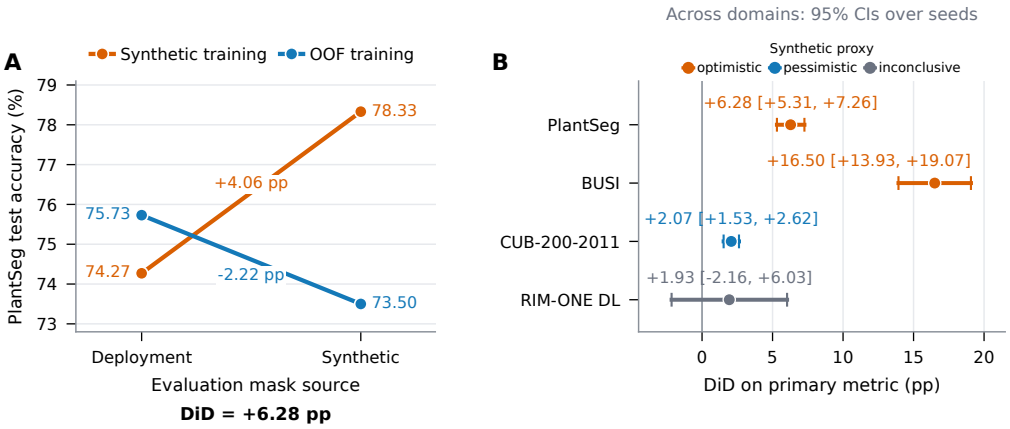


Figure 2: Primary source crossing and evidence across domains. (A) On the frozen PlantSeg test set, synthetic evaluation masks help the hard-zero classifier trained on synthetic masks but hurt the classifier trained on OOF masks. During training, both arms use targets defined by OOF predictions; evaluation synthetic masks are matched to the fixed deployment prediction for each test image. Values are means over six seeds. (B) DiDs on the primary metric with 95% intervals over seeds. Marker color denotes the practical direction of the source-matched comparison, not the sign of DiD. PlantSeg and BUSI are synthetic optimistic, CUB is synthetic pessimistic, and the RIM-ONE primary interval includes zero.

5.2 The interaction recurs across domains, but practical bias changes direction

Source dependence recurs across domains, but the practical bias is not universally optimistic (Table 1). BUSI reproduces the PlantSeg pattern with $\mathrm{DiD} = +16.50\,\mathrm{pp}$ ($[13.93, 19.07]$) under five-fold OOF provenance. Its 98-image evaluation cohort and lack of an independent test limit this result to a preregistered medical domain replication.

CUB provides the complementary natural image result. Its ROI crop crossing gives $\mathrm{DiD} = +2.07\,\mathrm{pp}$ ($[1.53, 2.62]$), but the practical direction reverses: ROI cropping with predicted masks improves on

Table 1: Source crossing evidence across domains. Intervals are prespecified 95% intervals over seeds. The final column describes the practical direction of the source-matched comparison; RIM-ONE is inconclusive on its primary metric.

Dataset	Primary metric	DiD	95% CI	Reading
PlantSeg	accuracy	$+6.28$ pp	$[+5.31, +7.26]$	synthetic optimistic
BUSI	accuracy	$+16.50$ pp	$[+13.93, +19.07]$	synthetic optimistic
CUB-200-2011	accuracy	$+2.07$ pp	$[+1.53, +2.62]$	synthetic pessimistic
RIM-ONE DL	balanced accuracy	$+1.93$ pp	$[-2.16, +6.03]$	inconclusive

image-only by $+1.99$ pp ($[1.62, 2.37]$), whereas the synthetic estimate is $+0.03$ pp ($[-0.40, 0.47]$). Synthetic evaluation therefore fails to reveal a useful crop from predicted masks. By contrast, RIM-ONE's prespecified balanced accuracy DiD is inconclusive; both its seed interval $[-2.16, +6.03]$ pp and interval stratified by hospital and class $[-4.90, +8.13]$ pp include zero. The secondary AUROC result is not promoted over this frozen primary endpoint.

5.3 Synthetic validation can change downstream model selection

Synthetic evaluation changes which mask consumer appears useful. This preserved PlantSeg selection audit uses synthetic training masks matched to in-sample segmenter targets and is distinct from the OOF primary crossing. Among eight candidates, hard-zero and soft concat are $+1.96$ pp above image-only under synthetic masks but -2.08 and -2.48 pp below it under deployment masks (Figure 3A). ICGD shows a smaller reversal, while mask channel falls to a tie.

Synthetic validation consequently selects hard-zero concat, whereas deployment mask validation selects image-only. On frozen deployment masks, their selected checkpoints score 0.7347 and 0.7538, producing $+1.91$ pp regret (95% CI over seeds $[0.48, 3.34]$). Neither validation arm selects the optimum on the frozen test set, UASMG, so deployment mask validation is better in this experiment but is not an oracle. CUB exposes the opposite consequence (Figure 3C): synthetic evaluation fails to identify the $+1.99$ pp benefit from ROI cropping with predicted masks. The proxy can therefore create false enthusiasm or fail to detect a benefit, depending on the combination of generator and consumer.

5.4 The finding persists across tested segmenters and consumers

The PlantSeg interaction is not explained by one segmenter, residual IoU mismatch, or class imbalance. Zero-shot CLIPSeg and Grounding DINO with SAM 3 produce DiDs of $+10.56$ and $+13.51$ pp, with all six seeds positive for each. Restricting evaluation to the subset with tight overlap matching leaves DiD $= +5.86$ pp, and balanced accuracy with sequence clustering gives $+5.95$ pp with a 95% interval of $[4.85, 7.90]$ pp. The classwise DiD is positive for 84 of 115 classes. Because zero-shot mask quality and mask source change together, their larger magnitudes are observational rather than controlled comparisons.

Consumer design changes magnitude but not direction among the four tested fusion modes. DiD ranges from $+8.44$ pp for soft concat and $+6.16$ pp for hard-zero concat to $+1.03$ pp for ICGD and $+0.96$ pp for UASMG, with each seed interval excluding zero. The largest interactions occur when the mask can remove spatial content; consumers that preserve the image are less sensitive. Complete segmenter, metric, fusion, and spatial removal controls are in Appendices K through N.

5.5 Matching IoU or fragmentation does not ensure downstream fidelity

CUB exposes a concrete interaction between generator and consumer. ROI cropping converts a mask into a padded bounding box. Although predicted and synthetic masks have closely matched mean IoU, 99.55% of synthetic boxes cover more than 99% of the image, compared with 0.93% of boxes from predicted masks; their median area fractions are 1.000 and 0.421. The synthetic crop therefore behaves almost like the full image and fails to reproduce the field-of-view change induced by predicted masks.

PlantSeg shows that repairing one measured structural discrepancy is also insufficient. A prediction-anchored generator is closer on six measured statistics but remains far from deployment in component

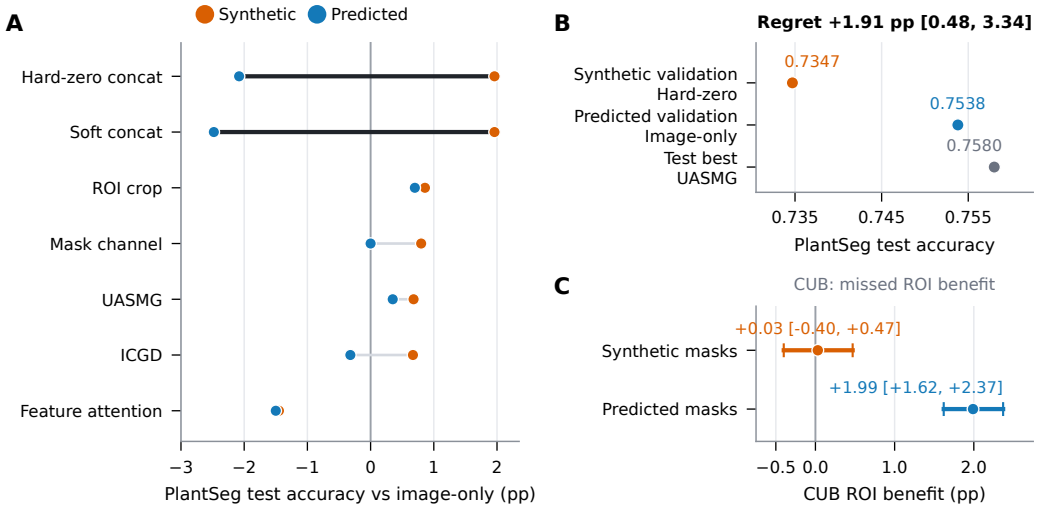


Figure 3: Decision consequences in PlantSeg and CUB. The PlantSeg selection audit uses synthetic training masks matched to in-sample segmenter targets, rather than the OOF targets in Figure 2. (A) Two concat consumers reverse relative to image-only under all six seeds; ICGD shows a smaller reversal and mask channel falls to a tie. Color denotes the evaluation mask source. (B) Synthetic and predicted mask validation select hard-zero concat and image-only, respectively, yielding $+1.91$ pp PlantSeg regret ($[0.48, 3.34]$). Neither selects UASMG, the best method on the frozen test set. (C) On CUB, synthetic evaluation estimates a $+0.03$ pp ROI benefit ($[-0.40, 0.47]$), whereas predicted masks give $+1.99$ pp ($[1.62, 2.37]$).

count and retains $\mathrm{SOG}_{\mathrm{eval}}(S) = +2.66$ pp. Across the fragmentation ladder, component count distance falls from 87.78 to 0.25 reference standard deviations, yet the most closely matched rung still has $\mathrm{SOG}_{\mathrm{eval}}(S) = +4.26$ pp ($[3.39, 5.14]$). The trend over seeds is -0.58 pp per decade ($[-1.17, +0.01]$), so closer fragmentation does not produce a corresponding decline in downstream bias. Full structural distributions and ladder results are in Appendix D.

5.6 Cross-fitting improves provenance alignment but is only a partial mitigation

OOF masks are closest to deployment predictions across all seven measured features beyond IoU, and their linear source probe AUC is 0.515, compared with 0.770 for in-sample predictions and nearly 1.0 for synthetic sources (Figure 5). In the control with three arms and six seeds, adding segmenter diversity without correcting provenance changes accuracy by -0.09 pp ($[-0.87, +0.68]$). OOF instead exceeds the in-sample mosaic from five models by $+1.06$ pp ($[0.60, 1.52]$) and the in-sample arm from one model by $+0.97$ pp ($[0.38, 1.55]$). Pooling the two preregistered cohorts gives $\Delta_{\mathrm{CF}} = +1.10$ pp ($[0.48, 1.73]$).

The benefit depends on architecture. ResNet50 improves by $+0.86$ pp, while ViT-Base improves by only $+0.10$ pp with mixed seed signs. ConvNeXt-Tiny returns to within 0.46 pp of its contemporaneous image-only baseline rather than establishing a clear net benefit from mask guidance. Cross-fitting therefore improves provenance alignment and recovers part of the associated loss, but it does not remove the need for a source-crossed audit.

6 Discussion and Limitations

Principal interpretation. The stable finding across domains is a lack of interchangeability, not a universally optimistic synthetic bias. Source-matched evaluation inflates the apparent value of mask guidance on PlantSeg and BUSI, whereas on CUB it hides a useful crop from predicted masks; RIM-ONE leaves the primary interaction unresolved. This pattern is consistent with the mask generator

becoming part of the classifier's input distribution. Its process determines where information is removed, retained, or reframed, and the consumer can adapt to recurring errors specific to a source. Such adaptation resembles shortcut learning (Geirhos et al., 2020), although our design identifies a provenance interaction rather than the specific features used by the classifier.

Relation to prior work. Our results support the view that overlap alone cannot characterize segmentation errors with distinct locations, boundaries, or components (Bernhard et al., 2024). They also qualify evidence that synthetic data can preserve model level reliability rankings (Loiseau et al., 2024): here, the downstream model is not fixed but learns from the proxy, so fidelity depends on the joint system of mask and consumer. This distinction helps reconcile our findings with mixed transfer from synthetic to natural corruptions (Taori et al., 2020; Agnihotri et al., 2025). It further extends work on error propagation in cascaded imaging systems (Mahbod et al., 2020; Mehta et al., 2022): provenance mismatch can change not only accuracy but also which downstream design appears preferable.

Implications for evaluation. Mask similarity should therefore be treated as a diagnostic, not an acceptance test. A minimal audit trains consumers on synthetic masks and masks representative of deployment, crosses both checkpoints with both evaluation sources, and reports the four cells and their DiD with uncertainty over seeds. Image-only performance, selection consequences, fusion mode, and mask provenance should accompany this comparison. Consumer geometry also matters: CUB shows that a sparse false positive can expand an ROI crop to nearly the full image despite matched IoU. OOF prediction can reduce provenance mismatch for a sensitive consumer, but its neutral or adverse effects on other backbones preclude treating it as a general correction. Likewise, GT benefit, morphology, and segmenter IoU can prioritize audits but cannot certify downstream fidelity; the CUB and Severstal screens illustrate both sides of this limitation.

Scope and limitations. PlantSeg supplies the primary result on the frozen test set. BUSI is preregistered but uses the same cohort of 98 images for checkpoint selection and evaluation. CUB combines three previously observed and three prospectively frozen seeds, tests one zero-shot segmenter and one ROI implementation, and may inherit ImageNet pretraining overlap noted by its release. RIM-ONE has aligned deployment IoU but wide primary intervals, so failure to reject is not evidence of equivalence; its AUROC and results stratified by hospital remain secondary. The PlantSeg selection experiment did not rerun the full candidate pool with synthetic training matched to OOF targets, and thus measures the consequence of the practiced protocol based on in-sample targets rather than isolating mask structure. The cross-fitting observation is post hoc, the structural ladder varies two statistics together, and PlantSeg's split defined at the image level lacks a guaranteed biological unit. We therefore do not claim validity for every generator, consumer, architecture, or end-to-end system.

Decisive next tests. Future work should preregister source-crossed evaluation on independent cohorts with splits disjoint by biological unit or patient; manipulate location, boundary, fragmentation, and crop extent independently while holding IoU fixed; and repeat model selection with training matched to OOF targets across multiple consumer geometries. These experiments would distinguish which error structures drive provenance adaptation and determine when cross-fitting is a sufficient mitigation rather than merely an improvement in alignment.

7 Conclusion

For the tested generators and consumers, matching IoU for each image does not make synthetic masks and masks predicted by a segmenter interchangeable downstream. The direction of the resulting discrepancy varies across domains and mask consumption mechanisms, indicating that proxy fidelity is a property of the joint mask source and consumer system rather than of mask overlap alone. This mismatch can alter model selection and persists after closer structural matching, while masks obtained by cross fitting provide only partial mitigation whose magnitude depends on architecture. Studies that use synthetic masks to support downstream design choices should therefore report mask provenance and evaluate their conclusions through source crossing with four cells under masks representative of deployment. Independent replications with disjoint units and controlled manipulation of individual error structures are needed to determine when simpler mask level proxies are sufficient.

Reproducibility Statement

The submission reports exact dataset snapshots and splits, mask provenance, training hyperparameters, checkpoint selection rules, effects over seeds, and analysis status. The supplementary appendix records the frozen logit protocol and principal run-level checks. An anonymized code and result bundle will include experiment scripts, tables for the frozen test set, figure data, and configuration registries. Canonical logits for each image are retained for the source crossing, selection, cross-fitting, foldin, prediction-anchored, and exclusion analyses. Image ID intervals are conditional on fixed checkpoints. DataLoader worker count was outside the original seed contract because its generator was not pinned; otherwise identical runs vary by roughly ± 0.5–0.7 pp. We therefore pool two replicates for the headline cross-fitting effect and use a seed-level interval for the fragmentation trend.

Ethics Statement

This study performs secondary analysis of public plant disease and breast ultrasound datasets and collects no new data from human participants. We disclose that the BUSI release page does not state an explicit dataset license, distinguish the article license from the data, exclude its normal class with an empty mask to avoid label leakage, and avoid clinical performance claims. PlantSeg is used under its non-commercial data license. All authors must confirm compliance with the applicable dataset terms and the ICLR Code of Ethics before submission.

AI Use Statement

Generative AI tools were used for language editing, manuscript organization, LaTeX conversion, literature metadata checks, figure drafting, code editing, consistency checks against the authors' experiment records, and feedback on experimental design and result interpretation. The authors made the final research and analysis decisions, executed the experiments, and manually verified all AI-assisted text, figures, code changes, numerical results, citations, and claims. No experimental data were generated or substituted by generative model output. The authors take responsibility for the final content.

SAM 3 Acknowledgment

This work uses Meta's Segment Anything Model 3 materials and acknowledges Meta in accordance with the SAM License. Funding, institutional, contributor, and acknowledgments that identify the authors are omitted from the anonymous submission.

References

Shashank Agnihotri, David Schader, Nico Sharei, Mehmet Ege Kaçar, and Margret Keuper. Are synthetic corruptions a reliable proxy for real-world corruptions?, 2025. URL https://arxiv.org/abs/2505.04835. CVPR Workshop on Synthetic Data for Computer Vision.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. doi: 10.1016/j.dib.2019.104863. URL https://doi.org/10.1016/j.dib.2019.104863.

Francisco José Fumero Batista, Tinguaro Diaz-Aleman, Jose Sigut, Silvia Alayon, Rafael Arnay, and Denisse Angel-Pereira. RIM-ONE DL: A unified retinal image database for assessing glaucoma using deep learning. *Image Analysis & Stereology*, 39(3):161–167, 2020. doi: 10.5566/ias.2346. URL https://doi.org/10.5566/ias.2346.

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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 968–977, 2024. URL https://openaccess.thecvf.com/content/WACV2024/html/Bernhard_

Whats_Outside_the_Intersection_Fine-Grained_Error_Analysis_for_Semantic_Segmentation_WACV_2024_paper.html.

Matthias Bruhns, Jan T. Schleicher, Maximilian Wirth, Marcello Zago, Sepideh Babaei, and Manfred Claassen. Effects of segmentation errors on downstream-analysis in highly-multiplexed tissue imaging. *PLOS Computational Biology*, 21(9):e1013350, 2025. doi: 10.1371/journal.pcbi.1013350. URL https://doi.org/10.1371/journal.pcbi.1013350.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719. Model and code release under the SAM License.

Yoonjae Cho, Sampa Misra, Ravi Managuli, Richard G. Barr, Jeongmin Lee, and Chulhong Kim. Attention-based fusion network for breast cancer segmentation and classification using multi-modal ultrasound images. *Ultrasound in Medicine & Biology*, 51(3):568–577, 2025. doi: 10.1016/j.ultrasmedbio.2024.11.020. URL https://doi.org/10.1016/j.ultrasmedbio.2024.11.020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Wilfrido Gómez-Flores, Maria Julia Gregorio-Calas, and Wagner Coelho de Albuquerque Pereira. BUS-BRA: A breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical Physics*, 51(4):3110–3123, 2024. doi: 10.1002/mp.16812. URL https://doi.org/10.1002/mp.16812.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Shilu Kang, Dongfang Li, Jiaxin Xu, Aokun Mei, and Hua Huo. Segmentation-assisted fusion-based classification for automated CXR image analysis. *Sensors*, 25(15):4580, 2025. doi: 10.3390/s25154580. URL https://doi.org/10.3390/s25154580.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2024.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.html.

Thibaut Loiseau, Tuan-Hung Vu, Mickael Chen, Patrick Pérez, and Matthieu Cord. Reliability in semantic segmentation: Can we use synthetic data? In *European Conference on Computer Vision*,

2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/03410.pdf.

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7086–7096, 2022.

Amirreza Mahbod, Philipp Tschandl, Georg Langs, Rupert Ecker, and Isabella Ellinger. The effects of skin lesion segmentation on the performance of dermatoscopic image classification. *Computer Methods and Programs in Biomedicine*, 197:105725, 2020. doi: 10.1016/j.cmpb.2020.105725. URL https://pubmed.ncbi.nlm.nih.gov/32882594/.

Raghav Mehta, Thomas Christinck, Tanya Nair, Aurélie Bussy, Swapna Premasiri, Manuela Costantino, M. Mallar Chakravarthy, Douglas L. Arnold, Yarin Gal, and Tal Arbel. Propagating uncertainty across cascaded medical imaging tasks for improved deep learning inference. *IEEE Transactions on Medical Imaging*, 41(2):360–373, 2022. doi: 10.1109/TMI.2021.3114097. URL https://doi.org/10.1109/TMI.2021.3114097.

Mariia Rizhko, Lauren Erdman, Mandy Rickard, Kunj Sheth, Daniel Alvarez, Kyla N. Velaer, Megan A. Bonnett, Christopher S. Cooper, Gregory E. Tasian, John Weaver, Alice Xiang, Armando J. Lorenzo, and Anna Goldenberg. Improving identically distributed and out-of-distribution medical image classification with segmentation-guided attention in small dataset scenarios. In *Proceedings of Medical Imaging with Deep Learning*, volume 250 of *Proceedings of Machine Learning Research*, pp. 1282–1296, 2024. URL https://proceedings.mlr.press/v250/rizhko24a.html.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d8330f857a17c53d217014ee776bfd50-Abstract.html.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009. doi: 10.1016/j.neunet.2009.06.042. URL https://doi.org/10.1016/j.neunet.2009.06.042.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-UCSD birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. URL https://www.vision.caltech.edu/datasets/cub_200_2011/.

Ying Wang, Jie Su, Qiuyu Xu, and Yixin Zhong. A collaborative learning model for skin lesion segmentation and classification. *Diagnostics*, 13(5):912, 2023. doi: 10.3390/diagnostics13050912. URL https://doi.org/10.3390/diagnostics13050912.

Tianqi Wei, Zhi Chen, Xin Yu, Scott Chapman, Paul Melloy, and Zi Huang. A large-scale in-the-wild dataset for plant disease segmentation. *Scientific Data*, 13:205, 2026. doi: 10.1038/s41597-025-06513-4. URL https://www.nature.com/articles/s41597-025-06513-4.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1. URL https://doi.org/10.1016/S0893-6080(05)80023-1.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html.

Yutong Xie, Jianpeng Zhang, Yong Xia, and Chunhua Shen. A mutual bootstrapping model for automated skin lesion segmentation and classification. *IEEE Transactions on Medical Imaging*, 39 (7):2482–2493, 2020. doi: 10.1109/TMI.2020.2972964. URL https://pubmed.ncbi.nlm.nih.gov/32070946/.

Hee-Jin Yu and Chang-Hwan Son. Apple leaf disease identification through region-of-interest-aware deep convolutional neural network. *Journal of Imaging Science and Technology*, 64(2): 020507–1–020507–10, 2020. URL https://arxiv.org/abs/1903.10356.

Yuliang Zhang, Fang He, Lulu Peng, Tianyu Yan, Pingping Zhang, Ting Song, Lili Du, and Dunjin Chen. 3D segment anything model with visual mamba for diagnosing placenta accreta spectrum, 2026. URL https://arxiv.org/abs/2606.00489.

Alex Zwanenburg, Stefan Leger, Linda Agolli, Karoline Pilz, Esther G. C. Troost, Christian Richter, and Steffen Löck. Assessing robustness of radiomic features by image perturbation. *Scientific Reports*, 9:614, 2019. doi: 10.1038/s41598-018-36938-4. URL https://www.nature.com/articles/s41598-018-36938-4.

A Dataset and Protocol Details

Table 2: Datasets and evidence roles. Numeric splits are image counts in train/validation/evaluation order. RIM-ONE lists hospital strata because its preregistered analysis keeps hospitals disjoint. PlantSeg metadata contains 2,295 test records, but one lacks a usable class label.

Dataset	Classification task	Train/val./eval.	Evidence role
PlantSeg	115-way disease	7,916/1,247/2,294	primary frozen test
BUSI	binary tumor	549/98/–	preregistered replication
CUB-200-2011	200-way species	5,994/–/5,794	extension with six seeds
RIM-ONE DL	binary glaucoma	HUC/–/HUMS+r1	preregistered boundary

Computational environment. Experiments ran on a workstation with four NVIDIA GeForce RTX 4090 GPUs, using Python 3.12 and PyTorch 2.9.1 with CUDA 12.8. Each PlantSeg classifier ran on one GPU; a ConvNeXt-Tiny hard-zero concat run for 15 epochs with batch size 64 took approximately 45–60 minutes (about one GPU-hour) and peaked at approximately 15 GB. The five SegFormer-B0 OOF folds were trained sequentially on one GPU, taking 44.8–84.9 minutes per fold and 5.55 GPU-hours in total. Wall-clock time was approximately 5.7 hours including dump merging.

PlantSeg data are CC BY-NC 4.0 and its article is CC BY-NC-ND 4.0. Its segmenter is fitted on training data and stopped early on validation; mask consumers are selected on validation and the 2,294 usable test images remain frozen until final evaluation. The supplied split is defined at the image level rather than grouped by plant or leaf. The class distribution has a long tail (approximately 5–271 images per class in repository metadata). The implementation with 116 outputs shifts the 115 disease indices by one and reserves output zero for background. No stronger guarantee against duplicates at the plant or leaf level is available.

The original BUSI data page states no explicit dataset license; the accompanying article is CC BY 4.0. BUSI uses `SPLIT_SEED=1234` and a stratified 15% holdout defined at the image level; the normal class is removed because its empty mask would reveal its label. Its validation cohort of 98 images is also the evaluation set, so it is not an independent external test. CUB uses the official split and evaluates all six seeds on the official test; each of its 200 classes is represented. It uses Grounding DINO with SAM 3 in zero shot mode and no segmenter fitted specifically to CUB. RIM-ONE DL follows its separate preregistered manifest, which keeps hospitals disjoint, rather than the screening split: HUC supplies training and OOF masks, while HUMS and r1 remain held out for evaluation (Appendix H).

B Canonical Uncertainty and Replicate Accounting

Canonical logits are available for all central contrasts on the frozen test set. The crossed design and the selection experiment run six seeds; the remaining rows retain the August cohort of three seeds and are marked accordingly. Intervals over images average the contrast for each image over seeds and resample 2,294 image IDs 20,000 times with checkpoints held fixed, so they cover image

sampling only. Intervals over seeds and hierarchical intervals resample seeds and are the basis for every significance statement in this paper.

Table 3: Locked estimates on the frozen test set. CI level is part of the estimand and must not be mixed across rows. Rows marked ‡ use three seeds and are not directly comparable with the rows using six seeds above them.

Estimand	Point	95% CI	Uncertainty level
$\mathrm{SOG}_{\mathrm{eval}}(S)$	$+4.06$ pp	$[+3.57, +4.55]$	seed, six runs
$\mathrm{SOG}_{\mathrm{eval}}(R)$	-2.22 pp	$[-3.13, -1.32]$	seed, six runs
DiD	$+6.28$ pp	$[+5.31, +7.26]$	seed, six runs
$\mathrm{SOG}_{\mathrm{match}}$	$+2.61$ pp	$[+1.93, +3.29]$	seed, six runs
Selection regret	$+1.91$ pp	$[+0.48, +3.34]$	seed, six runs
Δ_{CF} (pooled)†	$+1.10$ pp	$[+0.48, +1.73]$	six paired runs
OOF − foldin	$+1.06$ pp	$[+0.60, +1.52]$	seed, six runs
Arm anchored to predictions $\mathrm{SOG}_{\mathrm{eval}}(S)$‡	$+2.66$ pp	$[+1.95, +3.39]$	image ID
Fragmentation slope‡	-0.58 pp/decade	$[-1.17, +0.01]$	seed

†The conditional estimate from August is $+0.81$ pp; see Table 6. ‡August cohort with three seeds.

The two cross-fitting replicates differ only in DataLoader worker count (five versus six). Worker count was outside the seed contract because the DataLoader generator was not pinned; repeated accuracy on deployment test masks varies by roughly ± 0.5–0.7 pp. All six paired differences between OOF and in-sample runs are positive ($p = 0.006$). The foldin arm now has six seeds drawn from the same cohort as the other two, so all three arms are directly comparable.

Two calibration facts bound how finely this benchmark resolves differences between arms. Our own reimplementation of the probabilistic mix generator at zero correlation length, which is the same estimand as the canonical arm, scores 0.96 pp below it; differences of about one percentage point between generators are therefore not resolvable here. The cohorts with six and three seeds differ by 0.10 pp on $\mathrm{SOG}_{\mathrm{eval}}(S)$, which is why they are labelled separately rather than merged.

Table 4: Estimands on the frozen test set under OOF provenance, using six seeds. Intervals over images resample the 2,294 test image IDs with checkpoints held fixed; intervals over seeds treat the seed as the analysis unit; hierarchical intervals resample seeds and then images. Significance claims use the latter two.

Estimand	Point	Image 95%	Seed 95%	Hier. 95%
$\mathrm{SOG}_{\mathrm{eval}}(S)$	$+4.06$ pp	$[+3.33, +4.81]$	$[+3.57, +4.55]$	$[+3.19, +5.01]$
$\mathrm{SOG}_{\mathrm{eval}}(R)$	-2.22 pp	$[-2.89, -1.55]$	$[-3.13, -1.32]$	$[-3.23, -1.22]$
DiD	$+6.28$ pp	$[+5.39, +7.19]$	$[+5.31, +7.26]$	$[+5.03, +7.56]$
$\mathrm{SOG}_{\mathrm{match}}$	$+2.61$ pp	$[+1.79, +3.42]$	$[+1.93, +3.29]$	$[+1.51, +3.71]$

C IoU Matching and Possible Duplicate Audits

Table 5: Absolute residuals from the IoU target for each image.

Split	Median	95th percentile	Mean
Train	0.0093	0.0218	0.0104
Validation	0.0105	0.0595	0.0183
Test	0.0100	0.0713	0.0183

There are no exact filename stem overlaps from validation or test into train. At cosine similarity at least 0.99, the embedding audit flags 118/1,247 validation images (9.46%) and 195/2,295 test images (8.50%); the embedding audit includes the one test record that lacks a usable classification

label, hence 2,295 rather than 2,294. The corresponding 0.95 rates are 12.91% and 12.42%. Table 6 recomputes all available estimands after removing flagged test images.

Table 6: Sensitivity to excluding possible duplicates. All values are percentage points; validation selection remains fixed. This table is anchored on the August cohort of three seeds throughout, so its full test column differs slightly from the headlines using six seeds in Table 3; the comparison of interest is across columns within a row. The cross-fitting row is the conditional estimate from August; the pooled headline over six runs is $+1.10$ pp and cannot be recomputed under exclusion because the July replicate lacks logits for each image.

Estimand (three seeds)	Full test	Drop cosine ≥ 0.99	Drop cosine ≥ 0.95
$\mathrm{SOG}_{\mathrm{eval}}(S)$	$+4.04$	$+4.07$	$+4.11$
$\mathrm{SOG}_{\mathrm{eval}}(R)$	-2.28	-2.64	-2.70
DiD	$+6.32$	$+6.70$	$+6.82$
Δ_{CF} (August)	$+0.81$	$+0.83$	$+0.95$
OOF $-$ foldin	$+1.31$	$+1.29$	$+1.41$
Arm anchored to predictions $\mathrm{SOG}_{\mathrm{eval}}(S)$	$+2.66$	$+2.72$	$+2.80$
Selection regret	$+2.08$	$+2.35$	$+2.52$

The largest point estimate drift is 0.50 pp, on DiD, and every estimand keeps its sign and becomes at least as large. If leakage from possible duplicates influenced these contrasts at all, it diluted them rather than manufacturing them.

D Structural Matching and Fragmentation Ladder

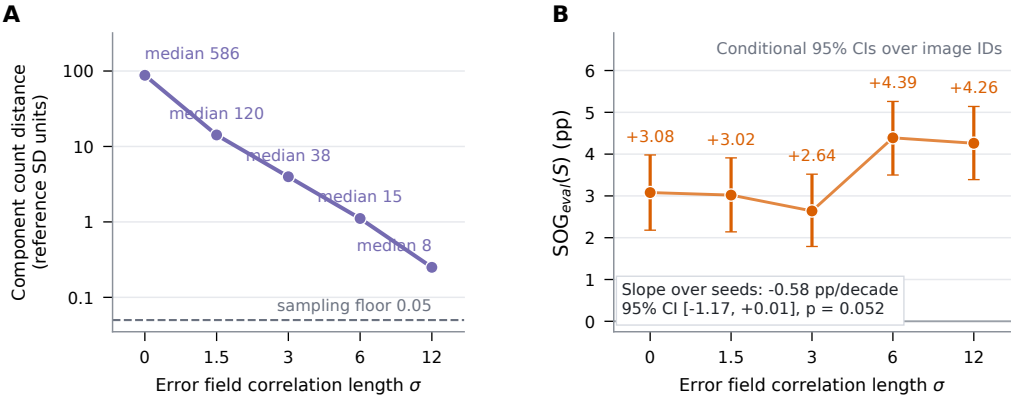


Figure 4: PlantSeg fragmentation control on the August cohort of three seeds. Each rung has the same IoU target for each image and the same marginal false-negative and false-positive rates. (A) Increasing error field correlation length reduces the component count distance by 350×, to five times the sampling floor. (B) The downstream gap does not decrease. Rung intervals are image ID bootstrap intervals conditional on checkpoints; the reported trend uses an interval over the three seeds.

Table 7 reports one-dimensional 1-Wasserstein distances to deployment masks, normalized by the reference distribution standard deviation. The sampling floor is obtained by randomly splitting the reference set in half.

Table 7: Structural distance for each feature (0 is matched).

Arm	Centroid	Outside	Components	FN/FP	Area	Boundary	Recall
Probabilistic mix	0.496	0.200	86.40	0.252	0.143	0.403	0.539
Anchored to predictions	0.501	0.212	53.92	0.606	0.144	0.320	0.538
In-sample predicted	0.447	0.415	0.095	0.320	0.169	0.466	0.304
OOF from cross fitting	0.025	0.025	0.085	0.059	0.118	0.021	0.037
Sampling floor	0.051	0.045	0.050	0.043	0.054	0.051	0.046

Table 8: Fragmentation ladder. Rung CIs are image ID intervals conditional on the three checkpoints; the trend interval in Table 3 is reported as an interval over seeds.

σ	Median components	Component distance	$\mathrm{SOG}_{\mathrm{eval}}(S)$	95% CI
0	586	87.78	$+3.08$ pp	$[+2.18, +3.98]$
1.5	120	14.21	$+3.02$ pp	$[+2.14, +3.91]$
3	38	3.97	$+2.64$ pp	$[+1.79, +3.52]$
6	15	1.11	$+4.39$ pp	$[+3.50, +5.26]$
12	8	0.25	$+4.26$ pp	$[+3.39, +5.14]$

The ladder does not vary only one variable. Four statistics are fixed by construction, but centroid displacement improves from 0.489 to 0.113 reference standard deviations while boundary distance worsens from 0.373 to 0.562. The five rungs therefore cannot separate features that change together.

E Dataset Screening and GT Mask Benefit

The dataset screen asked whether masks had measurable value before testing synthetic mask fidelity. Across the historical screen and later preregistered extensions, thirteen candidates carrying both classification labels and foreground masks were evaluated; only PlantSeg and BUSI passed. VGT, Brain and Thyroid were metadata level exclusions and are not part of this denominator. Table 9 reports the complete experimental accounting.

Table 9: Dataset screening pipeline over thirteen candidates carrying both classification labels and foreground masks. The screen is a scope device, not a claim that failed datasets have no useful mask representation. Benefit is the best consumption mode tried on that dataset; CUB, FungiTastic-Mini and Severstal were screened again across consumption modes rather than under hard-zero alone.

Dataset	Task	Best consumer	Image-only	GT mask	Benefit
PlantSeg v3	plant disease, 115	hard-zero concat	0.7495	0.7899	$+\mathbf{4.0}$ pp
BUSI binary	breast US, 2	hard-zero concat	0.9048	0.9660	$+\mathbf{6.12}$ pp
SegPPD-101	rice disease, 101	hard-zero concat	0.7330	0.7530	$+2.00$ pp†
CUB-200-2011	bird species, 200	ROI crop, pad 0.1	0.8422	0.8601	$+1.79$ pp*
BUS-BRA	breast US, 2	dual concat	0.8765	0.8889	$+1.23$ pp§
PAPILA	fundus glaucoma	hard-zero concat	0.8849	0.9008	$+1.59$ pp
ISIC 2016 3B	dermoscopy, 2	hard-zero concat	0.8505	0.8602	$+0.97$ pp
FungiTastic-Mini	fungi, 215	mask channel	0.6167	0.6238	$+0.71$ pp
HAM10000	dermoscopy, 7	hard-zero concat	0.9104	0.9154	$+0.51$ pp
Oxford-IIIT Pet	breed, 37	hard-zero concat	0.9388	0.9415	$+0.27$ pp
Severstal	steel defects, 4	ROI crop, pad 0.1	0.9669	0.9749	$+0.80$ pp‡
RiceSeg	rice disease, 3	hard-zero concat	≈ 1.00	≈ 1.00	≈ 0
Plant Pathology 2020	apple leaf, 2	—	1.000	—	≈ 0

†Directionally consistent but underpowered ($n = 5$, $p = 0.148$). *Historical gate value; the later CUB crossing with six seeds finds DiD $= +2.07$ pp without changing the original gate outcome. §Preregistered validation gate; the test set disjoint by patient remained locked. ‡The preregistered > 3 pp Severstal prediction failed; the 95% interval over seeds is $[-0.70, +2.30]$ pp. Rows are grouped as admitted, failed with headroom, and limited by a ceiling.

Four candidates are constrained by a ceiling: their image-only classifiers score 0.94–1.00, leaving little room for a mask to add. Six fail despite ample headroom. Five of those six contain a large, centred or predictable target, so localisation is not the main bottleneck; SegPPD-101 is the underpowered exception whose $+2.00$ pp benefit points in the admitted direction. Post hoc, the two admitted datasets are also the two with small, unpredictably located targets. This pattern is descriptive rather than predictive.

Preregistered Severstal forecast. Before any Severstal training result existed, we predicted that its small, spatially dispersed defects would pass the 3 pp gate. We used 6,239 images containing one defect and a nonempty mask, with a fixed stratified 4,991/1,248 train/validation split. Accuracy from images alone was already 0.9669; the best consumer, padded ROI crop, reached 0.9749, a GT benefit of only $+0.80$ pp (effects by seed $+1.44$, $+0.72$, $+0.24$ pp; 95% CI over seeds $[-0.70, +2.30]$). The prediction therefore failed. The available headroom was 3.31 pp, leaving only 0.31 pp above the gate rather than making passage mathematically impossible. A classifier using only masks reached 0.8868 versus a 0.7628 majority baseline, showing that mask geometry was strongly useful for class discrimination; nevertheless it did not produce an incorrect positive gate decision. The validation supports across the four classes were 154/39/952/103; balanced accuracy rose from 0.9128 using images alone to 0.9417 with ROI crop ($+2.89$ pp), while dual input reached 0.9453. We retain the preregistered accuracy gate rather than changing metrics after seeing the result. The failure demotes target geometry to a secondary descriptor: future screens must measure actual headroom from the image-only reference first.

This is measurable rather than impressionistic. On CUB, Grounding DINO with SAM 3 in zero shot mode reaches a mean IoU of 0.850 across images against ground truth (median 0.905; 0.853 on the train split, so its results across splits agree to within 0.31 pp), whereas SegFormer fine-tuned on PlantSeg reaches 0.604 on the PlantSeg deployment split. Where the segmenter barely errs there is little segmentation error for a mask source to disagree about, and this compounds with the absence of localisation signal rather than substituting for it.

Preregistered BUS-BRA gate. For BUS-BRA (Gómez-Flores et al., 2024), P13 froze a split disjoint by patient before training: 745/105/214 patients and 1,305/189/381 images for train/validation/test. Stage 1 used validation only. Accuracy from images alone was 0.8765 (balanced accuracy 0.8545); the best arm, dual concat, reached 0.8889 (0.8650), a GT benefit of $+1.23$ pp. Mask channel gave $+0.88$ pp, ROI crop -0.35 pp, and masks alone -2.29 pp. The dual gain was $0, +1.59, +2.12$ pp across seeds and did not pass the frozen 3 pp gate, so stage 2 stopped and the test set disjoint by patient remained unopened. This is a transparent screening failure, not an external replication.

Two limits of the screen itself. First, GT mask benefit is a proxy for downstream mask source sensitivity, not a sufficient condition for it. This paper reports a case where the two come apart: under deployment masks on PlantSeg the net utility of mask guidance is approximately zero and four of seven consumers are negative, yet DiD is $+6.28$ pp. A dataset could therefore show negligible GT benefit and still show a source effect, so we verified one rejection directly (Appendix F). Second, the screen is not a property of the dataset alone. A weaker segmenter produces lower IoU for each image and therefore more apparent room for mask source to matter, so screening the same dataset with a poorer segmenter makes it look more admissible. We fix the segmenter within each screen for this reason and report IoU for each split alongside any screening decision.

The historical screen used only hard-zero dual concat. A preregistered PlantSeg validation check therefore measured GT mask benefit under four consumption modes against the same image-only baseline (0.7485). Results in Table 10 span 5.45 pp, so the screen depends on the mode; however, hard-zero ranks first, meaning it was the most permissive tested mode on PlantSeg and could not have falsely rejected this dataset. This does not resolve the risk of an incorrect negative screen in domains that need global context. We avoid the term "oracle upper bound": the maximum over modes is biased by selection and is reported only alongside all four values.

Table 10: PlantSeg validation split GT mask benefit under four consumption modes, using the contemporaneous image-only baseline of 0.7485.

GT mask use	Accuracy	Benefit	Benefit by seed (pp)
Hard-zero dual concat	0.7880	$+3.96$ pp	$+3.61, +3.93, +4.33$
Hard mask channel	0.7749	$+2.65$ pp	$+2.01, +3.77, +2.16$
Padded ROI crop	0.7607	$+1.23$ pp	$+0.80, +1.04, +1.84$
Hard feature attention	0.7335	-1.50 pp	$-1.28, -2.41, -0.80$

F CUB Screening History and Source Crossing with Six Seeds

Because GT mask benefit is only a proxy for mask source sensitivity, we ran a crossed design on one rejected dataset. The historical P9 rule promoted a dataset only if $\mathrm{DiD} \geq 3$ pp with an interval over seeds excluding zero. The initial CUB result with three seeds did not pass that gate, and this screening outcome is retained. P15 subsequently froze a distinct extension with six seeds: seeds 0–2 were already observed, while seeds 3–5 and the interpretation rules were fixed before those new runs began. This extension tests source lack of interchangeability; it does not retroactively convert the original screen into an independent blind test.

CUB-200-2011 is the strongest available target. It has 15.8 pp of headroom, so its rejection is not a ceiling artefact, and among rejected datasets it has the highest and most stable GT mask benefit ($+1.79$ pp; effects by seed $+1.71$, $+2.03$, $+1.64$). It was screened under three consumption modes, not only hard-zero (hard-zero $+0.33$ pp, mask channel ≈ 0.00 pp, ROI crop $+1.79$ pp), so its rejection is not an artefact of a consumer that destroys context. The segmenter is Grounding DINO with SAM 3 in zero shot mode, which carries no in-sample fitting specific to CUB, so no cross-fitting step is required. We use the official $5{,}994/5{,}794$ split, matching the loader that produced the screening number so that benefit and DiD are comparable.

Table 11: CUB-200-2011 crossing with padded ROI crop using six seeds. Cells are mean accuracy on the official test set. Intervals are 95% t intervals over seeds.

	Eval synthetic	Eval predicted	Image-only
Train synthetic	0.8412	0.8448	0.8408
Train predicted	0.8364	0.8607	0.8408
Estimand	**Point**	**Seed 95%**	
$\mathrm{SOG}_{\mathrm{eval}}(S)$	-0.36 pp	$[-0.64, -0.08]$	
$\mathrm{SOG}_{\mathrm{eval}}(R)$	-2.43 pp	$[-2.80, -2.07]$	
DiD	$+2.07$ pp	$[+1.53, +2.62]$	
$\mathrm{SOG}_{\mathrm{match}}$	-1.96 pp	$[-2.26, -1.66]$	
Δ_{syn}	$+0.03$ pp	$[-0.40, +0.47]$	
Δ_{pred}	$+1.99$ pp	$[+1.62, +2.37]$	

The accuracy and macro balanced accuracy DiDs over six seeds are both $+2.07$ pp; the corrected hierarchical bootstrap over seeds, classes, and images for the latter is $[+1.25, +2.94]$ pp. NLL gives the same substantive ordering: $\mathrm{DiD}_{\mathrm{NLL}} = -0.0540$ nats with a 95% CI over seeds $[-0.0631, -0.0449]$ and hierarchical CI $[-0.0771, -0.0322]$; lower NLL is better. The three strict conditions for an incorrect negative result were not all met because $\Delta_{\mathrm{syn}} = +0.03$ pp has an interval spanning zero. We therefore do not claim that synthetic validation significantly rejected ROI. What the extension establishes is more basic: ROI cropping with predicted masks is beneficial, the synthetic protocol does not detect that benefit, and the two contrasts matched by source utilities differ by 1.96 pp with an interval excluding zero. Unlike PlantSeg and BUSI, the CUB proxy is pessimistic.

Sensitivity to IoU matching. The predicted and synthetic test masks have mean IoU 0.8462 and 0.8493 at the classifier resolution. Their absolute residual has median 0.0089 and 95th percentile 0.0193; 134 of 5,794 images exceed the 0.02 tolerance. Removing those images leaves 5,660 and slightly increases DiD to $+2.10$ pp ($[1.56, 2.63]$; hierarchical $[1.27, 2.97]$). The corresponding

$\mathrm{SOG}_{\mathrm{match}}$ is -2.01 pp ($[-2.31, -1.71]$; hierarchical $[-2.81, -1.22]$). The result is therefore not produced by the unmatched 2.31%.

Mechanism through crop boxes. The synthetic mask's scattered false positives interact with the bounding box consumer. Boxes from predicted masks occupy a median 0.421 of the image (5th–95th percentiles 0.173–0.866), whereas the synthetic median and both percentiles are 1.000. Boxes exceeding 0.99 of the image occur for 54/5,794 predicted masks (0.93%) but 5,768/5,794 synthetic masks (99.55%). Predicted-mask crops contain less background (0.721 versus 0.870), while both sources retain essentially all annotated bird pixels and visible parts. This supports crop expansion and background inclusion as the relevant mismatch for this generator and consumer pair; the weak part retention correlation ($\rho = 0.042$) is descriptive, not mechanistic proof.

G PREREGISTERED BUSI REPLICATION

BUSI passed the screen for GT mask benefit by $+6.12$ pp. We therefore preregistered the same hard-zero crossed design with six seeds. OOF segmenter outputs from five folds on the 549 training images have mean IoU 0.6966; the deployment dump on the 98 validation images has mean IoU 0.6963. Synthetic masks are generated for each image to match those OOF targets with probabilistic mixing.

Table 12: Preregistered BUSI OOF crossing using six seeds. Cells are accuracy on the validation cohort of 98 images; intervals are 95% intervals over seeds.

	Eval synthetic	Eval predicted
Train synthetic	0.9592	0.8333
Train OOF predicted	0.8571	0.8963
Estimand	**Point**	**Seed 95%**
$\mathrm{SOG}_{\mathrm{eval}}(S)$	$+12.59$ pp	$[+9.43, +15.74]$
$\mathrm{SOG}_{\mathrm{eval}}(R)$	-3.91 pp	$[-5.99, -1.83]$
DiD	$+16.50$ pp	$[+13.93, +19.07]$
$\mathrm{SOG}_{\mathrm{match}}$	$+6.29$ pp	$[+5.04, +7.54]$

The image ID paired bootstrap for DiD is $[+9.52, +24.15]$ pp; 34 of 98 images have a nonzero contribution. Both evaluation effects reverse sign, reproducing the PlantSeg pattern under OOF provenance in a medical domain. The earlier crossing based on in-sample masks is retained only as a provenance audit:

Table 13: BUSI provenance audit. OOF closes the segmenter IoU gap between training and validation while leaving the crossed interaction nearly unchanged.

Training mask provenance	Train IoU	Val. IoU	DiD	Seed 95%
In-sample	0.8578	0.6963	$+17.69$ pp	$[+13.11, +22.26]$
Five-fold OOF	0.6966	0.6963	$+16.50$ pp	$[+13.93, +19.07]$

Removing the 16.2 pp IoU gap between training and validation changes DiD by only -1.19 pp, so in-sample segmenter provenance is not the main explanation for the large interaction. In the in-sample audit, $\mathrm{SOG}_{\mathrm{eval}}(S) = +12.42$ pp, $\mathrm{SOG}_{\mathrm{eval}}(R) = -5.27$ pp, and $\mathrm{SOG}_{\mathrm{match}} = +3.40$ pp, each with an interval over seeds interval excluding zero. Two limitations remain. BUSI has no independent test; the same validation cohort of 98 images selects checkpoints and supplies the crossed evaluation, so this is a second domain replication rather than an independent external validation. In addition, probabilistic mixing produces less coherent errors than the segmenter. The magnitude is therefore a large, protocol specific estimate, not a general bound.

H PREREGISTERED RIM-ONE DL BOUNDARY ANALYSIS

RIM-ONE DL tests the audit under a protocol that keeps hospitals disjoint and uses regular optic disc masks of high quality. HUC supplies OOF training predictions; HUMS and r1 remain held out for deployment evaluation. The test deployment IoU is 0.913, closely aligned with the HUC OOF value of 0.918. The analysis plan names balanced accuracy as the primary endpoint and retains the frozen fusion, generator, seeds, and decision rule after the test set is opened.

Table 14: Source crossing estimates for RIM-ONE DL. Balanced accuracy is primary; accuracy and AUROC are secondary. The stratified interval resamples hospital and class.

Metric	Analysis	DiD	95% CI
Balanced accuracy	primary over seeds	$+1.93$ pp	$[-2.16, +6.03]$
Balanced accuracy	hospital$\times$class bootstrap	$+1.93$ pp	$[-4.90, +8.13]$
Accuracy	secondary	$+0.29$ pp	$[-4.04, +4.62]$
AUROC	secondary	$+1.96$ pp	$[+1.51, +2.40]$

Both primary intervals include zero, so this experiment does not establish a source interaction and cannot support equivalence. The AUROC interval excludes zero, but it cannot replace the frozen primary endpoint. Accuracy DiD also differs in sign across the HUMS and r1 strata (-0.44 and $+0.85$ pp); because there are only two hospital strata, this is reported as heterogeneity description rather than a subgroup conclusion. RIM-ONE therefore marks a boundary of the current evidence at high IoU.

I SPATIAL REMOVAL CONTROLS

Three controls further delimit the sensitive regime. In the original July same cohort extension, the accuracy difference between OOF and in-sample masks was $+1.39$ pp for hard-zero concat, $+0.20$ pp for ROI crop, and -0.86 pp for mask channel; the preregistered fusion granularity ordering therefore failed. That $+1.39$ pp is one replicate of the pooled $+1.10$ pp reported below and should not be read as a separate estimate. The remaining two controls are anchored on the August cohort of three seeds and so use its $\mathrm{SOG}_{\mathrm{eval}}(S) = +4.04$ pp rather than the $+4.06$ pp from six seeds. Hard-zero $\mathrm{SOG}_{\mathrm{eval}}(S)$ at mask resolutions 224, 56, 14, and 1 is $+4.04$, $+4.78$, $+3.84$, and $+1.42$ pp. The sequence is not monotonic and rejects a smooth resolution law. Instead, it shows an abrupt contrast between spatially selective and uniform masking. Third, filling removed pixels with the image mean improves accuracy under predicted masks by $+0.77$ pp but changes $\mathrm{SOG}_{\mathrm{eval}}(S)$ only from $+4.04$ to $+3.97$ pp, failing the numerical prediction. Average accuracy and evaluation optimism can therefore move independently.

J VALIDATION AUDIT WITH THREE CORRUPTIONS

All three PlantSeg validation configurations have positive mean $\mathrm{SOG}_{\mathrm{match}}$ (Table 15A): $+2.47$ pp ($[1.00, 3.94]$) under erosion, $+3.10$ pp ($[0.58, 5.63]$) under probabilistic mixing, and $+2.56$ pp under dilation, whose interval $[-0.73, 5.86]$ pp is inconclusive. We report the three corruptions individually rather than pooling them, because three paired configurations from one segmenter do not support a pooled estimand that a reader could mistake for evidence across segmenters.

The direction extends to ViT-Base ($+3.96$ pp) and ResNet50 ($+2.83$ pp), giving three backbones across two architecture families, not three families. Segmenter generality is established separately, on frozen test with two segmenters used in zero shot mode, not here. The preregistered BUSI OOF crossing is reported separately in Appendix G; the historical BUSI audit with two cells is superseded.

The means in the control with three arms and six seeds are 0.7476 for in-sample masks from one model, 0.7467 for the in-sample mosaic from five models, and 0.7573 for OOF masks from five models. ResNet50 improves from 0.7053 to 0.7139 under OOF training ($+0.86$ pp; seed effects $+0.65, +1.57, +0.35$ pp) but remains 0.15 pp below image-only. ViT-Base improves by only

A. Corruption audit

Corruption	Mean	Paired-t 95% CI
Erosion	$+2.47$	$[+1.00, +3.94]$
Prob. mix	$+3.10$	$[+0.58, +5.63]$
Dilation	$+2.56$	$[-0.73, +5.86]$

B. Backbone audit

Backbone	Mean	Seed min–max
ConvNeXt-Tiny	$+3.10$	$[+1.93, +3.69]$
ResNet50	$+2.83$	$[+1.28, +4.33]$
ViT-Base	$+3.96$	$[+3.77, +4.17]$

Table 15: Protocol gap matched by source, $\mathrm{SOG}_{\mathrm{match}}$ (pp), on PlantSeg validation with hard-zero concat. Panel A reports paired-t intervals; panel B reports ranges over seeds, not confidence intervals. The arm trained on predicted masks here uses in-sample predictions; Section 5.1 reports crossing on the frozen test set with both arms using targets derived from OOF predictions.

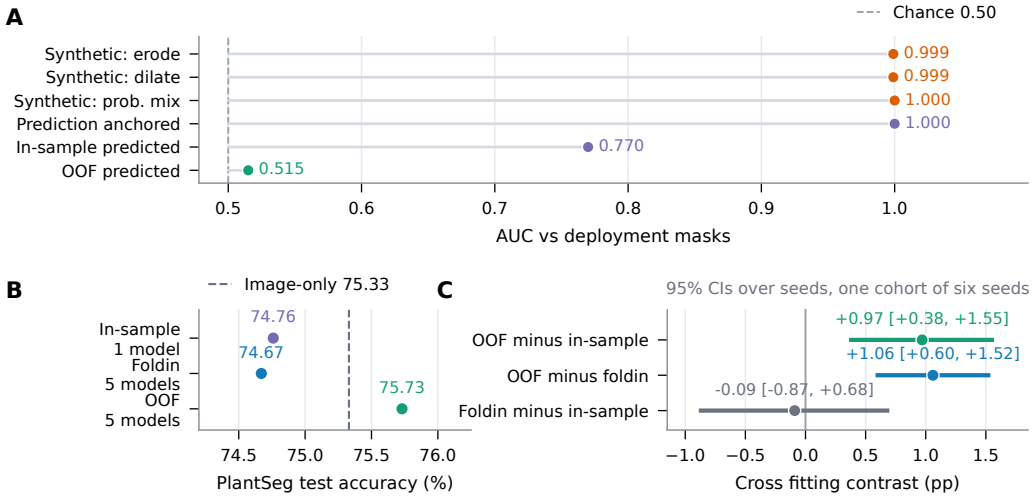


Figure 5: PlantSeg evidence from cross fitting and the diversity control. (A) Source discrimination from seven statistics beyond IoU; joint AUC saturates for all displayed synthetic arms. (B) ConvNeXt means over six seeds for in-sample masks from one model, in-sample masks from five models (foldin), and OOF masks from five models; the dashed line is the contemporaneous image-only baseline. (C) All three contrasts share one cohort of six seeds and one interval type over seeds, so they can be read against each other. Segmenter diversity without provenance correction contributes -0.09 pp.

$+0.10$ pp with mixed seed signs despite 1.48 pp of headroom relative to the image-only baseline. These architecture results delimit the cross-fitting claim rather than establishing a universal mitigation.

K Robustness Summary

Table 16: Each row removes one alternative explanation for the interaction. Intervals are calculated over six seeds unless noted.

Check	Rules out	DiD	Seed 95%
Primary: OOF targets, OOF predicted arm	—	$+6.28$ pp	$[+5.31, +7.26]$
Targets from training predictions	training IoU leakage	$+6.16$ pp	$[+4.99, +7.33]$
Predicted arm from training predictions	mask quality leakage	$+5.84$ pp	$[+4.32, +7.36]$
Overlap matched, 90.8%	residual IoU mismatch	$+5.86$ pp	$[+5.29, +6.44]$
Balanced accuracy, sequence clusters	weighting by long tail	$+5.95$ pp	$[+4.85, +7.90]^*$
CLIPSeg, zero shot	segmenter fitting leakage	$+10.56$ pp	$[+8.71, +12.41]$
GDINO+SAM3, zero shot	segmenter fitting leakage	$+13.51$ pp	$[+11.95, +15.08]$

*Widest of twelve intervals over three metrics and four resampling units; see Appendix P. Rows use the primary arm with OOF targets where applicable; the zero shot rows use IoU targets specific to each segmenter, and the two provenance rows intentionally substitute one arm.

L Replication with Zero Shot Segmenters

Both sources are used off the shelf without fine-tuning on PlantSeg. CLIPSeg is `CIDAS/clipseg-rd64-refined`; probabilities for each pixel are taken as the maximum over the prompts `disease lesion`, `diseased area on a leaf`, `leaf lesion`, `brown spot` and `necrotic tissue`, then thresholded at 0.5. Grounding DINO is `IDEA-Research/grounding-dino-base` with box and text thresholds 0.25 and 0.20 and at most five boxes, prompted with {`disease`} `lesion`, {`disease`} `spot` and `diseased lesion`; SAM 3 consumes the first prompt and we take the union of masks scoring at least 0.5. Images yielding no box or no mask are kept as empty masks (CLIPSeg 314/59/92 and Grounding DINO+SAM3 265/37/67 on train/validation/test). Model revisions, checksums, the generating command and a hash of the mask array are written into each dump, and packaging asserts that the recorded source matches the stored array.

For each segmenter the synthetic arm is matched per image to that segmenter's own IoU; the SegFormer synthetic masks are not reused. Both arms use hard-zero concat with ConvNeXt-Tiny and six seeds.

Table 17: Crossed design on the frozen test set for two segmenters in zero shot mode, using six seeds. Cells are accuracies; every estimand is reported in percentage points with its 95% interval over seeds. All eight intervals exclude zero.

	CLIPSeg	GDINO+SAM3
A_{SS}	0.7756	0.7810
A_{SR}	0.6957	0.7130
A_{RS}	0.7187	0.6972
A_{RR}	0.7443	0.7644
$\mathrm{SOG}_{\mathrm{eval}}(S)$	$+8.00$ pp	$+6.79$ pp
seed 95%	$[+6.25, +9.75]$	$[+5.50, +8.08]$
$\mathrm{SOG}_{\mathrm{eval}}(R)$	-2.56 pp	-6.72 pp
seed 95%	$[-3.85, -1.28]$	$[-7.50, -5.94]$
DiD	$+10.56$ pp	$+13.51$ pp
seed 95%	$[+8.71, +12.41]$	$[+11.95, +15.08]$
$\mathrm{SOG}_{\mathrm{match}}$	$+3.13$ pp	$+1.66$ pp
seed 95%	$[+2.56, +3.70]$	$[+0.75, +2.56]$

DiD by seed is $+10.99$, $+13.91$, $+9.42$, $+9.15$, $+9.68$, $+10.24$ pp for CLIPSeg and $+12.90$, $+15.48$, $+11.86$, $+12.55$, $+15.26$, $+13.03$ pp for Grounding DINO+SAM3. Only probabilistic mixing was run for these segmenters; erosion and dilation arms were not, so these results extend the segmenter axis but not the corruption axis.

M Subset with Matched Overlap

Table 18: DiD on the frozen test set, restricted to images whose synthetic and deployment masks agree in IoU for each image. Both intervals are 95%.

Subset	n	Share	DiD	Image 95%	Seed 95%
All test images	2,294	100%	$+6.28$ pp	$[+5.40, +7.18]$	$[+5.31, +7.26]$
$\lvert\Delta\mathrm{IoU}\rvert \le 0.05$	2,156	94%	$+5.97$ pp	$[+5.09, +6.85]$	$[+5.30, +6.64]$
$\lvert\Delta\mathrm{IoU}\rvert \le 0.02$	2,083	90.8%	$+5.86$ pp	$[+4.98, +6.77]$	$[+5.29, +6.44]$
$\lvert\Delta\mathrm{IoU}\rvert \le 0.01$	1,145	50%	$+5.28$ pp	$[+4.18, +6.42]$	$[+4.27, +6.30]$

Point estimate drift from the full test set to the 0.02 subset is -0.43 pp for $\mathrm{SOG}_{\mathrm{eval}}(S)$, -0.01 pp for $\mathrm{SOG}_{\mathrm{eval}}(R)$, -0.42 pp for DiD and -0.19 pp for $\mathrm{SOG}_{\mathrm{match}}$. Every interval over seeds excludes zero at every tolerance, so discarding the tenth of the test set with the poorest matching leaves the interaction intact. The full set median $|\Delta\mathrm{IoU}|$ is 0.010 with 95th percentile 0.071, consistent with the matching residuals in Table 23.

N CROSSED DESIGN BY FUSION MODE

Hard-zero concat uses six seeds; the other three modes use three. OOF is the primary reading.

Table 19: Crossed design by fusion mode and predicted mask provenance. Hard-zero uses six seeds; the other modes use three. All rows use a synthetic arm matched to in-sample predictions so that the four fusion modes are comparable with each other; the hard-zero row therefore reads $+6.16$ pp rather than the primary $+6.28$ pp, a difference of $+0.12$ pp that does not affect the ordering.

Fusion	Provenance	$\mathrm{SOG}_{\mathrm{eval}}(S)$	$\mathrm{SOG}_{\mathrm{eval}}(R)$	DiD	DiD seed 95%
Soft concat	OOF	$+4.45$	-4.00	$+8.44$	$[+4.81, +12.08]$
Soft concat	in-sample	$+4.45$	-2.72	$+7.16$	$[+6.63, +7.70]$
Hard-zero concat	OOF	$+3.94$	-2.22	$+6.16$	$[+4.99, +7.33]$
Hard-zero concat	in-sample	$+3.94$	-1.90	$+5.84$	$[+4.32, +7.36]$
ICGD	OOF	$+0.99$	-0.04	$+1.03$	$[+0.34, +1.73]$
ICGD	in-sample	$+0.99$	-0.64	$+1.63$	$[+1.08, +2.17]$
UASMG	OOF	$+0.33$	-0.62	$+0.96$	$[+0.85, +1.07]$
UASMG	in-sample	$+0.33$	-0.96	$+1.29$	$[+0.95, +1.64]$

O CANDIDATE PROVENANCE FOR THE SELECTION EXPERIMENT

Table 20: Candidate provenance and validation ranking, with means over six seeds from a single checkpoint for each (method, seed). Synthetic validation ranks hard-zero concat first; deployment mask validation ranks it last among mask consumers, below not using masks at all.

Candidate	Training masks	Uses mask	Val (synthetic)	Val (deployment)
Image-only	none	no	0.7523	0.7523
Hard-zero concat	synthetic	yes	**0.7729**	0.7223
Soft concat	synthetic	yes	0.7729	0.7229
Mask channel	synthetic	yes	0.7625	0.7423
ROI crop	synthetic	yes	0.7491	0.7370
Feature attention	synthetic	yes	0.7379	0.7368
ICGD	synthetic	yes	0.7593	0.7451
UASMG	synthetic	yes	0.7530	**0.7474**

P CLASS BALANCE AND CLUSTERED RESAMPLING

PlantSeg has 115 classes with a long tail (minimum 1, median 16, maximum 77 test images). Overall accuracy can therefore be carried by a few frequent classes, and resampling images assumes an independence the split unit does not guarantee. We recompute every headline contrast under three metrics and four resampling units. The image stem prefix identifies an acquisition sequence, which is the closest unit to a plant or a collection episode that the metadata supports: 2,294 test images fall into 285 sequences, median 6 images each, maximum 58, with 79 singletons. Clusters of possible duplicates (cosine among test images ≥ 0.95, single linkage) barely merge anything: 2,239 clusters, 2,186 of them singletons, and a largest cluster of 3. They give intervals close to the independent case. Acquisition sequences are the binding unit and give the widest intervals, and are what we report.

Table 21: Headline contrasts under metrics balanced across classes, with OOF provenance and six seeds. Switching to balanced weighting moves DiD by about half a point.

Estimand	Overall accuracy	Balanced accuracy	Macro-F1
$\mathrm{SOG}_{\mathrm{eval}}(S)$	$+4.06$ pp	$+3.63$ pp	$+3.46$ pp
$\mathrm{SOG}_{\mathrm{eval}}(R)$	-2.22 pp	-2.32 pp	-2.37 pp
DiD	$+6.28$ pp	$+5.95$ pp	$+5.83$ pp
$\mathrm{SOG}_{\mathrm{match}}$	$+2.61$ pp	$+2.12$ pp	$+2.16$ pp

All twelve DiD intervals (three metrics $\times$ four resampling units) exclude zero; the widest is acquisition sequence clustering on overall accuracy, $[+4.85, +7.90]$ pp. $\mathrm{SOG}_{\mathrm{eval}}(S)$, $\mathrm{SOG}_{\mathrm{eval}}(R)$ and $\mathrm{SOG}_{\mathrm{match}}$ are likewise $12/12$.

A decomposition by class answers whether frequent classes drive the result. DiD is positive for 84 of 115 classes, with median $+4.71$ pp and interquartile range $[+0.00, +9.15]$; $\mathrm{SOG}_{\mathrm{eval}}(S)$ is positive for 77. The five largest classes hold 14.0% of the test images; removing them moves DiD from $+6.28$ to $+6.52$ pp, that is, slightly up. The effect is therefore not an artefact of frequent classes.

Q Selection Protocol

Protocol A is the primary specification. One checkpoint exists per (method, seed); its epoch is chosen once, on synthetic validation. That single checkpoint is then scored on validation under both mask sources, so the two selection arms differ only in which method they rank first and the regret measures method choice alone.

```
# masks: synthetic = prob_mix matched in IoU
#        predicted = SegFormer deployment predictions
for m in candidates:                    # 8 consumers
    for s in seeds:                     # 6 seeds
        # the image only reference ignores masks entirely
        train(m, s, masks = synthetic)
        ckpt[m,s]   = argmax_epoch Acc(val, synthetic)
        v_syn[m,s]  = Acc(val,  ckpt[m,s], masks = synthetic)
        v_pred[m,s] = Acc(val,  ckpt[m,s], masks = predicted)
        t_pred[m,s] = Acc(test, ckpt[m,s], masks = predicted)

pick     = argmax_m mean_s v_syn[m,s]   # synthetic choice
dep_pick = argmax_m mean_s v_pred[m,s]  # deployment choice
regret = mean_s t_pred[dep_pick,s] - mean_s t_pred[pick,s]
# frozen test is read once, only to pay the regret.
```

Protocol B is diagnostic only. There the deployment mask arm also reselects the epoch, which requires separate runs validated on deployment masks. Those runs were never exported on frozen test, so B can rank methods but cannot pay for them at matching provenance. B selects ICGD rather than the image-only reference and yields a mixed provenance $+1.66$ pp. We report B as a ranking check and do not treat it as a bound on A. Neither arm is an oracle: the optimum on the frozen test set is UASMG, which neither validation source selects.

An earlier version of this experiment read the deployment validation column from a separate set of runs that had been validated on deployment masks. That made the oracle's identity and the price it paid come from different checkpoints. The present numbers evaluate one checkpoint under both mask sources and supersede it.

Reselecting the method within each seed gives $+1.69$ pp regret (95% CI over seeds $[0.22, 3.15]$). Across all 247 candidate subsets of size at least two, 225 incur positive regret, 13 tie, and 9 are negative by at most 0.49 pp. Dropping both concat methods reduces but does not eliminate regret ($+0.24$ to $+1.50$ pp). Because subsets overlap and share checkpoints, this is a candidate pool sensitivity analysis rather than 247 replications.

R GENERATOR AND SEGMENTER SPECIFICATIONS

Table 22: Segmenter quality for each image. Dice is $2\,\mathrm{IoU}/(1+\mathrm{IoU})$, exact for binary masks. In-sample training masks are far better than deployment masks; training masks obtained by cross fitting match the deployment distribution.

Segmenter	Split	n	Mean IoU	Median	p05	IoU $= 0$	IoU < 0.5
SegFormer	train	7,916	0.747	0.799	0.344	0.6%	11.3%
SegFormer	val	1,247	0.591	0.643	0.077	1.1%	33.5%
SegFormer	test	2,294	0.604	0.657	0.081	1.4%	31.1%
SegFormer OOF	train	7,916	0.600	0.647	0.079	1.0%	32.2%

Validation and test are outside segmenter training, so their OOF rows coincide with the deployment rows and are omitted.

For the primary source crossing, OOF predictions define a mean training IoU target of 0.6001 and the realized synthetic masks average 0.6102, compared with 0.747 for in-sample SegFormer predictions. The validation and test mask dumps are bitwise identical between the two training provenance arms, so the crossed manipulation changes training masks only.

The three generator families are not interchangeable and should not be described in one sentence, nor collectively called matched in IoU for each image. Probabilistic mixing approximately matches IoU for each image; erosion and dilation are discrete corruptions that target IoU but frequently miss it; the ladder matches expected overlap analytically. Erosion and dilation search a discrete grid of structuring element sizes, so more than half of the images cannot reach their IoU target at all, and erosion can return an empty mask. Probabilistic mixing bisects a continuous blend strength for 24 steps against a tolerance of 0.02. The ladder arms solve the target with an explicit formula, $t^* = G(1-\tau)/(0.6G + 0.1\tau B)$ for foreground area G, background area B, and target IoU τ. This equalizes the *expected* overlap across rungs before the mask is discretely realized; realized residuals are reported in Table 23 and are not zero.

Table 23: Generator fidelity to the IoU target for each image, tolerance 0.02.

Generator	Split	Median resid.	p95	Within tol.	Resid. > 0.05	Empty
Probabilistic mix	train	0.0093	0.0218	92.4%	0.5%	0%
Probabilistic mix	val	0.0105	0.0595	87.2%	5.6%	0%
Probabilistic mix	test	0.0100	0.0713	88.1%	6.1%	0%
Ladder rungs	test	0.0015	0.0771	90.5%	6.4%	0%
Erosion	all	0.025–0.028	0.13–0.22	37.8–43.5%	28–31%	0.5–1.3%
Dilation	all	0.020–0.022	0.07–0.39	47.2–50.7%	12–29%	0%

No split contains an empty ground-truth mask, so the branch for an empty target in the matching code (return the input unchanged and record IoU $= 1.0$) is never exercised on this data.

S ANALYSIS STATUS LEDGER

Table 24: Status boundaries retained in the paper.

Analysis	Status
Three SegFormer and corruption combinations	Validation $\mathrm{SOG}_{\mathrm{match}}$; hard-zero concat
SOG with fixed source, source crossing, selection	Canonical frozen logits from August
OOF versus in-sample	Pooled estimate over six July/August runs
OOF versus foldin	Cohort with six seeds; interval over seeds
Initial in-sample versus OOF observation	Post hoc finding from code audit
Fusion, structure, resolution, CNN/ViT	Declared before execution; failures retained
Mean fill experiment	Exploratory; declared after resolution result
Fragmentation ladder	Five arms; trend interval over seeds
Exclusion of possible duplicates	Post hoc audit; thresholds disclosed
BUSI OOF crossing with five folds	Preregistered before training; six seeds; no independent test
CUB crossing with six seeds	P15 extension; seeds 0–2 previously observed, seeds 3–5 prospectively frozen
RIM-ONE DL crossing	Preregistered test with hospitals kept disjoint; primary balanced accuracy interval includes zero
BUS-BRA gate for GT benefit	Preregistered validation screen failed; test disjoint by patient remains unopened
Severstal forecast for GT benefit	Preregistered before training; failed prediction retained