Supplementary Information

Vendor-Aware Three-Dimensional Missing-Tooth Localization in Partially Annotated Multi-Vendor Cone-Beam Computed Tomography: A Leakage-Controlled Evaluation of Semi-Supervision

Hana Asgari*[1], Faraneh Zarafshan†[2,3], Ahmadreza Talaeipoor‡[4], and Elham Mazinan§[1]

[1]Department of Computer, Ar.C., Islamic Azad University, Arak, Iran
[2]Department of Computer Engineering, Ka.C., Islamic Azad University, Karaj, Iran
[3]Institute of Artificial Intelligence and Social and Advanced Technologies, Ka.C., Islamic Azad University, Karaj, Iran
[4]Department of Oral and Maxillofacial Radiology, Faculty of Dentistry, Tehran Medical Sciences, Islamic Azad University, Tehran, Iran

Supplementary Note 1: Integrity and analysis lock

The source audit identified 158 scans and 114 released three-dimensional boxes. Meyer M38 and M61 were exact pixel duplicates with discordant released annotations and were prospectively excluded together. The retained cohort comprised 83 labeled scans with 111 objects (Kavo 14, Meyer 60, Newtom 9) and 73 scans without released box annotations (Kavo 21, Meyer 32, Newtom 20). All data roles were assigned at case level. Unlabeled scans were never coded as negative cases. Calibration and outer-test partitions were not accessed during model development, and the locked outer test was evaluated once after threshold selection in the dedicated fold-specific calibration partitions.

Table 1: Case roles in each outer fold.

Fold	Optimization	Validation	Calibration	Outer test
0	46	10	10	17
1	46	10	10	17
2	46	10	10	17
3	47	10	10	16
4	47	10	10	16

Supplementary Note 2: Repeated-seed analysis

Five new repeats were completed for each of five folds and three methods, yielding 75 of 75 successful formal runs. No seed was removed or replaced based on performance. Calibration and outer-test data were not accessed during repeated-seed training or selection. Dense ungated semi-supervision was not included in the 75-run audit and remains the initial completed negative-transfer ablation.

*Email: hana.asgari@iau.ir
†Correspondence: Fa.zarafshan@iau.ac.ir
‡Email: dr.artalaipour@gmail.com
§Email: Elham.mazinan@iau.ac.ir

Table 2: Repeated-seed developmental pFROC. Each method contains 25 fold–repeat outcomes.

Method	Mean	SD	Median	Q1	Q3	Minimum	Maximum
Pure supervised	0.890	0.087	0.923	0.821	0.954	0.679	1.000
Vendor-adversarial	0.899	0.086	0.923	0.804	0.972	0.714	1.000
Reliability-gated	0.872	0.106	0.923	0.786	0.945	0.625	1.000

Table 3: Matched fold–repeat comparisons. Wins, ties, and losses refer to the first named method relative to the second.

Comparison	Mean difference	SD	Median	Wins	Ties	Losses
Vendor − Pure	0.009	0.042	0.000	11	8	6
Gated − Pure	−0.019	0.035	0.000	2	14	9
Gated − Vendor	−0.028	0.065	−0.018	4	8	13

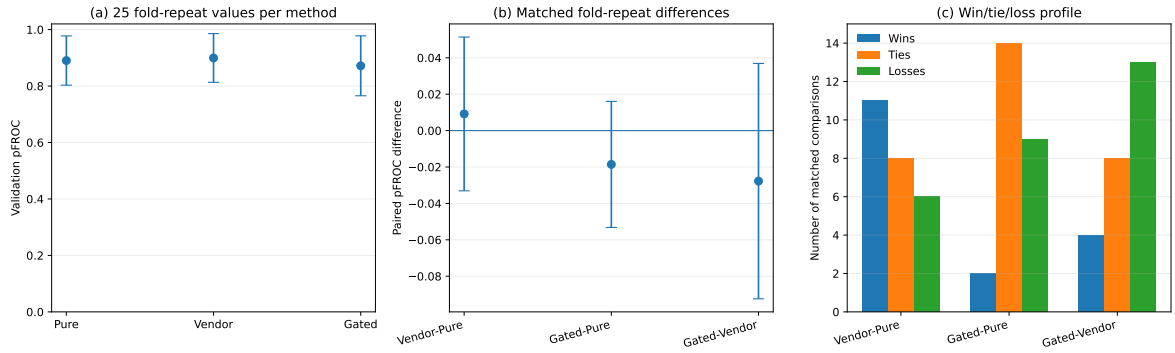

Figure 1: Repeated-seed developmental results. Fold–repeat observations reuse the same underlying patient cohort and are not independent clinical samples.

Supplementary Note 3: Post-lock detector benchmarks

The separately locked revision protocol specified a conventional single-scale 3D detector and a confidence-only Mean-Teacher-3D control after the original outer-test results were known. Both controls used the immutable development manifests, five repeats, five folds, and matched seeds. All 50 formal runs completed and passed checksum audit without calibration or outer-test access.

Table 4: Post-lock exploratory validation benchmarks.

Method	n	Mean	SD	Minimum	Maximum	Mean time, s
Conventional 3D	25	0.921	0.065	0.821	1.000	688.08
Mean-Teacher-3D	25	0.886	0.078	0.750	1.000	688.56

Table 5: Paired post-lock comparison. The exact sign-flip test uses five fold-level mean differences as independent blocks; the run-level Wilcoxon result is secondary.

Measure	Result
Contrast	Mean-Teacher-3D minus conventional 3D
Mean paired difference	-0.035
Hierarchical bootstrap 95% CI	-0.070 to -0.00005
Teacher better/tied/worse	3/9/14
Primary exact fold-blocked test	$p = 0.250$ (five independent blocks)
Secondary run-level Wilcoxon test	$p = 0.003$ (dependent repeats; supportive only)

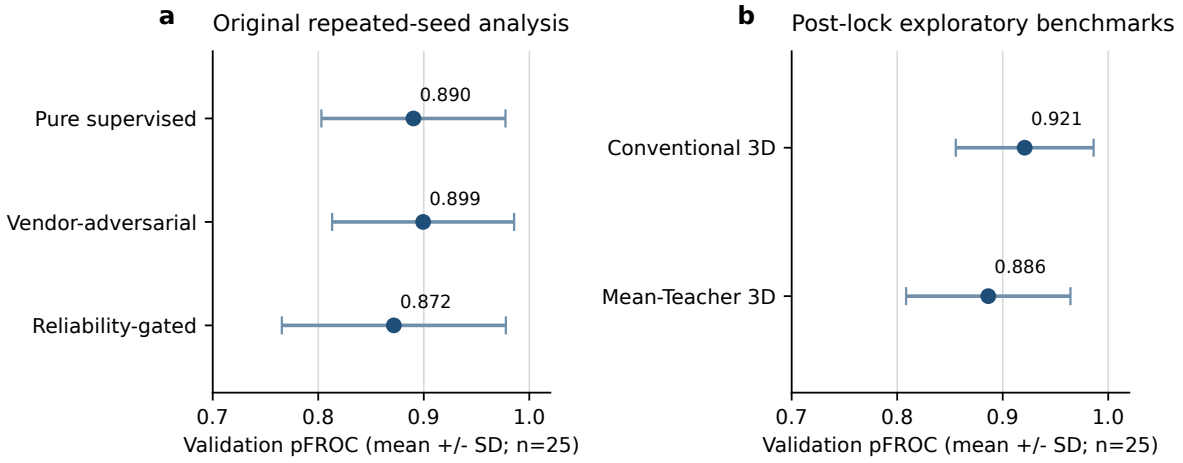

Figure 2: Original repeated-seed results and post-lock exploratory benchmarks. Error bars denote standard deviations, not confidence intervals.

Supplementary Note 4: Gate-component ablation

Ten new post-lock validation-only training runs compared confidence-only and confidence-plus-uncertainty gates across the five immutable folds. Five existing full-gate checkpoints were reused. Fifteen pseudo-label quality evaluations were completed without calibration or outer-test access. At the common confidence threshold 0.70, the full gate increased mean precision while reducing coverage. Uncertainty alone did not change downstream pFROC, and the full-gate pFROC difference was not significant in the exact five-block test.

Table 6: Exploratory cumulative gate ablation. Values are means across five folds.

Gate	pFROC	Pseudo precision	Target coverage	P–C AUC
Confidence only	0.891	0.785	0.710	0.525
Confidence + uncertainty	0.891	0.723	0.696	0.532
Full gate	0.898	0.826	0.680	0.527

Table 7: Fold-level downstream pFROC in the gate ablation.

Gate	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4
Confidence only	0.911	0.768	0.893	0.981	0.904
Confidence + uncertainty	0.911	0.768	0.893	0.981	0.904
Full gate	0.929	0.786	0.893	0.981	0.904

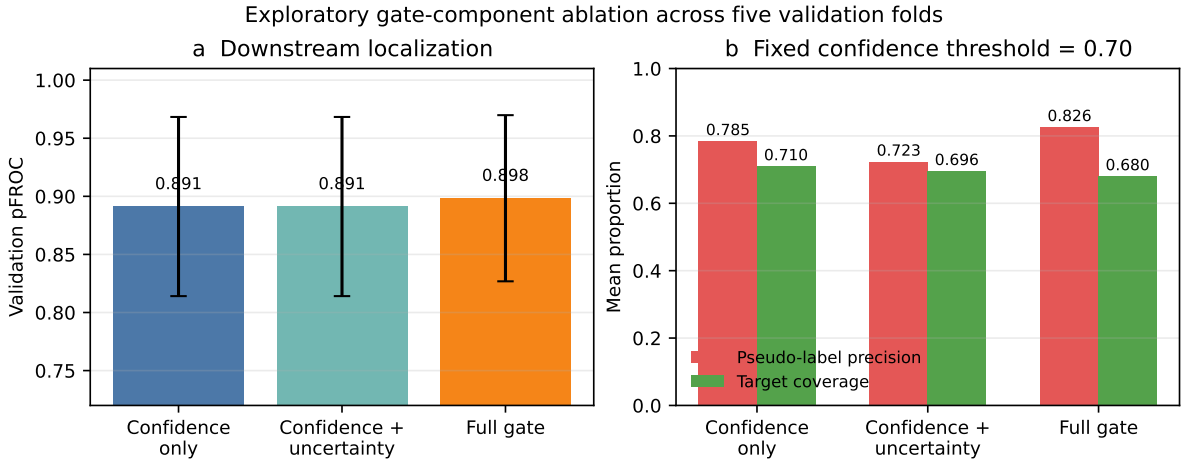

Figure 3: Gate-component ablation and pseudo-label precision–coverage trade-off at confidence 0.70. Error bars in panel a are fold-level standard deviations.

Supplementary Note 5: Locked outer-test operating point

Within each fold and method, the decision threshold maximized calibration sensitivity subject to no more than one false positive per volume. Each calibration partition contained 10 cases. Thresholds were then applied unchanged to outer-test partitions containing 17, 17, 17, 16, and 16 cases. Pooled estimates therefore used 83 case-level outer-test evaluations and 111 target objects. Confidence intervals were produced with 2,000 case-level bootstrap replicates. This procedure calibrated decision operating points rather than probabilities; temperature scaling, ECE, Brier score, and NLL improvement were not evaluated.

Table 8: Locked outer-test summary.

Method	Sensitivity	FP/volume	TP	FP	FN
Pure supervised	0.757	0.602	84	50	27
Vendor-adversarial	0.793	0.458	88	38	23
Dense ungated SSL	0.748	0.530	83	44	28
Reliability-gated	0.766	0.518	85	43	26

Table 9: Fold-specific thresholds selected on the disjoint calibration partitions.

Method	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4
Pure supervised	0.762	0.770	0.641	0.674	0.741
Vendor-adversarial	0.625	0.737	0.700	0.762	0.820
Dense ungated SSL	0.595	0.668	0.775	0.716	0.754
Reliability-gated	0.508	0.672	0.689	0.517	0.727

Supplementary Note 6: Computational profile

Evaluation of four methods across five folds required 640.5 s (10.7 min). Profiling used an NVIDIA GeForce RTX 3080 Ti with 11.63 GB VRAM, PyTorch 2.7.0, CUDA 12.8, mixed-precision FP16, batch size one, gradient checkpointing, and three sequential teacher passes. The first profiled call included warm-up and is not interpreted as steady-state latency.

Table 10: Patch-level implementation profile. Timings are not complete-volume clinical latency.

Patch $D \times H \times W$	Allocated memory, GB	Reserved memory, GB	Time, s
$64 \times 96 \times 96$	0.189	0.209	9.694 (warm-up)
$80 \times 112 \times 112$	0.273	0.299	0.384
$96 \times 128 \times 128$	0.382	0.439	0.288
$112 \times 144 \times 144$	0.531	0.604	0.340
$128 \times 160 \times 160$	0.723	0.781	0.347

Supplementary Note 7: Reproducibility artefacts

The submission archive contains the manuscript source, configuration files, summary tables, result figures, and SHA-256 manifests. Before submission, the full repeated-seed evidence archive should be verified against the following SHA-256 value:

```
e9e572fa39aa4827c1a768d81b822d41
eaa6af733078bfcac48eeed7ab7f0b27
```

A permanent public repository identifier for de-identified derived outputs and analysis code must be inserted into the Data Availability statement before acceptance.

Supplementary Note 8: Post-lock exploratory strict LOVO evaluation

The strict LOVO analysis was executed only after the original outer-test lock. One vendor was held out at a time, and both labeled and unlabeled cases from that vendor were excluded from source training. The compared methods were Pure supervised, vendor-adversarial, and full reliability-gated training. Five fixed repeats produced 45 formal held-out evaluations. Source calibration produced 45 frozen operating thresholds before the first target access. The target vendor was never used for hyperparameter, checkpoint, normalization, or operating-threshold selection.

Table 11: Strict LOVO source/target composition.

Held-out	Optimization	Validation	Calibration	Target cases	Source unlabeled	Target objects
Kavo	49	10	10	14	52	32
Meyer	17	3	3	60	41	68
Newtom	52	11	11	9	53	11

Table 12: LOVO descriptive summary with case-level bootstrap 95% intervals. Mean±SD is across five stochastic repeats; CIs are for the five-seed mean under 2,000 case-level bootstrap replicates.

Held-out	Method	Sensitivity	FP/volume	pFROC	Median center error, mm
Kavo	Pure	0.394 ± 0.075 [0.267, 0.512]	1.557 ± 0.559 [1.057, 2.043]	0.395 ± 0.049 [0.286, 0.504]	2.763 ± 0.352 [2.359, 3.124]
	Vendor	0.381 ± 0.092 [0.255, 0.494]	1.757 ± 0.577 [1.229, 2.286]	0.372 ± 0.036 [0.266, 0.482]	2.612 ± 0.101 [2.221, 3.047]
	Gated	0.369 ± 0.087 [0.237, 0.500]	1.014 ± 0.523 [0.657, 1.386]	0.398 ± 0.043 [0.280, 0.520]	2.584 ± 0.153 [2.399, 2.947]
Meyer	Pure	0.385 ± 0.181 [0.303, 0.467]	0.090 ± 0.072 [0.047, 0.147]	0.786 ± 0.046 [0.706, 0.855]	2.212 ± 0.075 [1.939, 2.480]
	Vendor	0.465 ± 0.186 [0.378, 0.545]	0.100 ± 0.095 [0.053, 0.157]	0.799 ± 0.046 [0.719, 0.874]	2.327 ± 0.065 [2.064, 2.543]
	Gated	0.412 ± 0.191 [0.335, 0.490]	0.187 ± 0.298 [0.123, 0.260]	0.734 ± 0.058 [0.657, 0.805]	2.227 ± 0.165 [1.968, 2.474]
Newtom	Pure	0.800 ± 0.119 [0.655, 1.000]	0.689 ± 0.372 [0.267, 1.178]	0.864 ± 0.066 [0.723, 1.000]	1.891 ± 0.229 [1.114, 2.947]
	Vendor	0.818 ± 0.064 [0.617, 1.000]	0.467 ± 0.363 [0.156, 0.822]	0.914 ± 0.019 [0.762, 1.000]	2.009 ± 0.105 [1.182, 2.883]
	Gated	0.800 ± 0.076 [0.617, 1.000]	0.556 ± 0.333 [0.222, 0.956]	0.855 ± 0.063 [0.700, 1.000]	1.978 ± 0.088 [1.194, 3.016]

Table 13: Per-seed strict LOVO held-out results. TP/FP/FN are exact counts at the frozen source-calibration-selected operating threshold.

Held-out	Method	Repeat	Seed	Sens.	FP/vol	TP/FP/FN	pFROC	Center, mm
Kavo	Pure	1	34026	0.312	1.071	10/15/22	0.383	2.509
Kavo	Pure	2	35026	0.344	1.857	11/26/21	0.320	2.671
Kavo	Pure	3	36026	0.375	0.857	12/12/20	0.453	2.429
Kavo	Pure	4	37026	0.438	1.857	14/26/18	0.406	2.904
Kavo	Pure	5	38026	0.500	2.143	16/30/16	0.414	3.304
Kavo	Vendor	1	34026	0.469	2.214	15/31/17	0.398	2.649
Kavo	Vendor	2	35026	0.344	2.143	11/30/21	0.320	2.534
Kavo	Vendor	3	36026	0.250	0.857	8/12/24	0.352	2.492
Kavo	Vendor	4	37026	0.375	1.500	12/21/20	0.406	2.641
Kavo	Vendor	5	38026	0.469	2.071	15/29/17	0.383	2.745
Kavo	Gated	1	54026	0.406	1.357	13/19/19	0.383	2.623
Kavo	Gated	2	55026	0.219	0.286	7/4/25	0.336	2.716
Kavo	Gated	3	56026	0.438	1.643	14/23/18	0.430	2.727
Kavo	Gated	4	57026	0.375	1.000	12/14/20	0.398	2.483
Kavo	Gated	5	58026	0.406	0.786	13/11/19	0.445	2.373
Meyer	Pure	1	34026	0.500	0.017	34/1/34	0.824	2.274
Meyer	Pure	2	35026	0.471	0.183	32/11/36	0.838	2.185
Meyer	Pure	3	36026	0.500	0.117	34/7/34	0.743	2.214
Meyer	Pure	4	37026	0.382	0.117	26/7/42	0.735	2.287
Meyer	Pure	5	38026	0.074	0.017	5/1/63	0.790	2.101
Meyer	Vendor	1	34026	0.721	0.233	49/14/19	0.860	2.375
Meyer	Vendor	2	35026	0.279	0.050	19/3/49	0.761	2.228
Meyer	Vendor	3	36026	0.471	0.017	32/1/36	0.805	2.318
Meyer	Vendor	4	37026	0.294	0.033	20/2/48	0.746	2.321
Meyer	Vendor	5	38026	0.559	0.167	38/10/30	0.824	2.394
Meyer	Gated	1	54026	0.544	0.083	37/5/31	0.787	2.109
Meyer	Gated	2	55026	0.353	0.067	24/4/44	0.768	2.025
Meyer	Gated	3	56026	0.324	0.000	22/0/46	0.695	2.384
Meyer	Gated	4	57026	0.176	0.067	12/4/56	0.651	2.398
Meyer	Gated	5	58026	0.662	0.717	45/43/23	0.768	2.218
Newtom	Pure	1	34026	0.636	0.667	7/6/4	0.750	2.081
Newtom	Pure	2	35026	0.727	0.444	8/4/3	0.909	1.683
Newtom	Pure	3	36026	0.909	1.333	10/12/1	0.864	1.736
Newtom	Pure	4	37026	0.909	0.444	10/4/1	0.909	1.763
Newtom	Pure	5	38026	0.818	0.556	9/5/2	0.886	2.192

Held-out	Method	Repeat	Seed	Sens.	FP/vol	TP/FP/FN	pFROC	Center, mm
Newtom	Vendor	1	34026	0.818	0.333	9/3/2	0.932	1.927
Newtom	Vendor	2	35026	0.818	0.667	9/6/2	0.886	1.936
Newtom	Vendor	3	36026	0.727	0.222	8/2/3	0.932	2.112
Newtom	Vendor	4	37026	0.909	1.000	10/9/1	0.909	1.934
Newtom	Vendor	5	38026	0.818	0.111	9/1/2	0.909	2.136
Newtom	Gated	1	54026	0.727	1.111	8/10/3	0.750	2.059
Newtom	Gated	2	55026	0.818	0.444	9/4/2	0.841	1.945
Newtom	Gated	3	56026	0.818	0.444	9/4/2	0.886	1.970
Newtom	Gated	4	57026	0.909	0.556	10/5/1	0.909	1.851
Newtom	Gated	5	58026	0.727	0.222	8/2/3	0.886	2.064

The source-only operating-point constraint was sensitivity maximization subject to FP/volume ≤ 1 on source calibration data, with ties resolved by lower FP/volume and then higher threshold. Target FP/volume was not constrained after threshold transfer and could therefore exceed 1. Target pFROC was separately computed from a fixed 1,001-point threshold grid at FP/volume budgets 0.5, 1, 2, and 4 and did not alter the frozen operating threshold. Descriptive uncertainty used 2,000 case-level bootstrap replicates (seed 20260817). No LOVO pairwise superiority test was prespecified.

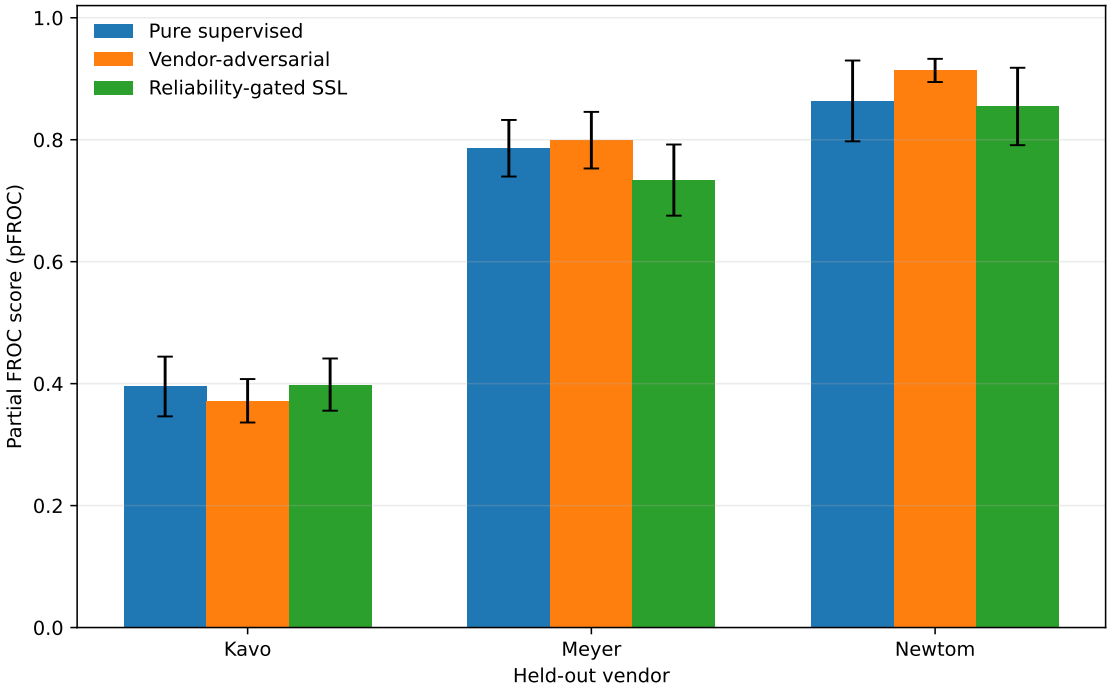

Figure 4: Post-lock strict LOVO mean pFROC with SD across five repeats.

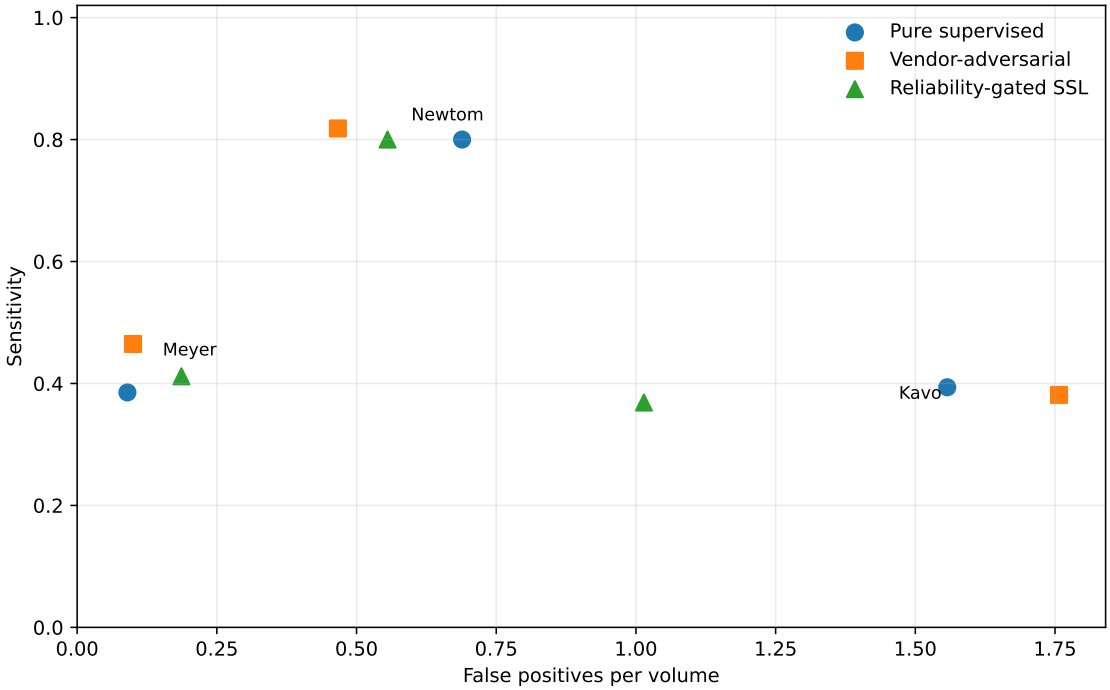

Figure 5: Post-lock strict LOVO mean sensitivity versus false positives per volume at frozen source-calibration-selected operating points.

Evidence lock. All 45 training reports, 45 source-calibration locks, 45 held-out evaluation files, 45 raw target-prediction files, 45 provenance files, and 45 threshold-count matrices passed the final evidence audit. The corresponding SHA-256 values are:

```
Training audit:  6c0f0b7fcde8f8cc03fd8f67d2105e28
                 4c3422017478238f839341c227e81336
Calibration master:  8b61310035a454986619e4aecf2d7734
                     773c192ad1488d4cc4c71875e2f0a517
Final LOVO summary:  75ec2305e4be46cbb83fcd156dcf1732
                     3f131acbcd63e231070861ce8a3cc27c
Final evidence manifest:  ddc035a362347958cd46048b4b5138c1
                          ec271d00e579d14eba12f8b06d93ead7
```

Machine-readable summaries are supplied under the supplementary directory:
`lovo_summary_v1.csv` and
`lovo_per_seed_v1.csv`.