FIELDSPACE WHITE PAPER

Deterministic Safety Evidence for Neural ADAS and Autonomy Systems

A public benchmark evidence package for evaluating FieldSpace as a low-compute, auditable safety observer beside existing ADAS and autonomy stacks.

> **FieldSpace turns part of ADAS validation from a fleet-scale neural training problem into a fast replay-and-audit problem.**

60,019
openpilot / comma.ai frames replayed across ten public driving segments

50
Waymo Open Motion Dataset scenarios processed in observer mode

64
official nuPlan mini scenarios completed against classical baselines

17x
lower mean trajectory compute time than PlanCNN on the first shared neural smoke scenario

Executive Summary

Neural driving models have moved ADAS and autonomy forward, but they create a business and safety problem that every supplier eventually has to face: the more complex the model becomes, the harder it is to explain, validate, and reproduce its behavior in edge cases.

FieldSpace is designed for the layer around that problem. It is a deterministic safety observer and validation layer that can run beside an existing ADAS or autonomy stack. It does not require fleet-scale model training before it can produce useful safety evidence. It takes scene state as input, evaluates risk and constraints deterministically, and emits repeatable go / slow / stop judgments with an audit trail.

The practical business idea is simple:

> **FieldSpace turns part of ADAS validation from a fleet-scale neural training problem into a fast replay-and-audit problem.**

Early public benchmark work now supports that direction across four evidence paths:

- openpilot / comma.ai replay: 60,019 real driving frames across ten public route segments, with sub-millisecond CPU latency.
- Waymo Open Motion Dataset observer mode: 50 public scenarios, 4,550 frames, and 15 exported trigger windows.
- nuPlan closed-loop classical comparison: 64 public mini scenarios, 192 total official simulations, and 0 runner failures across SimplePlanner, IDMPlanner, and FieldSpace.
- nuPlan neural smoke comparison: UrbanDriver and PlanCNN public neural checkpoints now run in the same official nuPlan path; on the first shared smoke scenario, FieldSpace matched or exceeded core safety-related scores and ran materially faster on mean trajectory compute time.

The current claim is intentionally focused and commercially useful: FieldSpace can be evaluated on public autonomy data as a deterministic safety layer, can produce inspectable outputs, and can begin direct comparison against accepted classical and neural planner baselines without requiring another training infrastructure program.

The Problem: Neural-Only Validation Is Expensive and Hard to Audit

Modern ADAS stacks increasingly depend on learned components. These models can be powerful, but validation becomes expensive in several ways:

- They require large volumes of labeled or curated driving data.
- They often require GPU-heavy training and retraining loops.
- They can change behavior when the model, weights, training set, or inference environment changes.
- They are difficult to explain to safety engineers, regulators, insurers, and OEM customers after a failure.

- Edge-case coverage is hard to argue because the model's decision boundary is learned, not explicit.

For a supplier, this is not only a technical issue. It affects cost, integration risk, customer trust, liability posture, and time to pilot.

The strongest near-term role for FieldSpace is not to replace every neural system in the vehicle. The stronger role is to sit beside those systems and answer a different question:

> **Given the current scene, does an independent deterministic layer agree that the planned behavior is safe, explainable, and within constraints?**

That question is valuable because it does not require Magna or an OEM to rip out an existing stack. It creates an evaluation path that is smaller, faster, and easier to review.

What FieldSpace Is

FieldSpace is a deterministic safety and validation layer for autonomy. It can be used in shadow mode, where it observes the same scene state available to an ADAS or autonomy stack and produces an independent safety judgment.

At a high level:

1. Perception, map, or log data provides scene state.
2. FieldSpace converts objects, lanes, goals, speed, and constraints into a structured risk field.
3. Obstacles, boundaries, and rules repel unsafe motion.
4. Goals and route intent attract allowable motion.
5. The system emits a deterministic action judgment or trajectory candidate.
6. The output can be audited because the inputs, constraints, derived risk quantities, and decision path are inspectable.

The important distinction is that FieldSpace does not need to learn a driving policy from fleet-scale examples before it can start producing safety evidence. It can run as a deterministic observer on logs, public benchmarks, or Magna-selected edge cases.

Why This Matters Commercially

For a business audience, the value is not only "better autonomy." The immediate value is lower-friction validation.

Lower Evaluation Cost

FieldSpace can be tested on existing logs or public benchmark data. That lowers the initial burden because a partner does not need to stand up a full training pipeline, labeling workflow, retraining loop, or GPU cluster just to evaluate the core safety-observer value.

Faster Pilot Path

The first useful pilot does not need to be a full autonomy replacement. A focused pilot can be:

- select public or internal edge cases,
- replay the scenes through FieldSpace in shadow mode,
- compare FieldSpace outputs against baseline ADAS behavior or agreed benchmark metrics,
- review the audit traces and failure cases.

That is a smaller ask than “give us your full autonomy stack and months of engineering support.”

More Explainable Safety Evidence

A neural model can produce a good action without making the reasoning easy to inspect. FieldSpace is designed to produce the safety evidence directly: scene state, active constraints, risk level, warning timing, and recommended action.

That matters for:

- safety review,
- internal engineering debates,
- insurance discussions,
- regulator-facing evidence,
- OEM customer confidence.

Less Platform Disruption

FieldSpace can be positioned as an independent observer first. That is commercially important because large automotive organizations rarely want to replace an existing ADAS stack just to test a new idea. A shadow-mode observer can be evaluated with lower organizational resistance.

Public Evidence Path

The current evidence is intentionally staged. It starts with replay and observer mode, then moves into official closed-loop planning.

1. openpilot / comma.ai Replay

FieldSpace was run on public openpilot-style driving logs from comma.ai’s `openpilotci` bucket.

Run size:

- 10 one-minute engaged public driving segments
- 10 vehicle platforms
- 60,019 frames
- FieldSpace observer running at 100 Hz

Headline result:

- Average processing time: 0.30-0.39 ms per frame
- p95 processing time: 0.41-0.56 ms per frame
- Aggregate clear rate: 98.9%
- Caution-or-higher rate: 1.1%
- On routes containing closing-traffic events, FieldSpace warned 13-15 seconds before openpilot committed a sustained hard brake

Interpretation:

This is supporting evidence that FieldSpace can process real driving logs quickly on CPU and remain quiet on clean highway driving while still producing early warning signals on closing-traffic routes.

2. Waymo Observer Mode

FieldSpace was then exercised on the Waymo Open Motion Dataset in observer mode.

Run size:

- 50 public Waymo scenarios
- 4,550 frames
- 2 validation shards
- 15 trigger windows exported

Alert distribution:

- Clear: 4,458 frames, 97.98%
- Caution: 67 frames, 1.47%
- Warning: 25 frames, 0.55%
- Critical: 0 frames, 0.00%

Curated close-interaction result:

- True positives: 8
- False positives: 0
- True negatives: 29
- False negatives: 13
- Precision: 100.0%
- Recall: 38.1%

Interpretation:

This shows a conservative observer profile. When FieldSpace raised caution-or-higher, the input-derived close-interaction heuristic agreed that the scenario was relevant. That supports false-positive control and auditability. Recall is the next engineering target.

3. nuPlan Closed-Loop Classical Baselines

FieldSpace was then run inside the official nuPlan closed-loop reactive-agents simulation path.

Run size:

- 64 public mini scenarios
- Same scenario tokens across all planners
- 3 planners compared: SimplePlanner, IDMPlanner, FieldSpace
- 192 total official simulations
- 0 failed simulations

Scenario mix:

- 21 high-magnitude-speed scenarios
- 18 medium-magnitude-speed scenarios
- 13 intersection scenarios
- 4 near-multiple-vehicles scenarios
- 3 stationary scenarios
- 2 near-construction-zone-sign scenarios
- 1 following-lane-with-slow-lead scenario
- 1 near-pedestrian-on-crosswalk scenario
- 1 stopline-traffic-light scenario

Key results:

Metric	FieldSpace	IDMPlanner	SimplePlanner
Route progress ratio	0.9660	0.9471	0.6210
Comfort score	0.9375	0.9844	0.5156
No-at-fault collision score	0.9766	0.9688	0.9531
TTC within bound	0.9219	0.9219	0.9375
Mean trajectory compute time	0.0084 s	0.0231 s	0.0034 s

Interpretation:

On this bounded public mini slice, FieldSpace completed every scenario, led on route progress, was competitive on safety-related metrics, and ran faster than IDMPlanner on mean trajectory compute time. SimplePlanner remained faster, as expected, because it is intentionally minimal.

4. nuPlan Neural Smoke Comparison

The next evidence gate was to compare against public neural planner baselines. FieldSpace now runs in the same official nuPlan path as public tuPlan Garage neural checkpoints.

Neural baselines executed:

- UrbanDriver

- PlanCNN

Shared smoke scenario:

- Scenario token: `000d90717e5e569d`
- Scenario type: intersection
- Runner failures: 0 across SimplePlanner, IDMPlanner, FieldSpace, UrbanDriver, and PlanCNN

Key smoke results:

Metric	FieldSpace	UrbanDriver	PlanCNN
Route progress ratio	0.9576	1.0000	0.9813
Comfort score	1.0000	1.0000	1.0000
No-at-fault collision score	1.0000	0.0000	1.0000
TTC within bound	1.0000	0.0000	1.0000
Speed-limit compliance	1.0000	0.7546	1.0000
Mean trajectory compute time	0.0303 s	0.0927 s	0.5250 s

Interpretation:

This is a smoke gate, not a broad neural benchmark. It is still important because the neural comparison path now exists. On the first shared scenario, FieldSpace matched PlanCNN on core safety-related scores, exceeded UrbanDriver on collision, TTC, and speed-limit compliance, and ran about 3x faster than UrbanDriver and about 17x faster than PlanCNN on mean trajectory compute time.

What The Evidence Supports Today

The current evidence supports a bounded but commercially meaningful claim:

> **FieldSpace can run on real public autonomy data, produce deterministic safety outputs, complete official closed-loop nuPlan simulations, and begin direct comparison against classical and neural planner baselines with lower compute and inspectable behavior.**

This is enough to justify technical review and a focused validation pilot.

It is not yet enough to claim general autonomy superiority across all neural models, all environments, or all nuPlan scenarios. The stronger claim should be earned by scaling the neural comparison from one shared smoke scenario to the same 64-scenario public mini slice, then beyond that to broader public splits or customer-selected scenarios.

Why This Is Different From “Another Planner”

FieldSpace should not be framed as just another planner competing for the same slot as every neural model. The more useful framing is:

> **FieldSpace is an independent deterministic safety-evidence layer.**

That means its output is valuable even before it controls the vehicle:

- It can challenge a neural stack’s planned action.
- It can flag scenes where the learned policy appears unsafe or uncertain.
- It can provide a replayable explanation after a warning or event.
- It can quantify how often it agrees or disagrees with the baseline.
- It can help decide which edge cases deserve deeper engineering review.

In short, FieldSpace can be useful as a validation layer before it is ever asked to be the primary driving policy.

Integration Model for an Automotive Partner

A practical partner evaluation can be staged in four steps:

1. **Public benchmark review.** Review the existing openpilot, Waymo, and nuPlan artifacts.
2. **Partner-selected edge cases.** Select a narrow set of scenarios: cut-in, hard braking, pedestrian crossing, intersection conflict, traffic-light behavior, or false-positive-sensitive highway scenes.
3. **Shadow-mode replay.** Run FieldSpace beside the existing stack on logs or benchmark scenarios.
4. **Pass/fail evidence package.** Compare warning timing, false positives, missed events, route progress, collision metrics, TTC, comfort, and compute cost.

This avoids a broad POC. It creates a focused technical review with measurable acceptance criteria.

Current Claim Boundary

The current work shows:

- FieldSpace can run on real public driving logs.
- FieldSpace can ingest public Waymo scenarios in observer mode.
- FieldSpace can execute inside official nuPlan closed-loop simulation.
- FieldSpace has completed a 64-scenario public mini-slice comparison against classical baselines.
- FieldSpace has started direct comparison against public neural planner baselines on a shared official nuPlan smoke scenario.
- FieldSpace can produce low-latency deterministic outputs without training a neural driving policy.

The next claim requires more evidence:

- Run UrbanDriver and PlanCNN across the same 64-scenario public mini slice.
- Preserve identical scenario tokens across all planners.
- Track checkpoint hashes and environment provenance.
- Break down results by scenario type and failure mode.
- Add richer traffic-light, route-goal, lane-graph, and map semantics in the FieldSpace wrapper.

Conclusion

The reason FieldSpace is interesting is not that it asks the industry to throw away neural models. The reason it is interesting is that neural models need an auditable safety-evidence layer around them.

FieldSpace gives automotive teams a way to evaluate safety behavior without starting with a fleet-scale training program. It can run on public logs, replay selected edge cases, execute in closed-loop simulation, and produce deterministic outputs that engineers can inspect.

The first public evidence is now in place. The next milestone is clear: scale the neural comparison from smoke to the 64-scenario public mini slice. If FieldSpace continues to match or exceed neural baselines on safety-related metrics while using materially less compute and producing auditable traces, the commercial case becomes much stronger.