

A Benchmark Protocol for Measuring Error in Video-Based ADAS Crash Reconstruction

Dhruv Hegde

Sri Balaji

NADIR Research

Technical Report NADIR-TR-2026-03

October 2026 — **Draft for founder review. Pre-registered protocol; contains no results.**

Abstract

Courts increasingly hear disputes about whether a vehicle’s driver-assistance system could have perceived a hazard and should have intervened, and the evidence available after a crash is usually a single dashcam or onboard video plus the event data recorder (EDR) pre-crash record. Reconstructions built from such evidence are only as useful as their known error. This report pre-registers the protocol NADIR Research will use to measure the error of its video-based reconstruction pipeline before any accuracy figure is published. We specify the quantities estimated (ego and actor speed, position, relative gap, time to collision, event timing, and the probability that a modelled sensor suite would have detected a given actor), the ground-truth sources in three stages (public driving datasets with lidar and inertial ground truth, instrumented closed-track runs with RTK-GNSS on every vehicle, and blind comparison against credentialed reconstruction experts on closed cases), the metric definitions and interval coverage tests, the reporting rules, and the threats to validity. Acceptance thresholds are deliberately left as explicitly marked placeholders to be fixed by the team before data collection. No results are reported here.

1 Status and scope

This document is a protocol, not a results paper. It exists so that the metrics, data splits, and reporting rules are fixed *before* we look at outcomes, which is the main defence against choosing the flattering metric after the fact. Every threshold in this report is marked **[TBD by team: value]**; the team will fix those values, record the date, and version this document before any stage-2 data is processed. Any later change to a frozen item will be logged in Section 10 with its reason.

Out of scope: claims about any manufacturer’s internal perception software, safety certification of any system, and real-time or in-vehicle operation. The pipeline under test is an offline, post-crash reconstruction tool.

Engineering rule carried over from prior work: synthetic data is used only for debugging. No figure derived from synthetic data will be reported as a measure of forensic accuracy.

2 Problem

Given (i) a video V from a camera of unknown or partially known intrinsics mounted in or on a vehicle, and (ii) optionally the EDR pre-crash record E (under 49 CFR Part 563, at minimum

vehicle speed, accelerator position and service-brake status for the 5.0 s before the trigger, sampled at 2 Hz), estimate:

1. the camera model (focal length, distortion, rolling-shutter line delay, mount pose, frame timestamps);
2. the ego trajectory $\mathbf{x}^e(t) \in SE(3)$ at metric scale;
3. the trajectory $\mathbf{x}^a(t)$ of each relevant actor a ;
4. derived quantities: speeds, relative gap, closing speed, time to collision, and the time of key events (braking onset, first visibility, impact);
5. for a specified ADAS sensor suite $\mathcal{S}$ (field of view, range, detection envelope, latency), the probability $p_{\mathcal{S}}(a, t)$ that $\mathcal{S}$ would have detected actor a at time t , and the latest time at which an intervention could have avoided or mitigated impact.

Every estimate must carry an uncertainty statement (a standard deviation or a central interval at stated level). An estimate without an interval is scored as a failure on the coverage metrics in Section 4.

3 Ground-truth sources

Ground truth is collected in three stages of increasing realism. A stage is not started until the previous stage’s report is written.

Stage	Source	What it provides
1	KITTI raw / odometry	Front camera video with OXTS GNSS/INS poses and Velodyne lidar; ego speed and pose ground truth; tracked actors with 3D boxes on labelled subsets.
1	nuScenes	Six-camera video, lidar, ego localisation, 3D actor annotations at 2 Hz keyframes.
1	Waymo Open Dataset (Perception)	Camera, lidar, ego poses, 3D actor tracks.
2	Instrumented closed-track runs (target: Mcity, Ann Arbor)	RTK-GNSS/INS on the ego vehicle and every target (soft target or guided vehicle), synchronised to a common time base; production dash-cams and a phone mounted at varied poses; EDR downloads where the test vehicle supports them; near-miss and AEB-style scenarios that do not require a real collision.
3	Blind closed-case comparison	Closed, adjudicated cases with video and EDR, reconstructed independently by NADIR and by ACTAR-accredited reconstructionists who do not see each other’s work; an adjudicator compares both against physical evidence of record.

Table 1: Ground-truth stages. Stage 2 venue is a target, not an agreement.

Stage 1 tests geometry and tracking on real imagery but with clean, factory calibrated cameras and no crash dynamics. Stage 2 is the primary error-rate measurement: real consumer cameras, real mounts, real timing jitter, and centimetre-level truth on every vehicle. Stage 3 does not have

instrument ground truth; it measures agreement with the current expert standard of care and is reported separately, never pooled with stage 2.

4 Metrics

Let i index evaluation instances (a clip and a query time or event), y_i the ground-truth value, $\hat{y}_i$ the estimate and $\hat{\sigma}_i$ its reported standard deviation.

Point error. For each scalar quantity we report

$$e_i = \hat{y}_i - y_i, \quad (1)$$

$$\text{MAE} = \frac{1}{n} \sum_i |e_i|, \quad \text{RMSE} = \sqrt{\frac{1}{n} \sum_i e_i^2}, \quad \text{bias} = \frac{1}{n} \sum_i e_i, \quad (2)$$

plus the median and 95th percentile of $|e_i|$. Means are reported with bootstrap 95% confidence intervals resampled at the clip level (not the frame level), because frames within a clip are not independent.

Speed error. $y_i = v(t_i)$, the ground-truth speed in m/s of the ego vehicle or an actor at query time t_i . Query times are fixed per clip in advance: impact (or closest approach), braking onset, and 1, 2 and 3 s before impact. Reported separately for ego and for actors, and as relative error $e_i/v(t_i)$ for $v(t_i) > [\text{TBD by team: minimum speed, m/s}]$.

Position error. For ego and each actor, the Euclidean error in the road plane at each query time,

$$e_i^{\text{pos}} = \|\hat{\mathbf{p}}(t_i) - \mathbf{p}(t_i)\|_2, \quad (3)$$

computed *without* similarity alignment: scale and pose must be recovered by the pipeline, not fitted to ground truth after the fact. We separately report the relative-gap error $|\hat{d}_i - d_i|$ where $d_i = \|\mathbf{p}^a(t_i) - \mathbf{p}^e(t_i)\|_2$, since gap matters more than absolute position in most disputes.

Time-to-collision error. With closing speed $\dot{d}(t) < 0$, $\text{TTC}(t) = -d(t)/\dot{d}(t)$. We report $e_i^{\text{TTC}} = \widehat{\text{TTC}}(t_i) - \text{TTC}(t_i)$ for instances with $\text{TTC}(t_i) \leq [\text{TBD by team: horizon, s}]$; instances with non-closing motion are excluded and counted.

Event-timing error. For each event (braking onset, first frame in which the actor is visible, impact), $e_i^\tau = \hat{\tau}_i - \tau_i$ in milliseconds, with ground truth from RTK/INS acceleration (stage 2) or EDR (where it is the only source, flagged as such).

Video–EDR alignment error. Where both are available, the error of the estimated clock offset between video and EDR time bases, and the agreement between video-derived speed and EDR speed resampled to 2 Hz.

Interval coverage. For a nominal central interval at level $1-\alpha$, $\hat{C}_i = [\hat{y}_i - z_{1-\alpha/2}\hat{\sigma}_i, \hat{y}_i + z_{1-\alpha/2}\hat{\sigma}_i]$ (or the pipeline’s own interval if non-Gaussian),

$$\text{cov}_{1-\alpha} = \frac{1}{n} \sum_i \mathbf{1}[y_i \in \hat{C}_i], \quad (4)$$

reported at $1 - \alpha \in \{0.5, 0.8, 0.9, 0.95\}$ with clip-level bootstrap intervals. Under-coverage is treated as more serious than over-coverage because it overstates certainty in a legal setting. We also report mean interval width so that coverage cannot be bought with uninformative intervals.

Detection-envelope calibration. Stage 2 only, using test vehicles whose ADAS warnings or interventions can be logged. For each modelled sensor-suite query, $p_i = p_S(a, t_i)$ and the observed outcome $o_i \in \{0, 1\}$ (warning or intervention observed by t_i):

$$\text{Brier} = \frac{1}{n} \sum_i (p_i - o_i)^2, \quad (5)$$

$$\text{ECE} = \sum_{b=1}^B \frac{|I_b|}{n} \left| \frac{1}{|I_b|} \sum_{i \in I_b} o_i - \frac{1}{|I_b|} \sum_{i \in I_b} p_i \right|, \quad (6)$$

with $B = \text{[TBD by team: number of bins]}$ equal-mass bins, and a reliability diagram. This measures only whether our *model* of the sensor suite is calibrated against the behaviour we can observe; it says nothing about the internal software of any manufacturer.

Failure accounting. Clips on which the pipeline declines to produce an estimate (for example, insufficient lane markings to recover scale) are counted and reported as a rate, broken down by reason. Declining is preferable to a confident wrong answer, but a high decline rate is itself a finding.

5 Acceptance thresholds

The following are placeholders and must be fixed by the team before stage 2 data is processed. They are not targets we have met.

Quantity	Threshold
Ego speed MAE at impact (stage 2)	[TBD by team: m/s]
Actor speed MAE at impact (stage 2)	[TBD by team: m/s]
Relative-gap error, 95th percentile, at -1 s	[TBD by team: m]
TTC error, 95th percentile	[TBD by team: s]
Braking-onset timing error, 95th percentile	[TBD by team: ms]
95% interval coverage, lower bound of bootstrap CI	[TBD by team: fraction]
Detection-envelope ECE	[TBD by team: value]
Maximum decline rate	[TBD by team: fraction]

Table 2: Acceptance thresholds (placeholders).

6 Data splits and blinding

- Stage 1: official validation splits only; the pipeline is developed on training splits. Test splits with withheld labels are submitted once.
- Stage 2: runs are assigned to development or held-out sets by a seeded random draw recorded before the runs take place. Held-out ground truth is stored by a team member who does not work on the pipeline and is released once per frozen pipeline version, identified by commit hash.

- Stage 3: cases are selected by the partner experts; NADIR receives only the evidence that would have been available to a retained expert.

7 Reporting rules

1. Every reported number names its stage, dataset, split, pipeline commit, and sample size in clips.
2. All metrics in Section 4 are reported, including those that look bad. No metric is dropped after outcomes are seen.
3. Stage 1, 2 and 3 results are never pooled.
4. Results are published as a numbered NADIR Research technical report with a DOI, and the website states the same numbers with a link to that report.

8 Threats to validity

- **Dataset camera bias.** Public datasets use rigidly mounted, factory-calibrated cameras; consumer dashcams have wide-angle lenses, rolling shutters, compression, and variable frame timing. Stage 1 error will understate field error.
- **Track realism.** Closed-track runs avoid real collisions and use soft targets; actor appearance and dynamics differ from real crashes.
- **Scale cues.** Metric scale from lane-marking standards and vehicle dimensions fails where markings are worn, non-standard, or absent, or where the vehicle model is misidentified.
- **EDR limitations.** EDR speed is wheel-speed derived and can be biased under braking lock-up or wheel slip; the 2 Hz sampling limits timing. EDR is a ground-truth source only where nothing better exists, and is flagged.
- **Sensor-suite models.** Our models of production sensor suites are built from public specifications and observed behaviour, not from manufacturer internals; the calibration metric measures that model only.
- **Expert comparison.** Stage 3 measures agreement with experts, whose own error is unknown; agreement is not accuracy.
- **Selection.** Cases with usable video may differ systematically from cases without it.

9 Data and licence constraints

- KITTI is distributed under CC BY-NC-SA 3.0; nuScenes is CC BY-NC-SA 4.0 for non-commercial use with a separate commercial licence; the Waymo Open Dataset has its own non-commercial licence. Benchmarking for a published research report must be reviewed against each licence, and no model trained on these datasets ships in a commercial product without a licence that permits it.

- MAST3R and DUST3R (dense multi-view reconstruction models) are released under CC BY-NC-SA 4.0, i.e. non-commercial. They may be used, if at all, only as research baselines in this benchmark and not inside the product.
- Closed-case video and EDR data belong to the parties and their counsel. They are used only under written agreement, stored encrypted, and never published; aggregate metrics only.
- Faces and licence plates in any released clip are blurred.

Licence terms above are as understood at drafting and must be re-checked by the team before use.
[TBD by team: legal review sign-off]

10 Change log

Date	Change
2026-10	Initial draft for founder review.

References to standards

- 49 CFR Part 563, Event Data Recorders.
- FMVSS No. 127, Automatic Emergency Braking Systems for Light Vehicles (compliance required from September 2029).
- Federal Rule of Evidence 702; *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993).