Obsidian: Annotation-Efficient Wafer and SEM Defect Inspection - An Evidence-Aware Architecture, Evaluation Plan, and Integration Roadmap

Aaron Singh

***Abstract* - Obsidian adapts self-supervised and active-learning methods to wafer maps and scanning-electron-microscope imagery. The central result will be an accuracy-versus-annotation-cost curve under wafer-, lot-, or acquisition-isolated evaluation. Public data support research comparisons but do not establish production inspection performance. This paper defines the system model, evidence contract, risks, and staged validation plan. No physical-system or performance result is claimed without a committed reproducibility artifact.**

Index Terms - wafer inspection, SEM defects, segmentation, self-supervised learning, active learning

I. INTRODUCTION

Defect inspection must detect rare events while controlling the cost and consistency of expert annotation. Random image splits can overstate performance when neighboring samples share acquisition conditions. Obsidian therefore treats data provenance, group-isolated splitting, uncertainty, and annotation budget as first-class experimental variables.

Research contribution

The contribution is not a performance claim. It is a reproducible reference architecture, a falsifiable evaluation protocol, and an evidence schema that records inputs, decisions, outputs, failure modes, and claim boundaries.

II. TECHNICAL CONTEXT

Inspection performance can be inflated by random image splits, duplicate fields of view, or neighboring samples from one wafer. Obsidian makes acquisition group, wafer or lot identity, label budget, and uncertainty part of the experimental contract before training begins.

III. RESEARCH QUESTIONS

1 Can self-supervised pretraining reduce the required defect labels without losing recall?

2 Does k-center selection outperform random annotation?

3 Can calibrated uncertainty identify unseen defect morphologies?

IV. SYSTEM MODEL AND ARCHITECTURE

A provenance-aware registry feeds a self-supervised encoder. Classification and U-Net-style segmentation heads share the representation. An active selector ranks unlabeled samples, while an uncertainty layer flags ambiguous or out-of-distribution images for review.

Minimize expected defect error plus calibration error under a fixed annotation budget B, with train and test groups isolated by acquisition provenance.

The model is a design and test abstraction. It establishes invariants and expected-value checks; it does not convert simulated variables into measured hardware behavior.

V. DETAILED OPERATION AND RATIONALE

1 Build loaders and immutable split manifests for each public data set.

2 Establish supervised classification and segmentation baselines.

3 Pretrain on unlabeled images and compare label fractions.

4 Compare random and k-center annotation while tracking calibration and rare-class behavior.

VI. DATA AND EVIDENCE CONTRACT

Candidate data include Carinthia semiconductor SEM images, WM-811K wafer maps, and MVTec AD industrial images. Data-set licenses, class distributions, acquisition groups, and preprocessing steps must be recorded before training.

Each run must preserve a machine-readable configuration, random seed where applicable, source revision, environment description, raw result, summary method, and known limitations. Derived plots must point back to the exact records used to produce them.

VII. IMPLEMENTATION BLUEPRINT

Provenance registry. Tracks data origin, wafer groups, licenses, and transforms. Its acceptance artifact is immutable split manifest.

Representation encoder. Learns features from labeled and unlabeled imagery. Its acceptance artifact is checkpoint and training record.

Task heads. Classifies wafer maps and segments SEM defects. Its acceptance artifact is per-class predictions and masks.

Active selector. Ranks samples by diversity and uncertainty. Its acceptance artifact is annotation sequence and budget curve.

VIII. EXPERIMENTAL METHODOLOGY

Evaluation proceeds in four gates. Gate 1 checks deterministic behavior and invariants. Gate 2 applies controlled noise, tolerance, and injected-failure sweeps. Gate 3 repeats measurements and reports distributions rather than a favorable single run. Gate 4 introduces an external data source or bounded lab adapter only after the first three gates are reviewable.

Primary metrics are reported with sample counts, aggregation rules, uncertainty, and failure cases.

IX. REPRODUCIBILITY AND ACCEPTANCE CRITERIA

A result is accepted only when another clean environment can reproduce the reference behavior from a named revision. Every comparison uses identical budgets and held-out conditions. Failed runs remain part of the evidence record. Unknown hardware, calibration, and production conditions are labeled unknown rather than assigned synthetic values.

X. FAULT INJECTION AND NEGATIVE TESTING

Negative testing is part of the main evaluation rather than an appendix. Each risk is converted into an explicit scenario with an expected detection, rejection, abstention, or safe-state outcome. A silent success under missing or contradictory state is treated as a failed test.

Group leakage. The failure consequence is optimistic test scores. The required control is wafer- or acquisition-isolated splits.

Class imbalance. The failure consequence is rare defects are ignored. The required control is macro metrics and per-class recall.

Uncalibrated confidence. The failure consequence is novel defects pass unnoticed. The required control is calibration and abstention tests.

Domain shift. The failure consequence is public data do not match fab imagery. The required control is separate cross-domain reporting.

XI. ANALYSIS AND INTERPRETATION

Analysis reports distributions, failure counts, and representative traces rather than a single favorable score. Baselines share the same inputs, budgets, stopping rules, and held-out conditions. When multiple configurations are explored, the selection procedure is recorded so the test set is not used as a tuning signal.

The primary metric families are Macro F1 and per-class recall, AUPRC, Dice and IoU, Expected calibration error, Labels and annotation hours. Each is reported with its unit, aggregation rule, sample count, spread, and failed or undefined scenarios. An improvement is meaningful only when reference correctness is preserved and the result repeats across declared evaluation seeds or groups.

XII. INTEGRATION INTERFACES

Integration occurs through versioned records rather than hidden shared state. Inputs declare units, coordinate frames or topology identifiers, timestamps, validity windows, and provenance. Outputs include a status, result payload, limitation flags, and a stable link to the run configuration.

Adapters are introduced in read-only or simulated form first. A live adapter cannot inherit the trust assigned to the reference model; it requires separate authentication, calibration, timeout, retry, and safe-state evidence.

XIII. EXTERNAL VALIDITY AND RELEASE CRITERIA

The initial conclusions are bounded by the declared simulator, public data, protocol mocks, and sampled conditions. They do not automatically generalize to other equipment, optical assemblies, wafers, stages, power systems, operating ranges, or production environments. External validity is evaluated as a new experiment, not assumed from architectural similarity.

A portfolio release requires the paper source, generated PDF, configuration schema, reproducible entry point, tests, raw or referenced evidence, and a concise limitations record. If one of these artifacts is missing, the project may remain a research design but cannot be presented as an evidence-backed implementation.

XIV. CLAIM BOUNDARIES, SAFETY, AND RISKS

Public benchmarks do not reproduce a production wafer-inspection line. No yield, fab, or deployment claim is made. Synthetic defects and cross-domain pretraining must be labeled and evaluated separately from real defect imagery.

The primary research risk is a category error: treating simulation, public benchmark data, or protocol mocks as proof of physical-system performance. The mitigation is provenance at every boundary, fail-closed behavior for missing state, and an explicit review before any live interface is enabled.

XV. GOALS, MILESTONES, AND INTEGRATION

1 Establish provenance-aware data loaders and split manifests.

2 Reproduce full-label CNN and U-Net baselines.

3 Add self-supervised pretraining.

4 Compare random and k-center annotation budgets.

5 Integrate inspection evidence into the FabSystems digital twin.

XVI. CONCLUSION

Obsidian now has a research plan comparable in rigor and presentation to the established portfolio papers. Its architecture is inspectable, its evaluation is falsifiable, and its limitations are explicit. Credibility will come from completing the staged evidence artifacts, not from promoting unverified numbers.

REFERENCES

[1] SJTU Intelligent Optics Lab, 'Annotation-efficient learning for OCT segmentation,' GitHub.

[2] MVTec Software GmbH, 'MVTec Anomaly Detection Dataset.'

[3] Aaron Singh, 'Obsidian research specification,' local portfolio repository, 2026.

TABLE I
SYSTEM COMPONENTS AND EVIDENCE RESPONSIBILITIES

Component	Responsibility	Required evidence
Provenance registry	Tracks data origin, wafer groups, licenses, and transforms	Immutable split manifest
Representation encoder	Learns features from labeled and unlabeled imagery	Checkpoint and training record
Task heads	Classifies wafer maps and segments SEM defects	Per-class predictions and masks
Active selector	Ranks samples by diversity and uncertainty	Annotation sequence and budget curve

TABLE II
PLANNED EVALUATION MATRIX

Metric family	Acceptance record
Macro F1 and per-class recall	Repeated result with configuration, seed, revision, and limitations
AUPRC	Repeated result with configuration, seed, revision, and limitations
Dice and IoU	Repeated result with configuration, seed, revision, and limitations
Expected calibration error	Repeated result with configuration, seed, revision, and limitations
Labels and annotation hours	Repeated result with configuration, seed, revision, and limitations

TABLE III
RISK REGISTER AND CONTROLS

Risk	Failure consequence	Control
Group leakage	Optimistic test scores	Wafer- or acquisition-isolated splits
Class imbalance	Rare defects are ignored	Macro metrics and per-class recall
Uncalibrated confidence	Novel defects pass unnoticed	Calibration and abstention tests
Domain shift	Public data do not match fab imagery	Separate cross-domain reporting

TABLE IV
PROJECT GOALS AND REQUIRED PROOF

ID	Goal	Completion evidence
G1	Establish provenance-aware data loaders and split manifests.	Reviewable source, test, trace, or measurement artifact
G2	Reproduce full-label CNN and U-Net baselines.	Reviewable source, test, trace, or measurement artifact
G3	Add self-supervised pretraining.	Reviewable source, test, trace, or measurement artifact
G4	Compare random and k-center annotation budgets.	Reviewable source, test, trace, or measurement artifact
G5	Integrate inspection evidence into the FabSystems digital twin.	Reviewable source, test, trace, or measurement artifact