ICSE 2027 Paper Progress

Behavior Vector Coverage

BVS Coverage Project

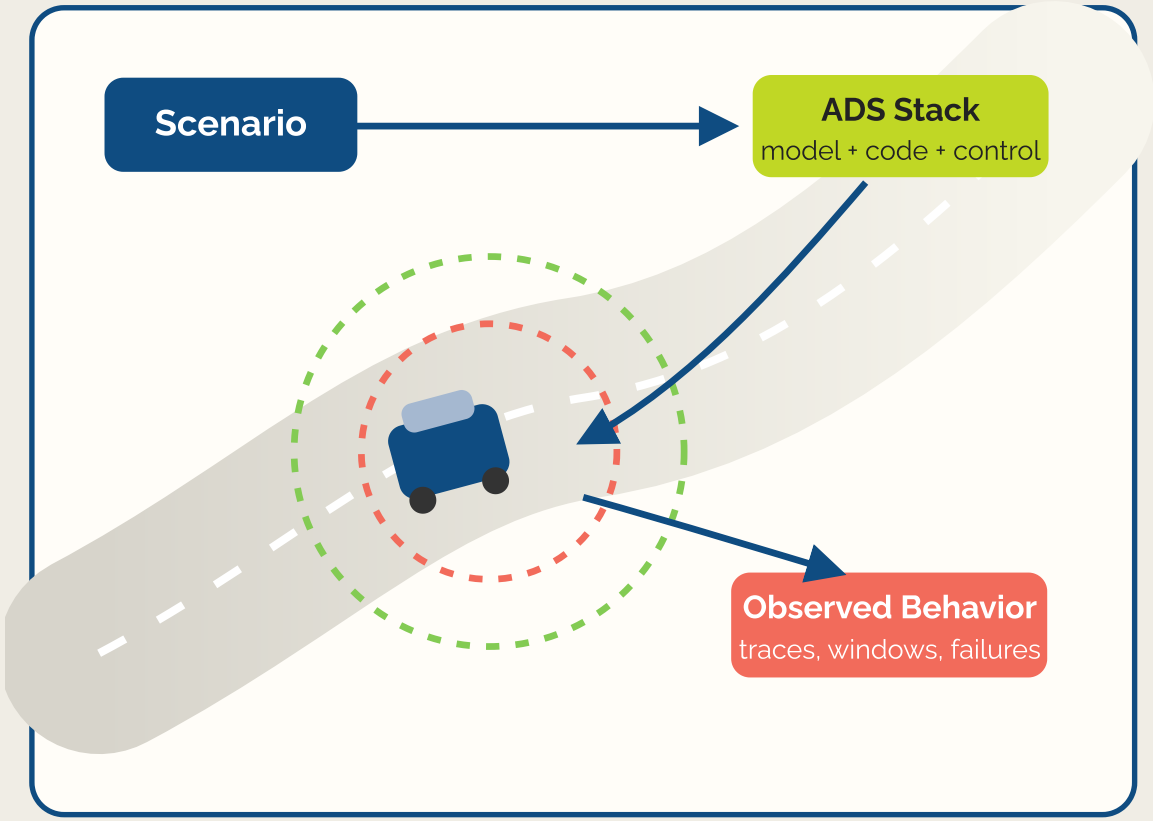
ADS test adequacy progress update from current implementation and InterFuser RQ1 artifacts

June 5, 2026

Motivation

Ml-Enabled Testing Is Hard

- ADS behavior emerges from model + code + traffic
- Failures are rare, contextual, and expensive to reproduce
- DNN coverage alone misses system-level behavior
- A scenario can pass while still exercising little behavior

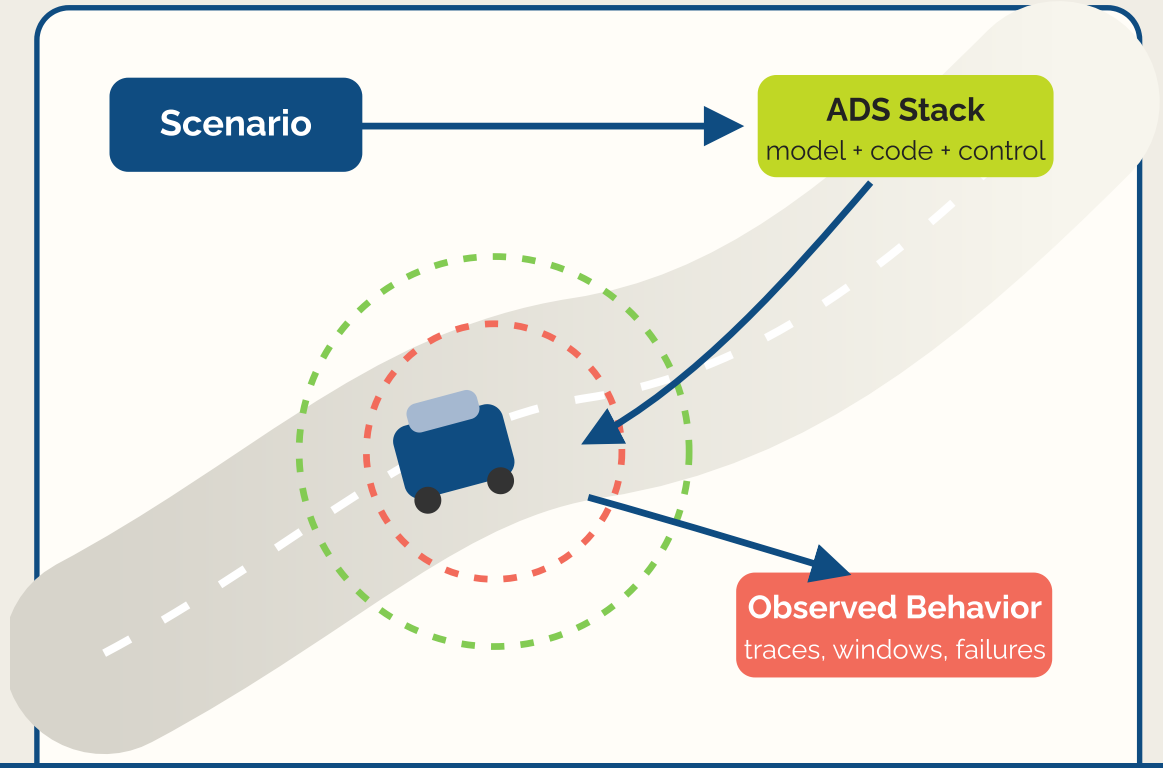


Refs: DeepXplore; DeepGauge; InterFuser

Motivation

Ml-Enabled Testing Is Hard

- **ADS behavior emerges from model + code + traffic**
- **Failures are rare, contextual, and expensive to reproduce**
- **DNN coverage alone misses system-level behavior**
- **A scenario can pass while still exercising little behavior**



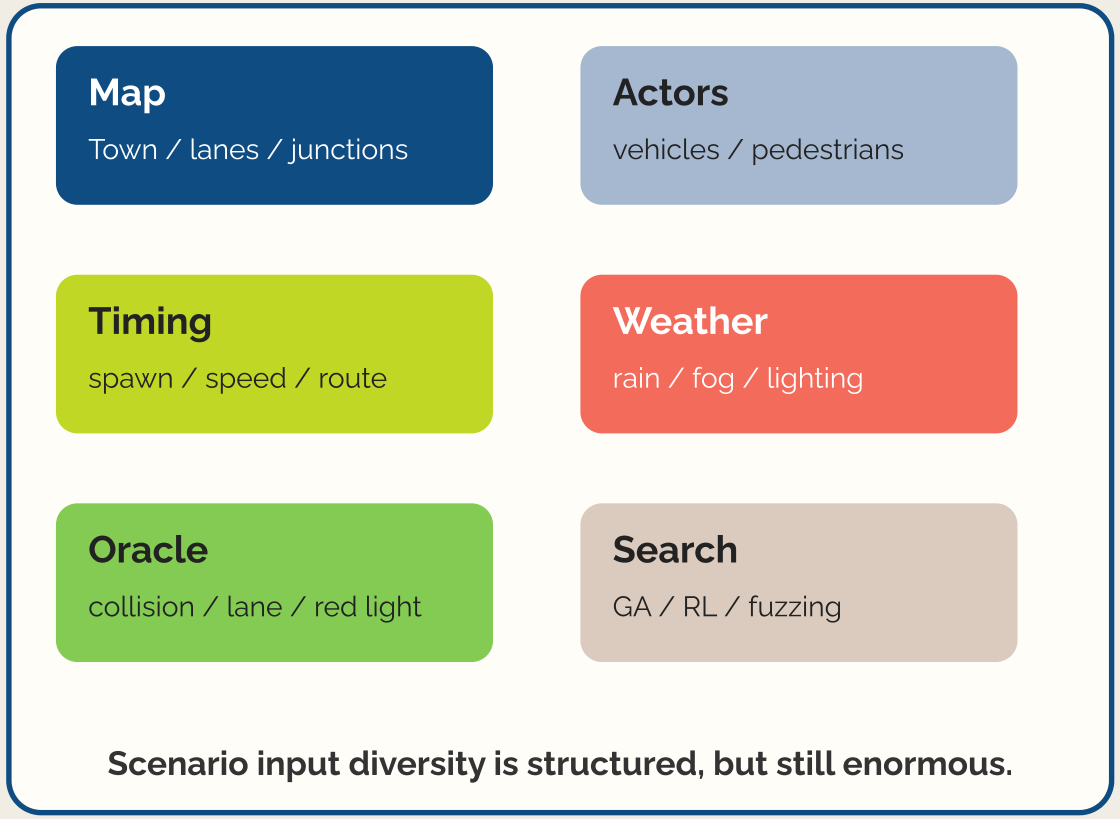
Adequacy should talk about exercised behavior, not only inputs.

Refs: DeepXplore; DeepGauge; InterFuser

Motivation

Scenario Inputs Drive ADS Testing

- **Simulation turns scenarios into executable inputs**
- **Search mutates maps, actors, routes, timing, weather**
- **Diversity is often defined over scenario parameters**
- **The goal is to expose more critical behaviors**

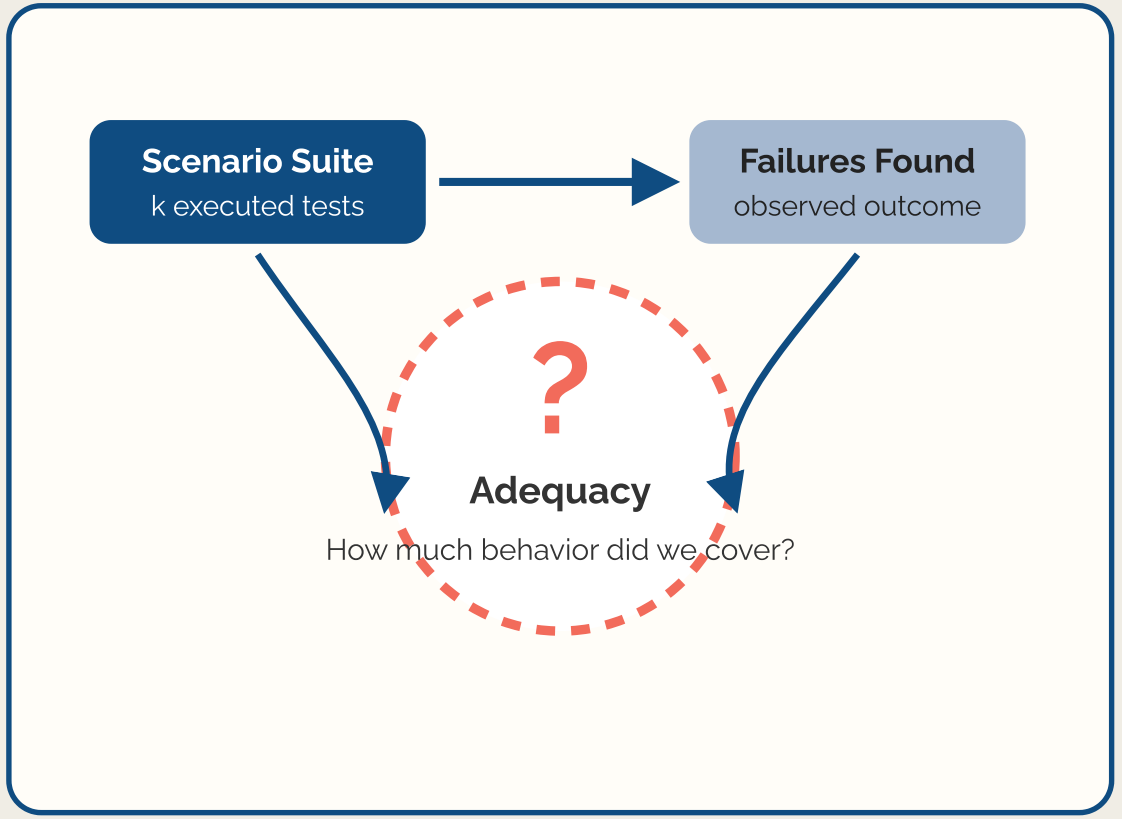


Refs: CARLA; DeepTest; SAMOTA; scenoRITA; DoppelTest; BehAVExplor

Motivation

Input Diversity Still Leaves A Gap

- **Critical-scenario search finds bugs efficiently**
- **But it rarely answers: when is a suite adequate?**
- **Scenario coverage can ignore ADS-internal behavior**
- **More scenarios can still be behaviorally redundant**

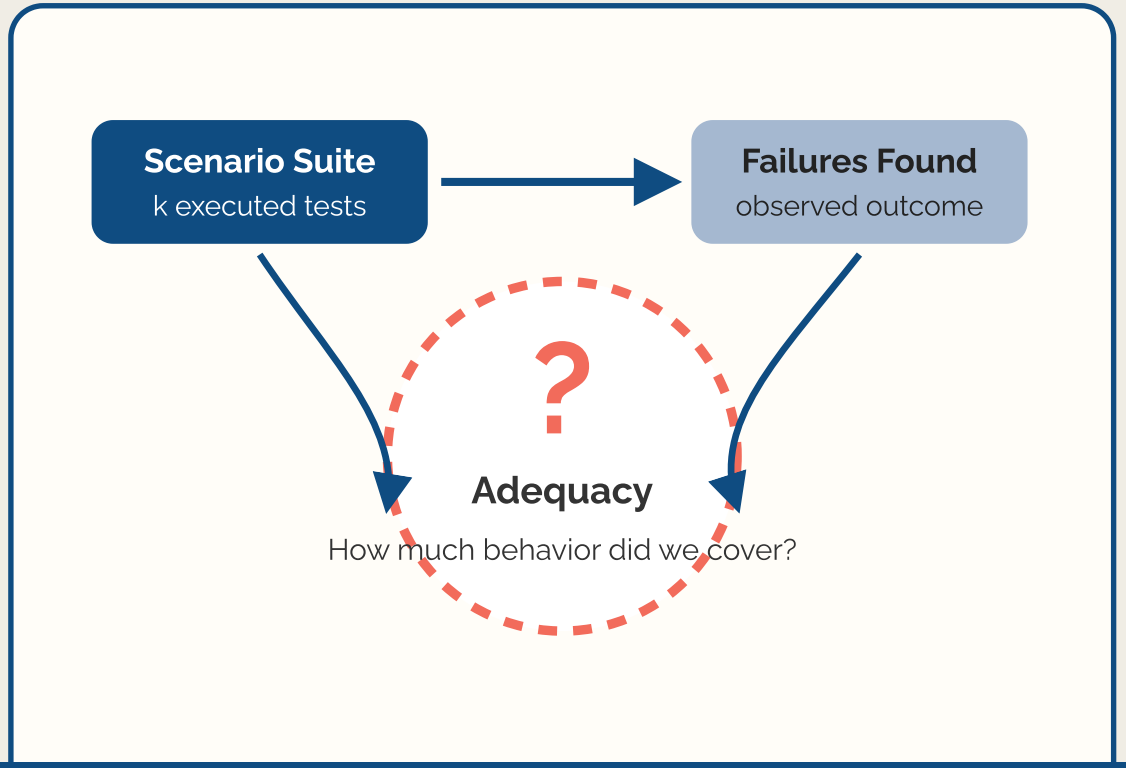


Refs: RAND Driving to Safety; BehAVExplor; scenoRITA

Motivation

Input Diversity Still Leaves A Gap

- **Critical-scenario search finds bugs efficiently**
- **But it rarely answers: when is a suite adequate?**
- **Scenario coverage can ignore ADS-internal behavior**
- **More scenarios can still be behaviorally redundant**



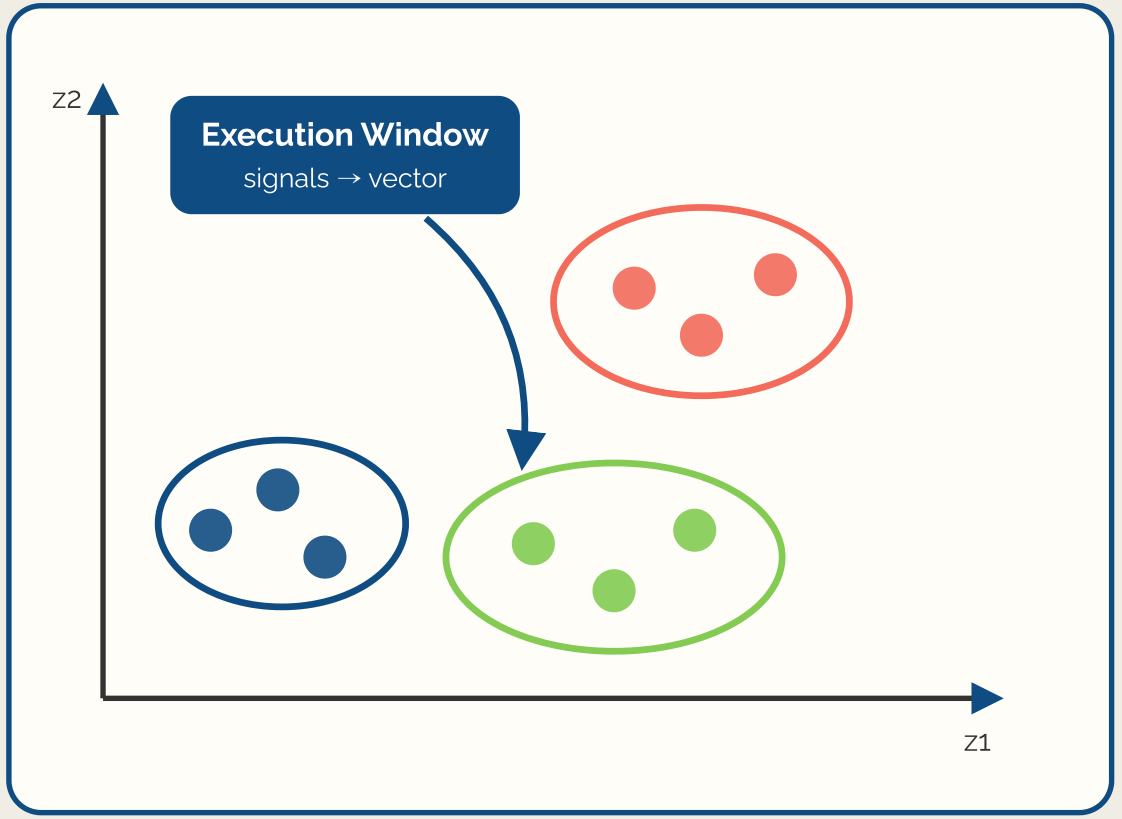
We need a stopping/assessment signal after execution.

Refs: RAND Driving to Safety; BehAVExplor; scenoRITA

Motivation

Behavior Vector Space

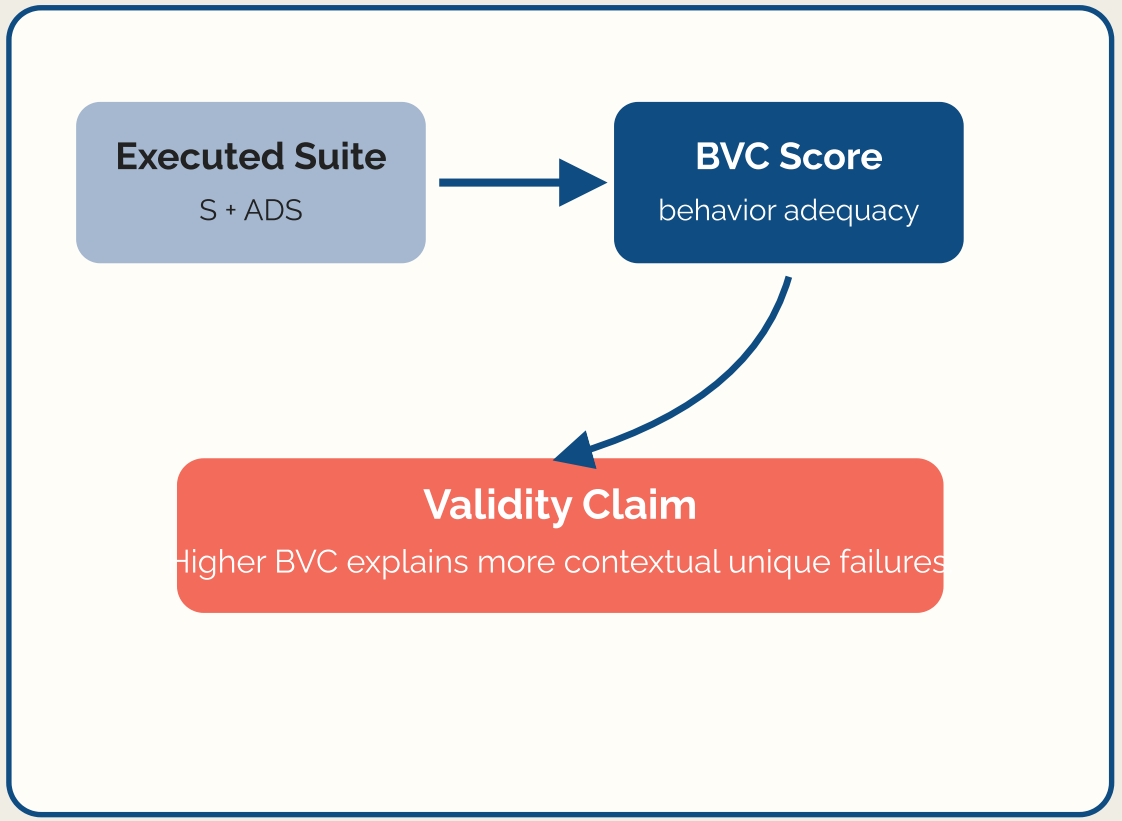
- Collect signals during each execution window
- Map each window to a normalized behavior vector
- Define coverage over the observed vector universe
- Keep the claim empirical, not global reconstruction



Basis: ICSE DS 2026 BVS concept; this deck uses current ICSE 2027 plan.md formalization

Motivation and Goal

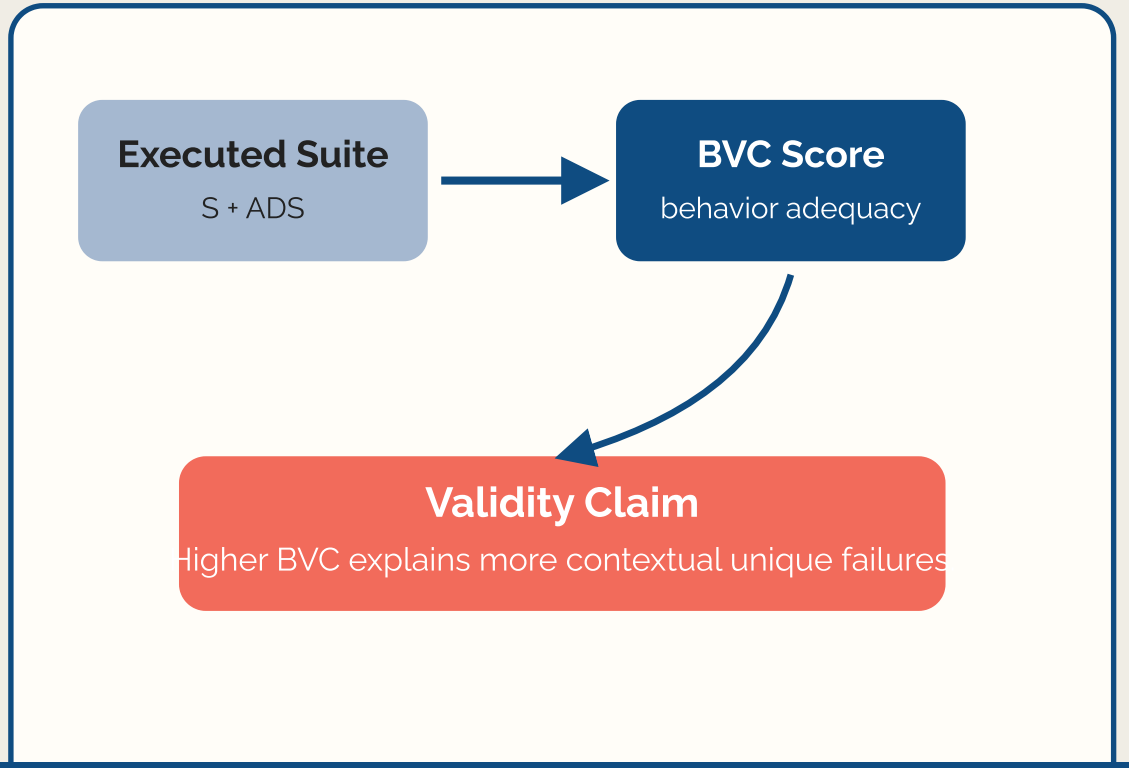
- Input: executed scenario suite and ADS implementation
- Output: behavior-vector coverage and validity evidence
- Question: did the suite exercise diverse ADS behavior?
- Evidence: higher BVC should explain unique failures



Motivation

Motivation and Goal

- **Input: executed scenario suite and ADS implementation**
- **Output: behavior-vector coverage and validity evidence**
- **Question: did the suite exercise diverse ADS behavior?**
- **Evidence: higher BVC should explain unique failures**

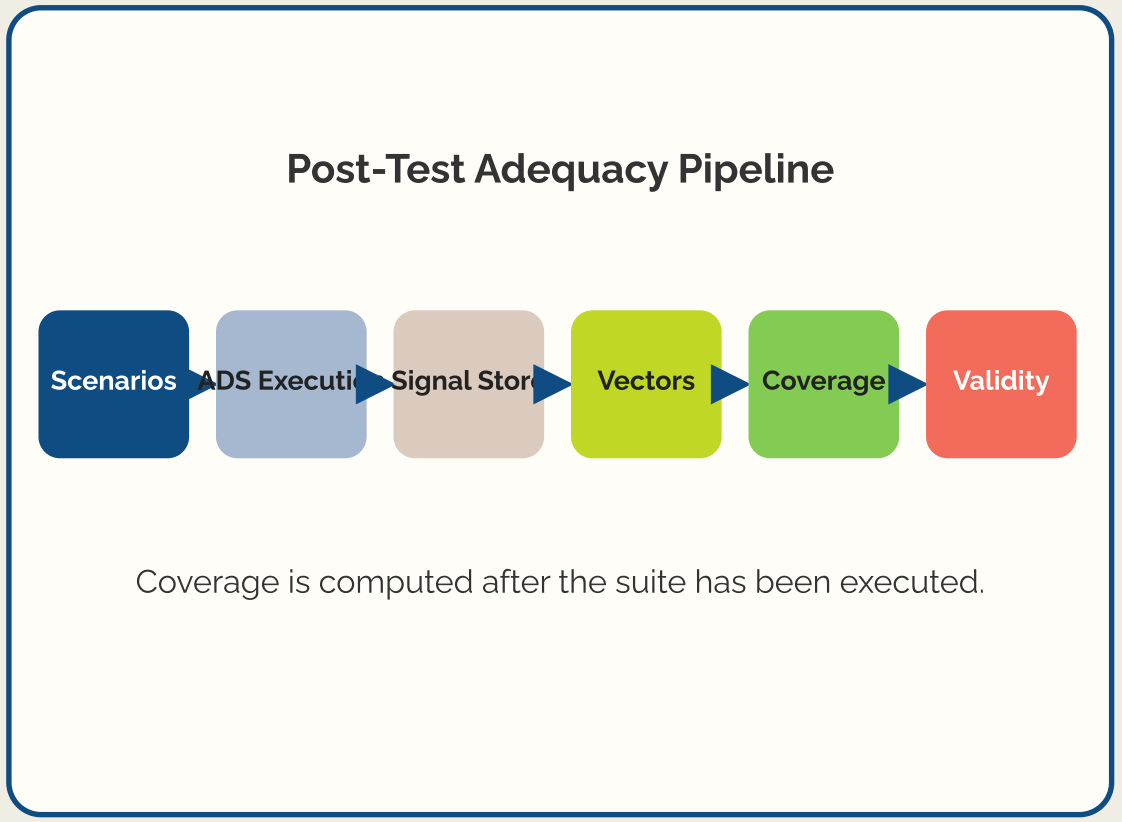


Core thesis: BVC estimates how much observed ADS behavior a suite covers.

Approach

Overall Approach

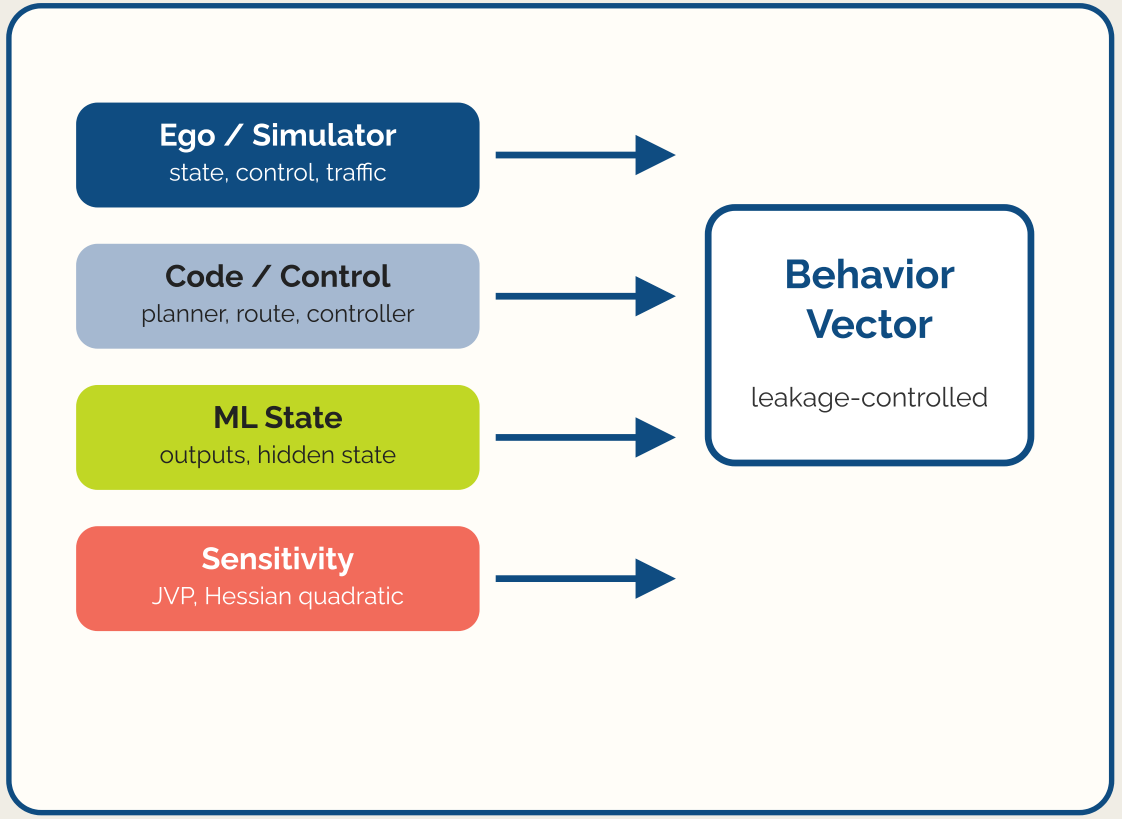
- Execute scenarios on a fixed ADS
- Record simulator, code, ML, sensitivity signals
- Aggregate windows into behavior vectors
- Score suite coverage over observed behavior
- Validate against unique failure discovery



Approach

Signal Fusion

- Ego/simulator: state, control, traffic context
- Code/control: planner and controller decisions
- ML state: outputs, hidden representations
- Sensitivity: JVP and Hessian-quadratic probes
- Leakage control removes shortcut failure proxies

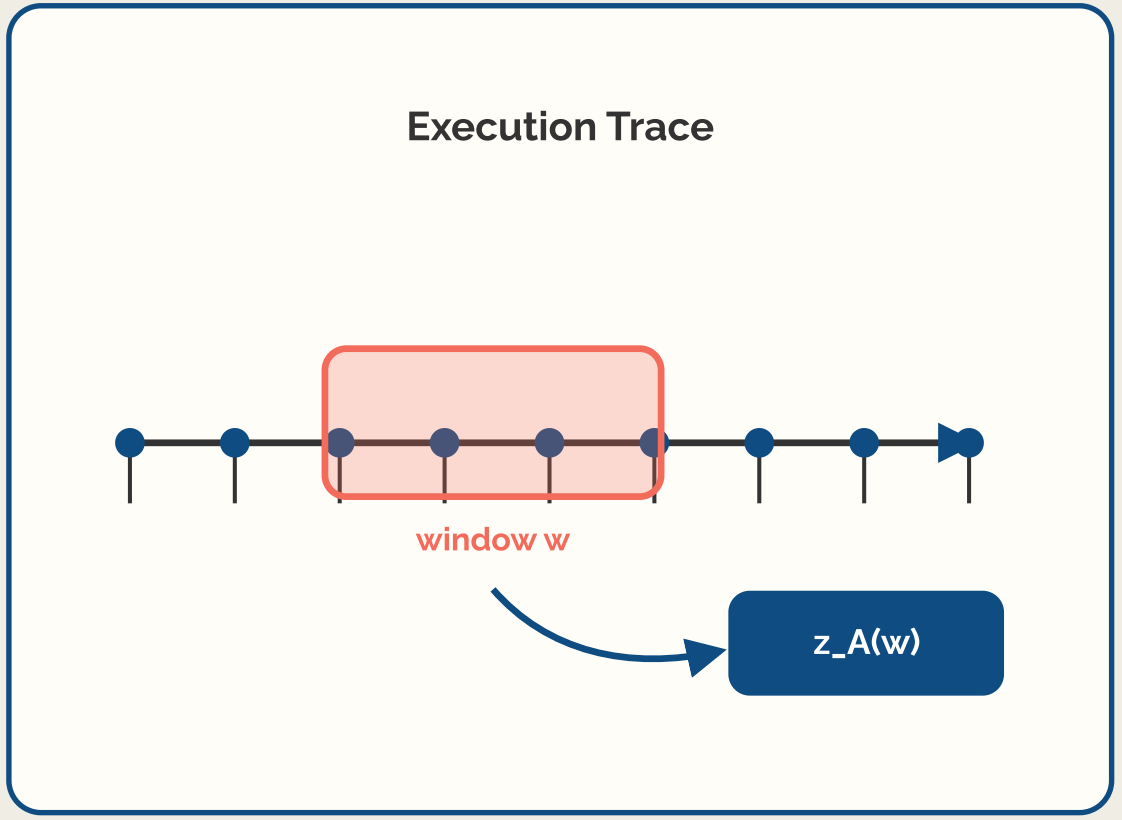


Approach

Scene Behavior Vector

$$z_A(w) = T(\varphi_A(w)) \in R^d$$

- Windowed aggregation over execution traces
- Z-score normalization over the corpus
- Zero-variance removal and clipping
- Optional block-wise PCA for high-dimensional blocks



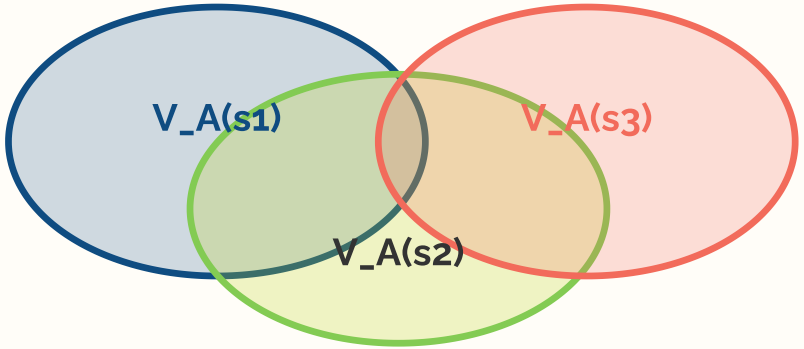
Approach

Scenario-Level BVs

$$V_A(S) = U_{s \in S} V_A(s)$$

$$U_A = U_{s \in D} V_A(s)$$

- Each scenario induces many behavior windows
- A suite is the union of its scenario vectors
- Coverage is measured against the executed pool

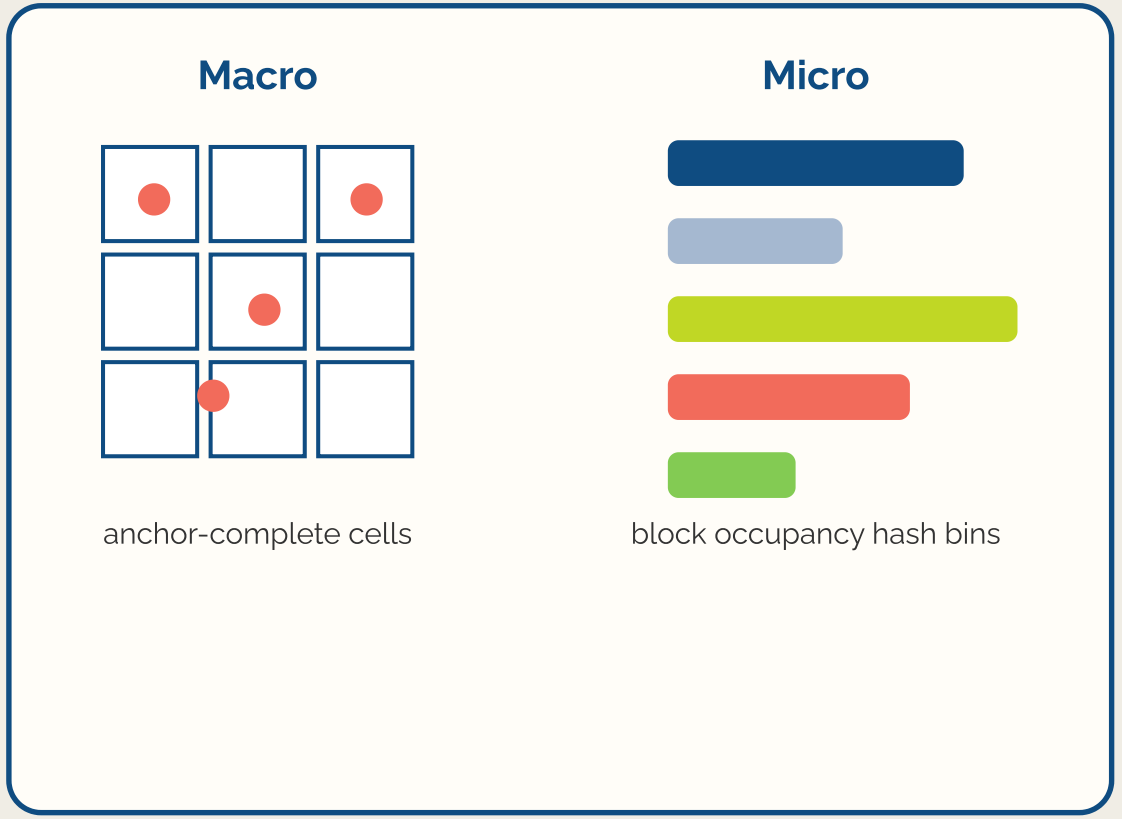


and universe are finite unions of behavior ve

Approach

Macro and Micro Coverage

- Macro: anchor-complete cells in full vector space
- Micro: block-level occupancy proxy
- Radius from nearest-other-scenario quantiles
- Main conservative metric: macro q25
- Robustness reports q05, q10, q25, q50



Evaluation

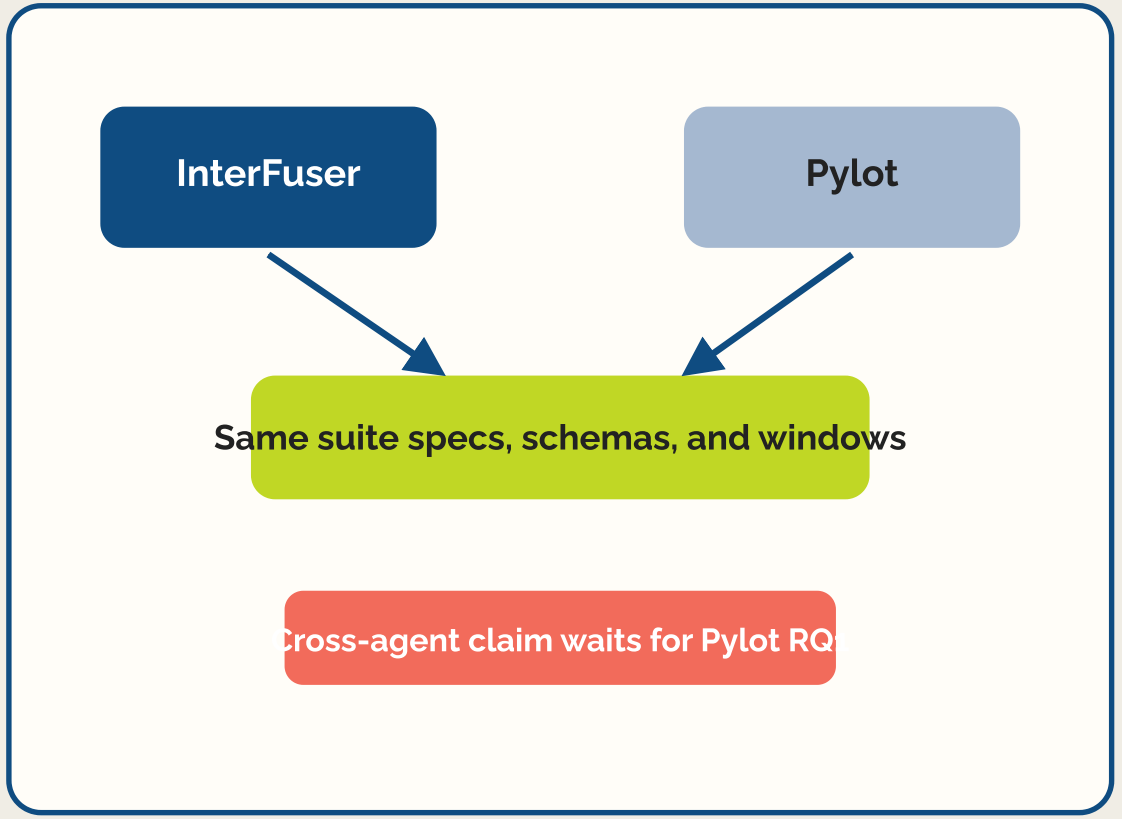
Research Questions

RQ1. [Validity]	**Is BVC positively associated with unique failure discovery?**
RQ2. [Incremental Value]	**Does BVC explain failures beyond scenario/input and simpler execution baselines?**
RQ3. [Construct Validity]	**Which signal families make BVC useful, and do they align with failure semantics?**
RQ4. [Robustness]	**How sensitive and practical is BVC after signal collection?**

Evaluation

Experimental Setup

Component	Current plan
ADS	InterFuser complete; Pylot running
Simulator	CARLA / Drivora scenario runner
Suite sizes	5, 10, 25, 35, 50, 75, 100
Suite specs	100 random suites per size
Windows	0.25s, 0.5s, 1.0s
Schemas	code_only, ml_only, all

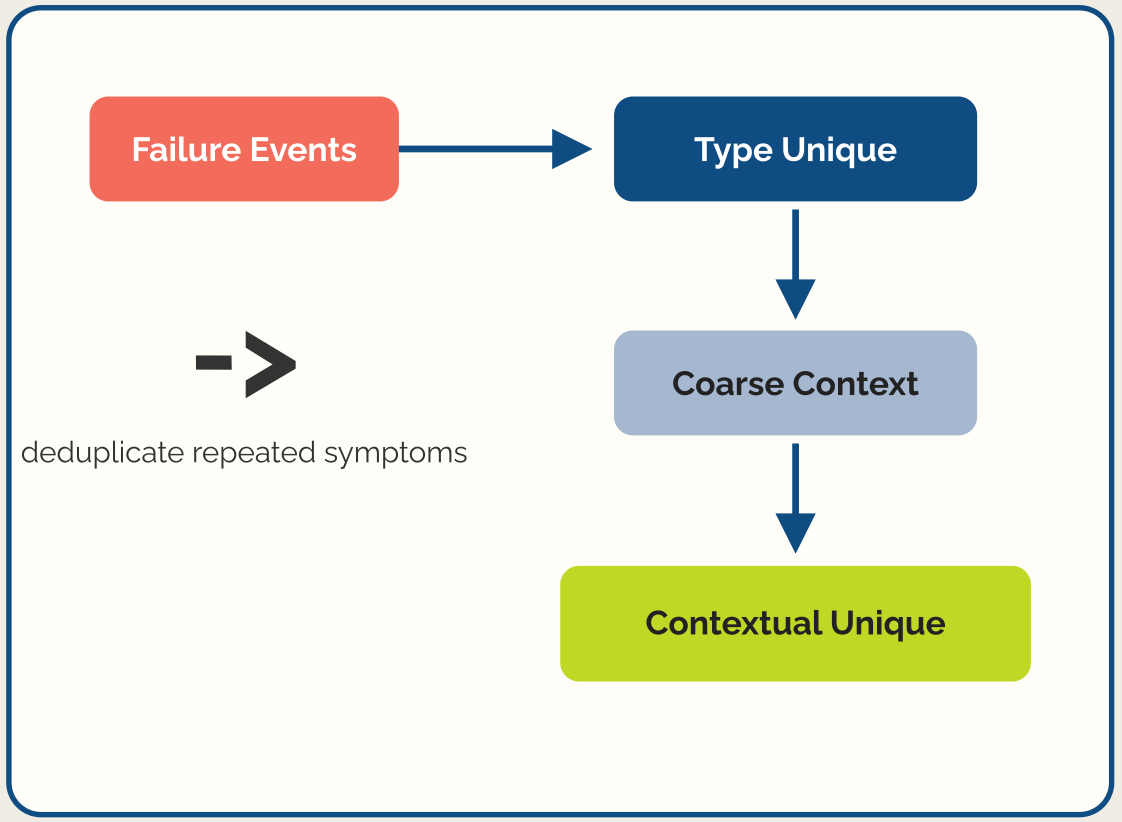


Refs: CARLA; InterFuser; Pylot

Evaluation

Unique Failure Definition

- Failure event: one oracle/infraction occurrence
- Type-unique: failure family only
- Coarse unique: type plus coarse location
- Contextual unique: type plus scenario context
- Primary target: contextual unique failures



Evaluation

RQ1 InterFuser Experiment

- **Mode: adequacy_fullset**
- **700 fixed suite specs reused everywhere**
- **3 schemas × 3 windows × 7 suite sizes**
- **Macro/micro metrics at q05, q10, q25, q50**
- **Correlation and OLS HC3 regression**

Primary choice	Value
Adequacy	macro_cell_coverage_q25
Failure target	contextual_unique_failure_count
Distance	L2 / sqrt(dim)
Controls	suite_vector_count
Rows	63 primary combinations

Evaluation

RQ1 Dataset Snapshot

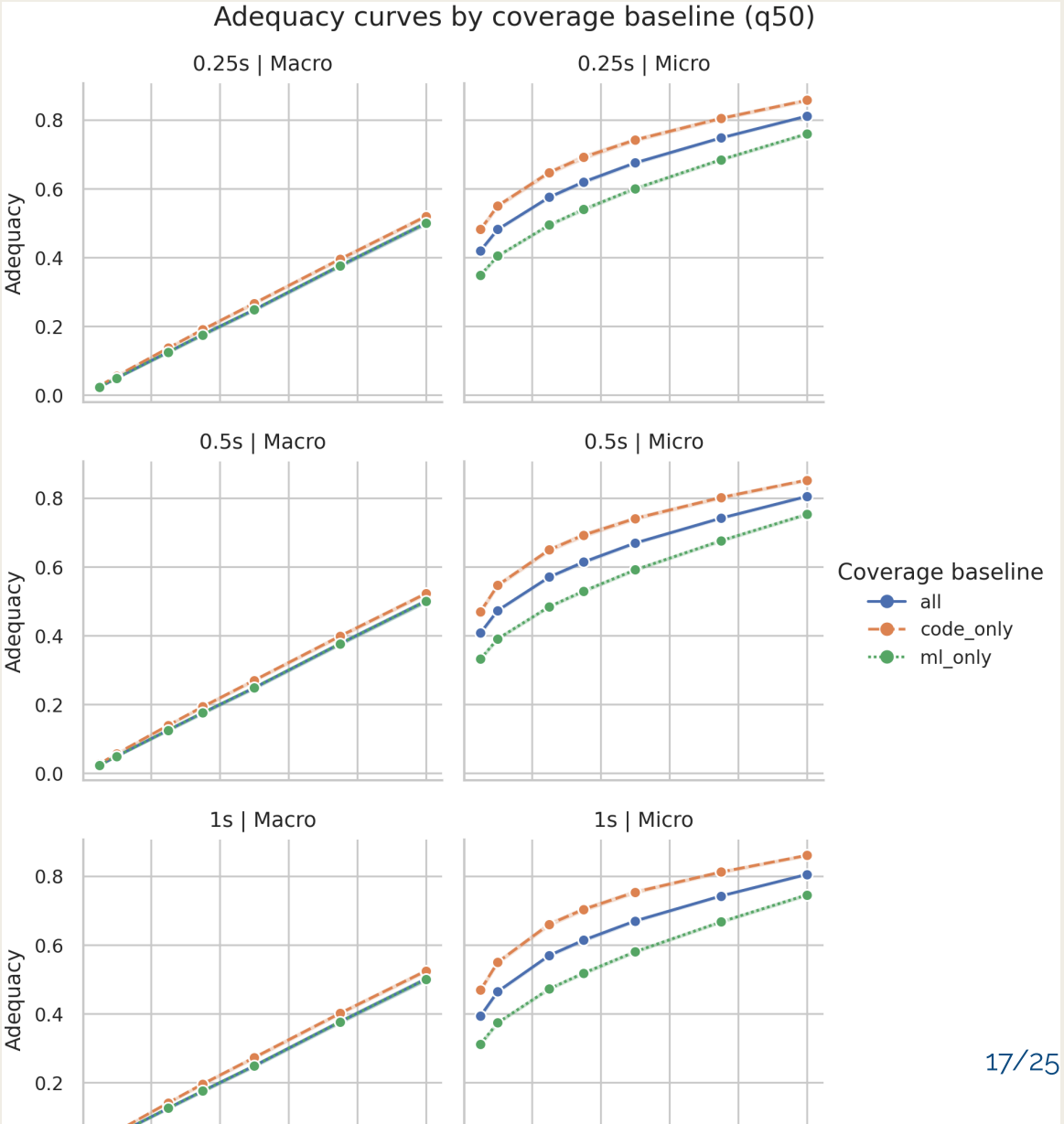
200 Successful scenarios	588 Ego traces	514 Failure events	249 Contextual unique
700 Fixed suites	1.38M Vector rows	3,637 All normalized dim	63 Primary rows

Source: results/paper/tables/interfuser_rq1_dataset_summary.csv; results/interfuser/rq/rq1/multi_window_summary.json

Evaluation

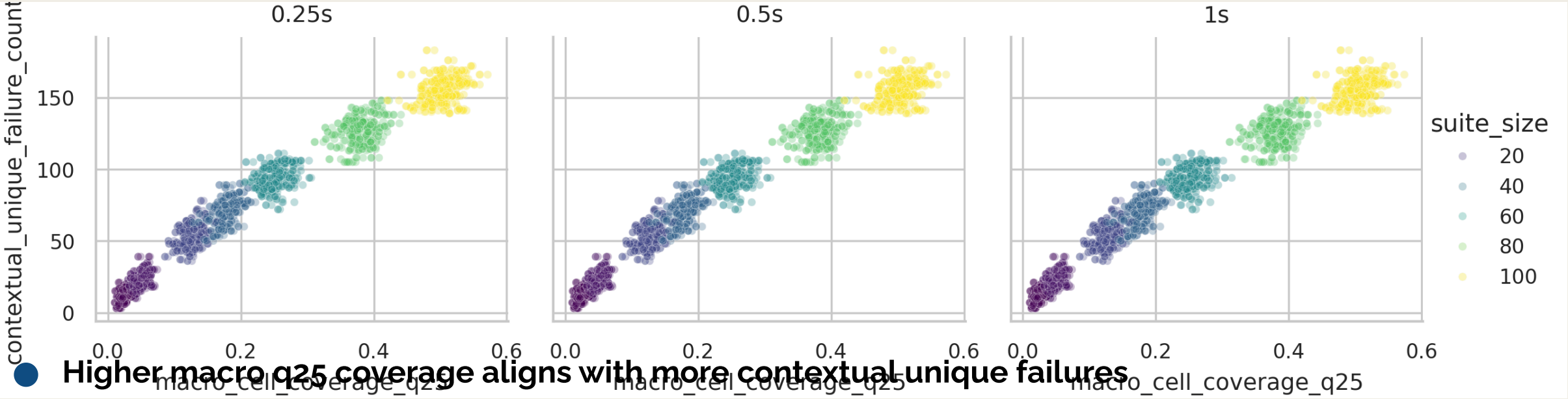
Adequacy Rises with Suite Size

- Coverage increases monotonically with suite size
- Macro grows conservatively from near zero
- Micro starts higher and saturates faster
- code_only is strong in macro q50
- all is strongest in micro q50



Evaluation

Coverage Tracks Unique Failures

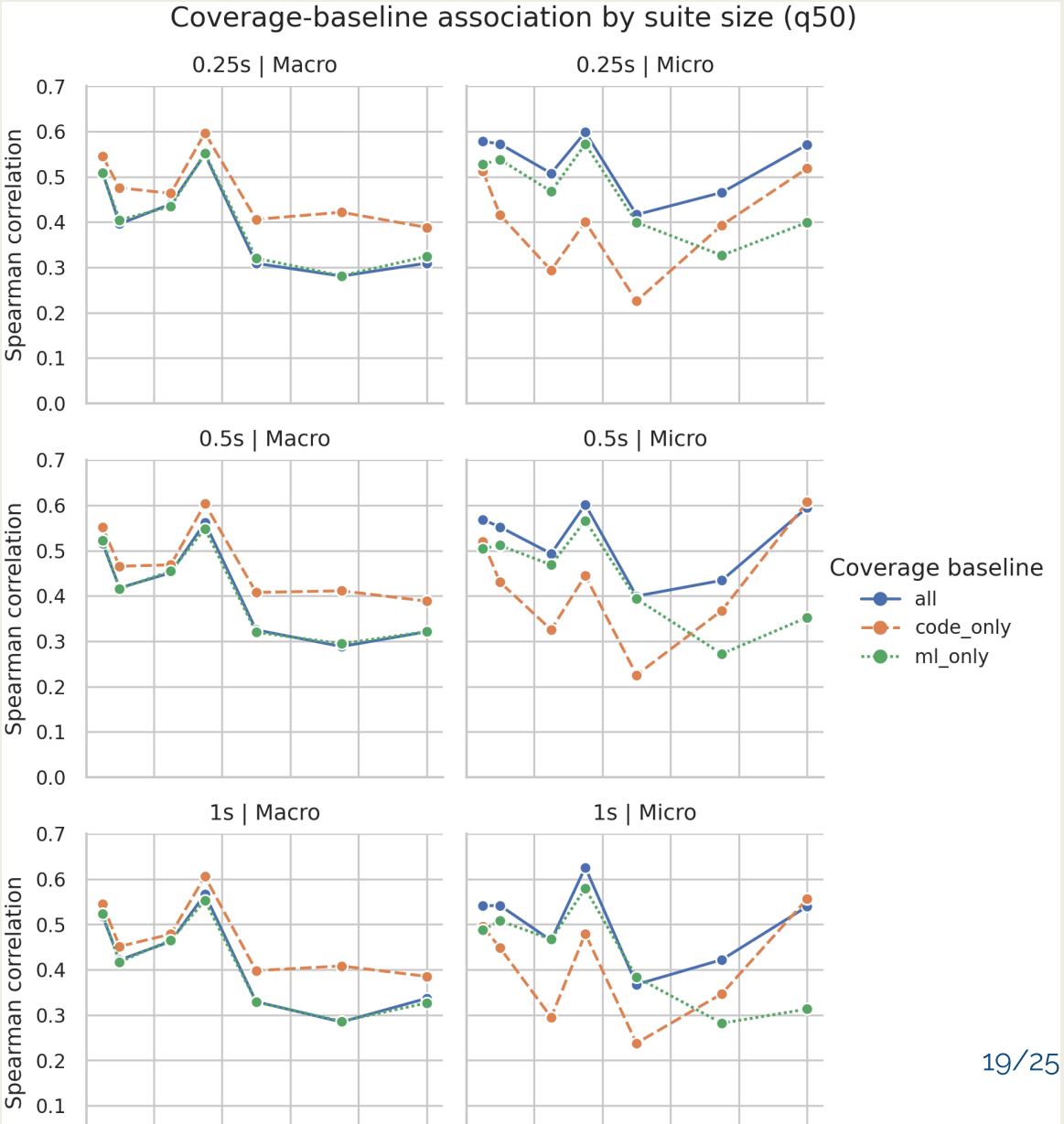


- **Higher macro q25 coverage aligns with more contextual unique failures**
- **The pattern repeats across 0.25s, 0.5s, and 1.0s windows**
- **Suite size explains level, BVC explains ordering within size**
- **This is the main validity picture**

Evaluation

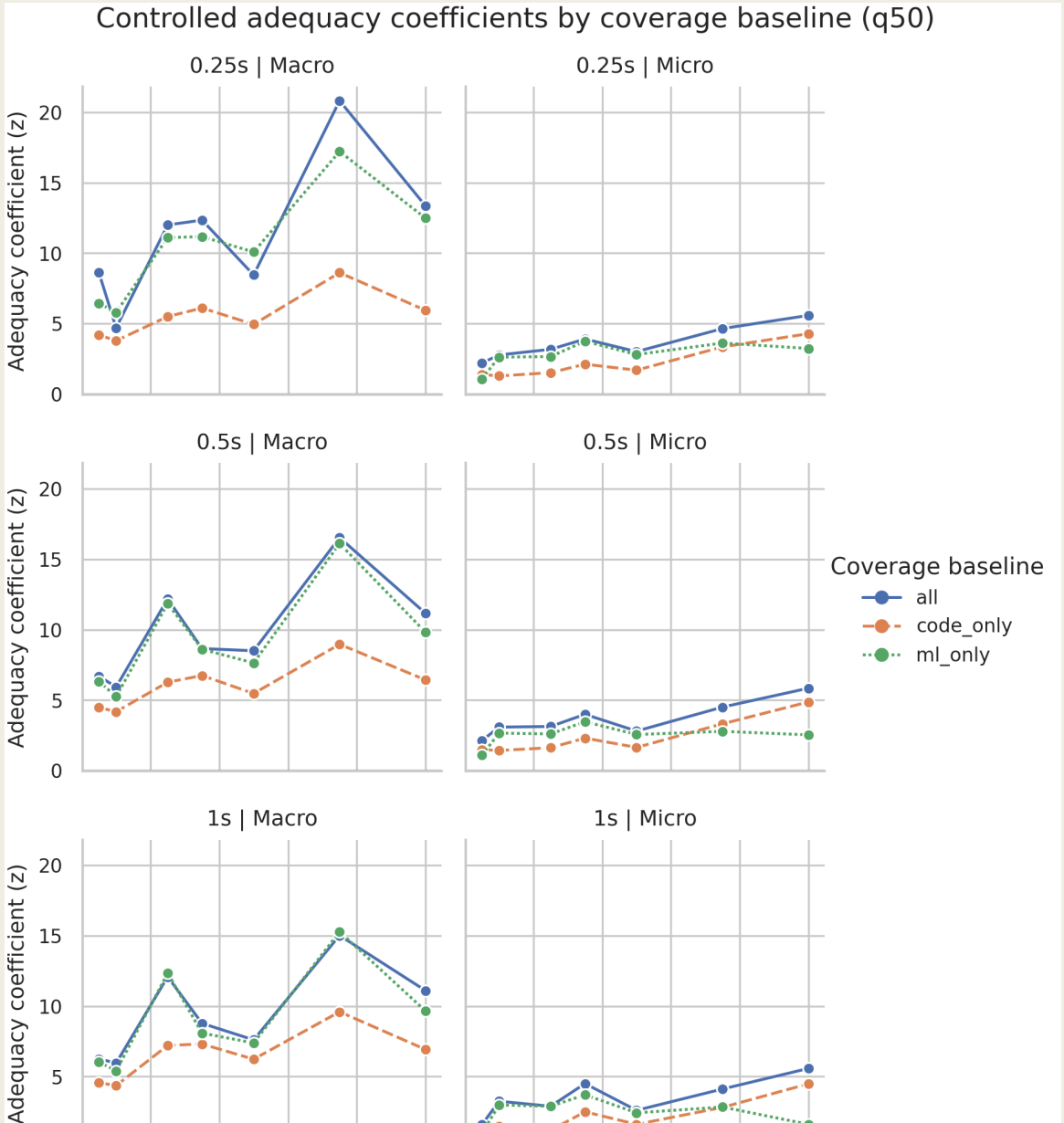
Association Holds Across Sizes

- **Primary Spearman: 63/63 positive**
- **Mean Spearman: 0.389**
- **Range: 0.232 to 0.555**
- **Positive in every schema-window-suite cell**
- **Strong support for InterFuser validity**



Evaluation

Controlled Models Support BVC

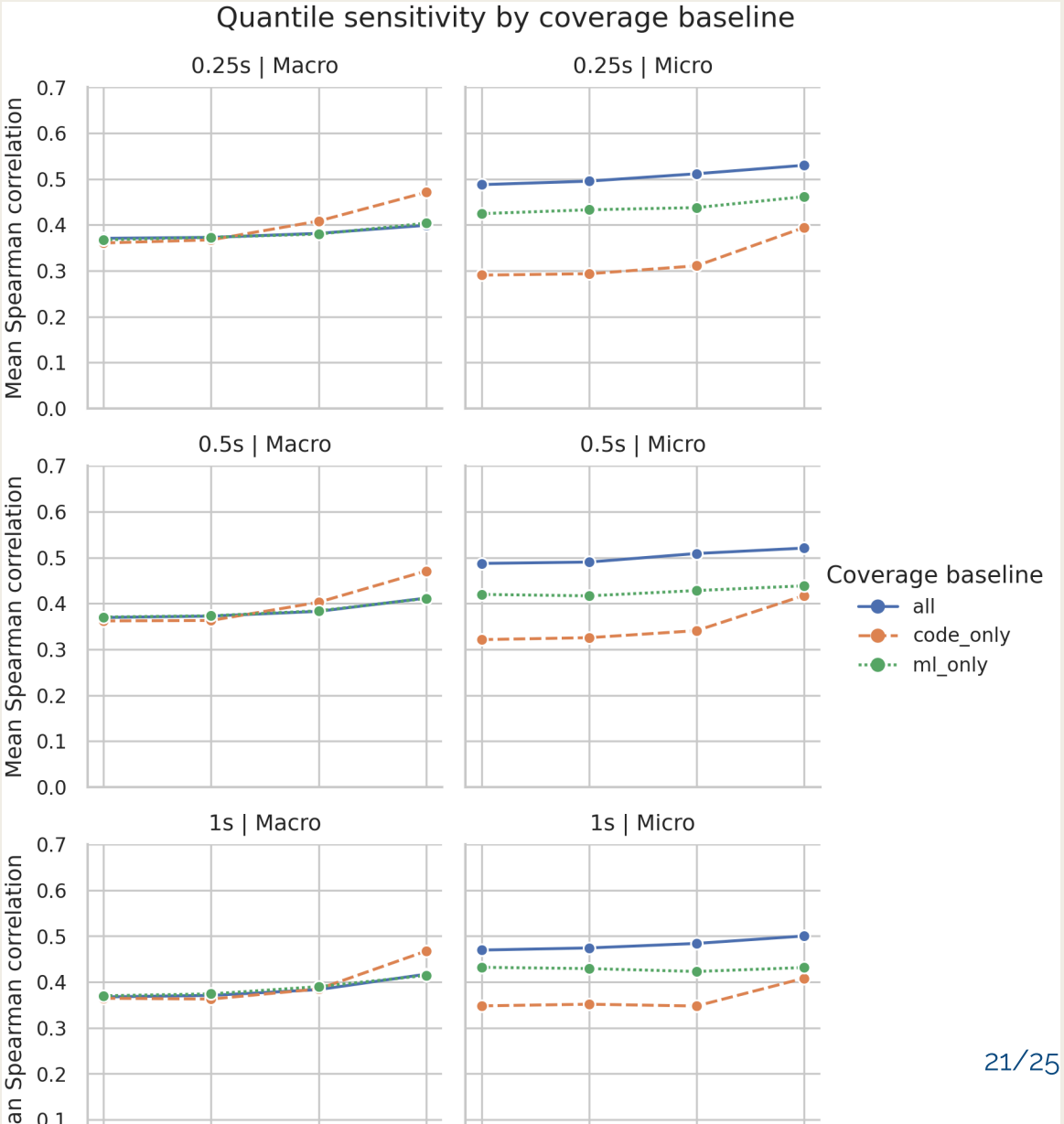


- **Primary BVC coefficients: 63/63 positive**
- **Significant positive: 54/63 at p < 0.05**
- **Significant negative: 0/63**
- **BVC remains explanatory after execution volume control**
- **Interpretation: not just more logged vectors**

Evaluation

Quantiles and Families Matter

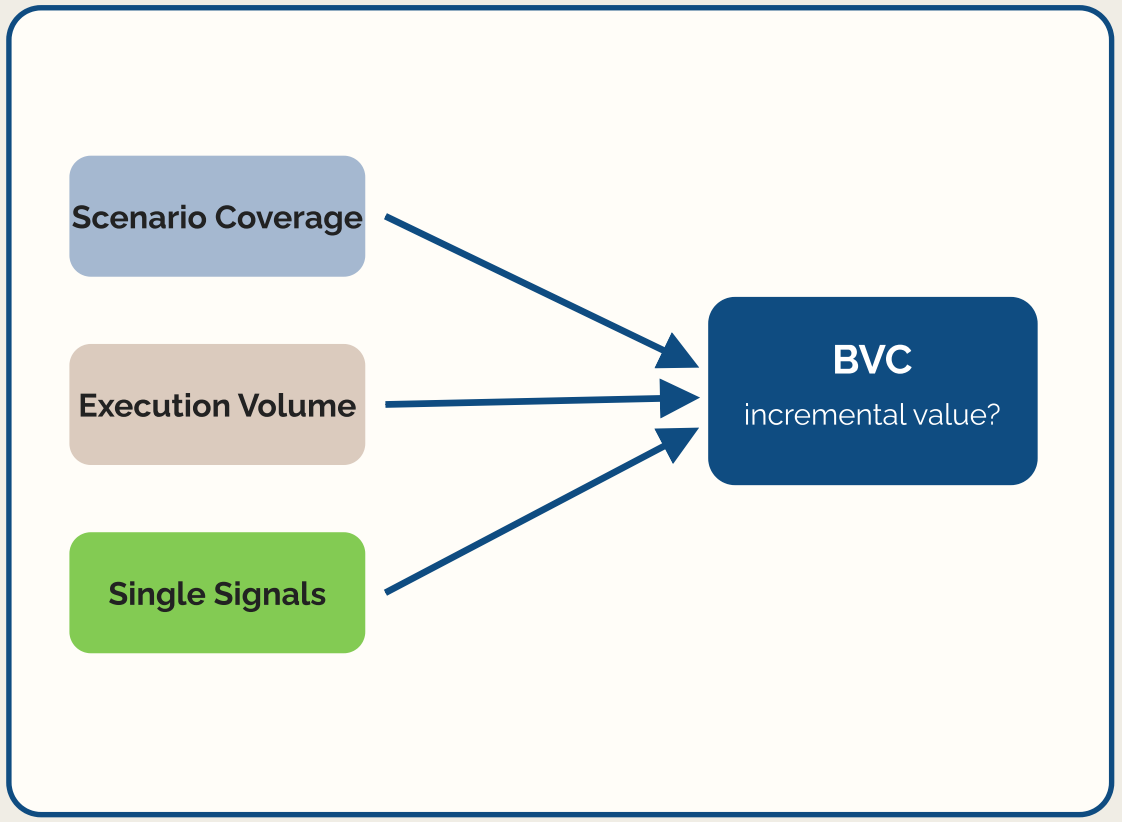
- **Macro q25 is the conservative primary metric**
- **q50 often gives stronger empirical association**
- **Micro/all is strong but less geometrically exact**
- **code_only, ml_only, all are separate baselines**
- **Do not collapse everything into one BVC score**



Evaluation

RQ2 Baseline Comparison Plan

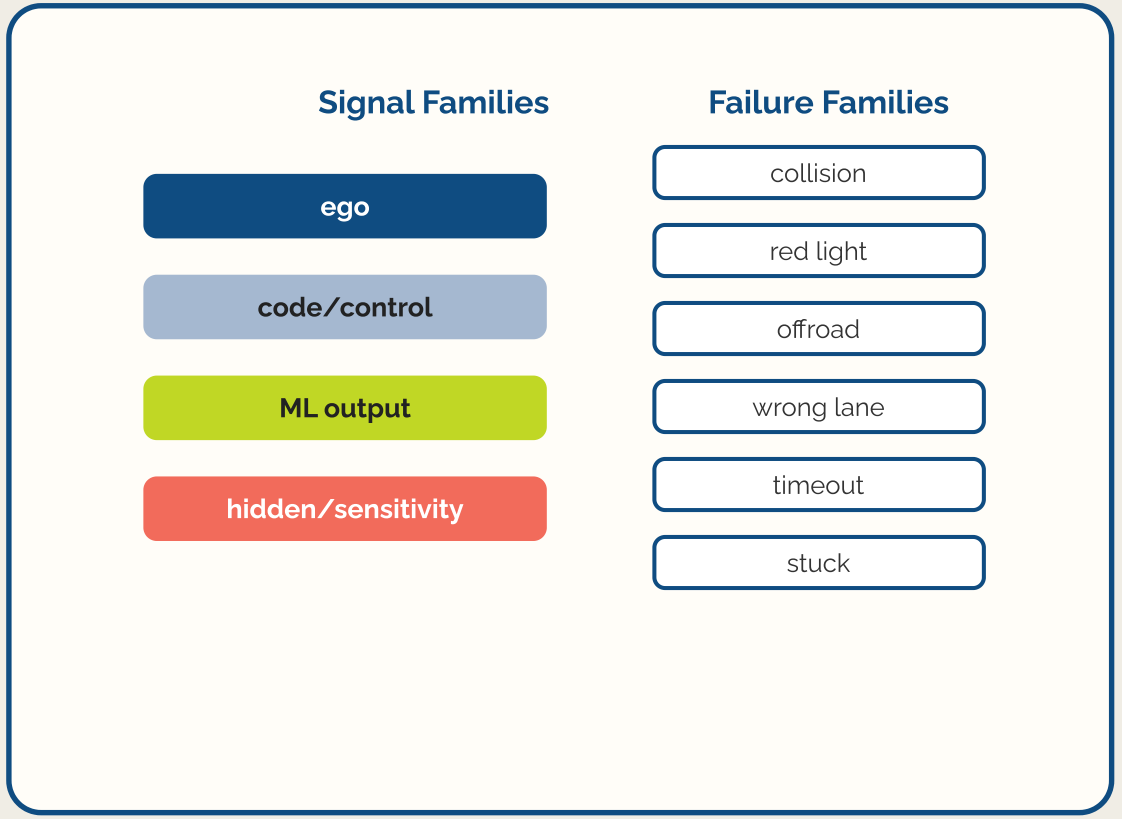
- Reuse RQ1 fixed suite specs
- Compare against random suite sampling
- Scenario/input-space coverage baselines
- Execution-volume and trace-count baselines
- Single-signal or family-only adequacy baselines



Evaluation

RQ3 Signal Semantics Plan

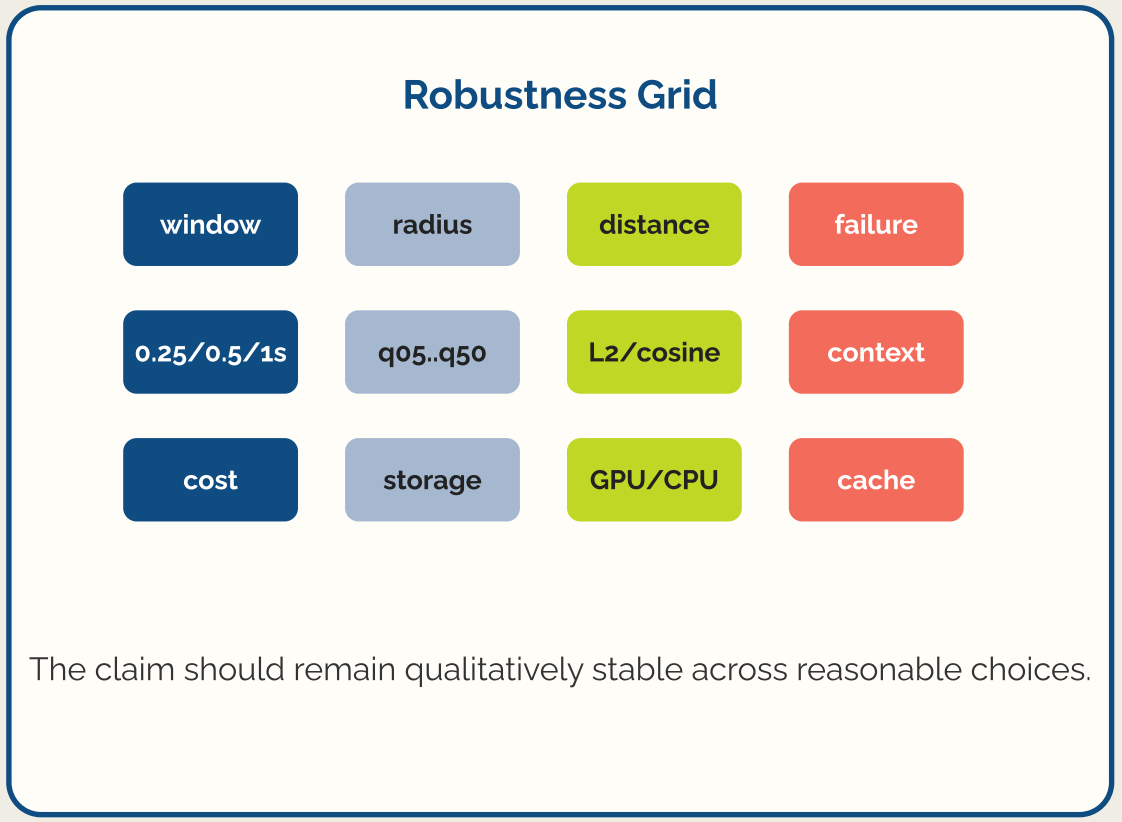
- Ablate signal families and schema variants
- Measure association with contextual failures
- Map signal families to failure families
- Use GA/search diagnostics as interim evidence
- Expected output: family × failure heatmap



Evaluation

RQ4 Robustness Plan

- Window sizes: 0.25s, 0.5s, 1.0s
- Radius quantiles: q05, q10, q25, q50
- Distance metrics: L2 now; cosine/group-weighted planned
- Failure granularities: event, type, coarse, contextual
- Cost: signal collection, vector build, cached scoring



The claim should remain qualitatively stable across reasonable choices.

Next

TODO Through June 30

Jun 5-6	Finish Pylot RQ1 run and sanity-check artifacts
Jun 7-9	Summarize Pylot RQ1 and align cross-agent tables
Jun 10-18	Implement and run RQ2 baselines
Jun 15-23	Run RQ3 signal-family semantics
Jun 20-27	Run RQ4 robustness and cost analysis
Jun 12-30	Write paper sections in parallel

- Pylot RQ1: in progress, target June 6
- RQ2/RQ3/RQ4 share RQ1 suite specs
- Summaries and figures must stay agent-scoped
- By June 30: full experiment narrative draft

Next

TODO Through June 30

Jun 5-6	Finish Pylot RQ1 run and sanity-check artifacts
Jun 7-9	Summarize Pylot RQ1 and align cross-agent tables
Jun 10-18	Implement and run RQ2 baselines
Jun 15-23	Run RQ3 signal-family semantics
Jun 20-27	Run RQ4 robustness and cost analysis

- Pylot RQ1: in progress, target June 6
- RQ2/RQ3/RQ4 share RQ1 suite specs
- Summaries and figures must stay agent-scoped
- By June 30: full experiment narrative draft

Priority: finish cross-agent RQ1 first, then reuse caches aggressively.

Current Progress

Behavior Vector Coverage for ADS Test Adequacy

InterFuser RQ1 supports the core validity claim

{'Next checkpoint': 'Pylot RQ1 and baseline comparison'}

June 2026