Technical report · IF-TR-2026-04 · version 1.1

A confidence-gated annotation and data engine for driving data: LabeloxAV v0.1.1 and its public evidence

Sherin Joseph Roy 11 November 2026 LabeloxAV v0.1.1 P2 · Data engine Funded entirely by BluRabbit Mobility Private Limited.

Abstract

LabeloxAV is an annotation and data engine for driving data. It machine-labels camera frames through three detection paths, calibrates and gates each proposal so that only confident, agreed detections are accepted without a person, routes the rest to human review, tracks objects across frames, lifts boxes into 3D, builds HD map layers with Lanelet2 and OpenDRIVE export, and exports datasets under a governance layer that keeps model output apart from human annotation. This report describes the design of release v0.1.1 and its evidence. The author reproduced the CPU unit tier and the blocking static gates at the release tag; one unit test failed because it opens a database connection without the marker that excludes it. Results reported in the repository cover a full-infrastructure test run, two detectors trained and evaluated on the public India Driving Dataset, and LiDAR ground-plane and HD map checks on public KITTI data, whose map geometry the documentation itself does not yet trust. Version 1.1 adds results on the lab's own driving corpus: machine-judged label precision by class, detectors trained on it, and a champion detector whose two gold-set scores conflict.

1 Problem

Labelling driving footage by hand is slow and expensive, and the cases that matter most are the hardest to find in it. Public driving datasets under-represent much of what appears on Indian roads, including autorickshaws, cattle on the carriageway, overloaded two-wheelers, hand carts and potholes, so models trained on them perform poorly on that long tail. LabeloxAV [1] addresses the labelling side of this problem: it turns raw dashcam video and recorded sensor drives into training data by machine-labelling them, sending uncertain and safety-critical labels to a person, and allowing a retrained model to replace the current one only through a governance gate.

The engineering question this report examines is narrower than the product. It asks what, at release v0.1.1, can be shown about the engine's correctness and quality using evidence that is either reproducible from the public repository or measured on public data. Quality numbers computed on private footage cannot be checked by a reader, and an engine whose stated purpose is to make label quality auditable should be judged first on what can be audited.

2 System design

The system is a Python 3.11 FastAPI backend over PostgreSQL, MinIO, Redis and Redpanda, with a Next.js web application that reads and writes only through the API. It runs as Docker containers, installed by script or Debian package.

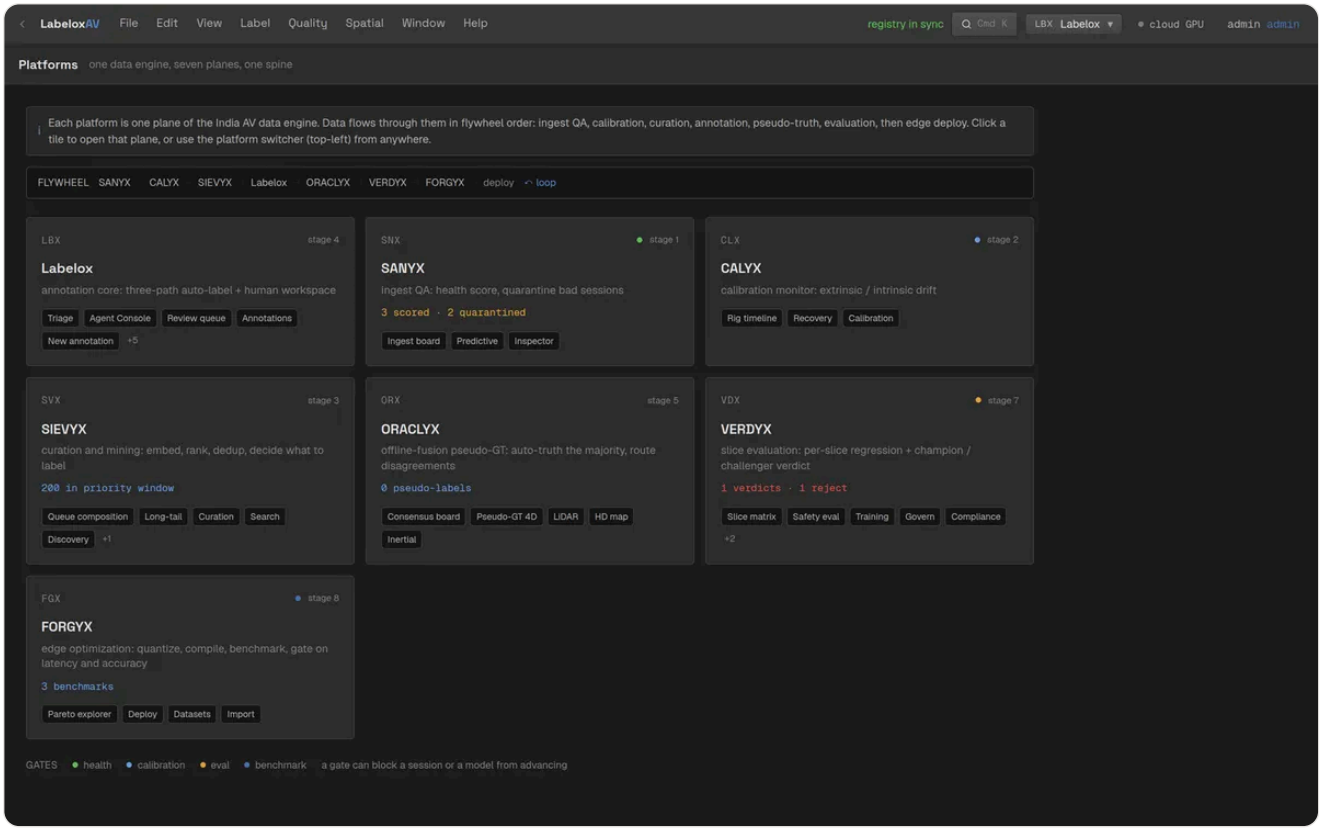


Figure 1. The platform launcher: the planes of the data engine in flywheel order. Screenshot from the repository (docs/screenshots/20-ui-platforms.png).

Data model

Everything hangs off three tables: a session (one recording), a frame (one camera image at one timestamp) and an object (one annotation on one frame, with class, geometry, confidence, state, source and provenance). An object becomes human-owned once a person touches it. Model output is therefore kept in a separate, append-only prediction plane (an inference run and its predictions) that review never writes to; evaluation scores a named inference run, not live corpus state. The decision is recorded in `docs/adr/0001-prediction-plane.md`.

Auto-labelling behind a confidence gate

Three paths propose labels: a closed-set YOLO11 detector [10], an open-vocabulary detector paired with SAM 2 segmentation [11], and a vision-language verifier served locally. Fused proposals are calibrated with isotonic regression [14] and routed by `services/autolabel/gate.py` to auto-accept, review or annotate. Auto-accept requires cross-path agreement as well as confidence, safety-critical classes (vulnerable road users and animals) use a higher threshold, and a rare or fallback class can be auto-accepted only with agreement and verifier confirmation. The gate's source and the README both state that, except where a per-class fit exists, the thresholds are configured constants rather than measured precision floors. A governance kill switch disables auto-accept entirely.

Review, tracking and active learning

Review confirms, reclassifies, adjusts or rejects an object and writes the before-state for audit. Tracking (`services/autolabel/track/`) is a BoT-SORT-style tracker [12] with a constant-velocity Kalman model, appearance features from the stored object embeddings and Hungarian assignment; it flags likely identity switches, and only detector output is fed back into it. Active learning ranks candidates by uncertainty, diversity, rarity and error-proneness to build the next review batch.

Lifting boxes into 3D

With a synchronised point cloud, `services/lidar/detect3d/lift.py` forms a frustum from the 2D box, gathers the enclosed points, fits an oriented cuboid in the bird's-eye view and snaps it to a ground plane fitted by RANSAC [13]. Without LiDAR, `services/oraclyx/mono_depth.py` estimates metric range from a ground-contact prior and a known-class-height prior, fused by inverse variance, and returns nothing when the camera is uncalibrated.

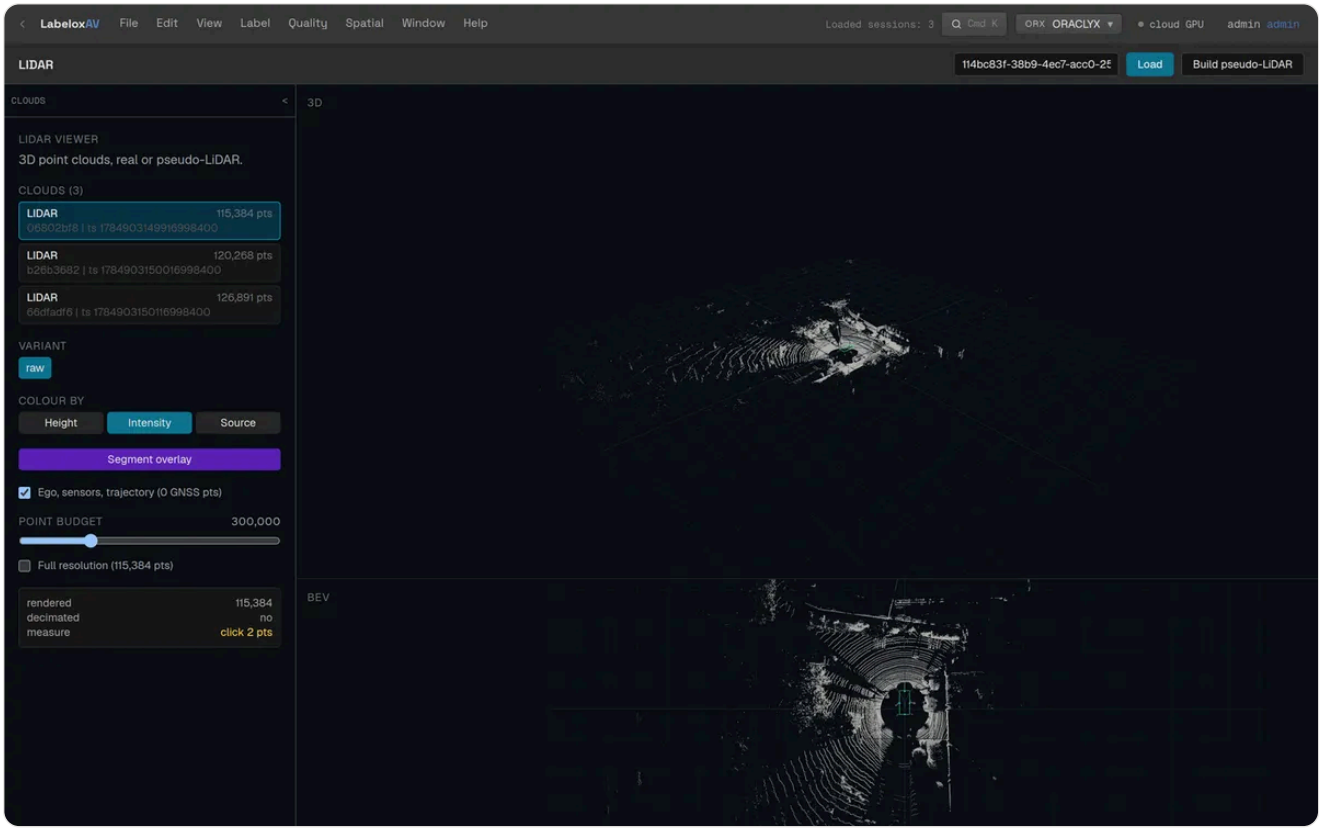


Figure 2. The LiDAR viewer showing a public KITTI scan in 3D and bird's-eye view. Screenshot from the repository (docs/screenshots/42-lidar-viewer.png); KITTI data, CC BY-NC-SA 3.0.

HD map and dataset export

`services/hdmap/` georeferences lane elements using each session's resolved camera calibration, fuses them into boundaries, pairs boundaries into lanelets under conservative heading, overlap, separation and sign-consistency conditions, and exports Lanelet2 [5] and OpenDRIVE [6]. Dataset export (`services/export/`) has adapters for COCO [9], YOLO, Pascal VOC, KITTI, BDD100K, Mapillary Vistas, OpenLABEL [7], nuScenes [8] (including 3D cuboids), CVAT, Label Studio and Parquet, and ships a coverage datasheet with each export.

Governance

A promotion gate (`services/govern/champion.py`) is a pure decision over a metrics dictionary. It refuses a challenger that does not beat the champion, regresses a safety-class floor, was reconstructed from review history, or shows divergence between the two evaluation harnesses (an Ultralytics validation pass and the prediction-plane scorer, which computes a 101-point interpolated AP). Ingest writes a per-frame anonymisation audit, with faces and plates blurred before storage. An import-linter contract keeps the domain-neutral engine from importing a concrete domain pack. The README counts the ontology at one hundred and seventy-eight governed classes.[1]

1. The engineering log's feature list gives one hundred and seventy classes; the README and the release notes give one hundred and seventy-eight.

3 Method

The author checked out the repository at tag v0.1.1 (commit 9f2fc6f) and read its README, documentation and changelog as text. No media, dataset or model weights were opened. Two kinds of evidence result.

Reproduced. On a CPU-only workstation with no Postgres, MinIO, Redis or Docker running, the author ran the repository's unit tier, which `make test-unit` defines as the marker filter `not db and not gpu and not infra`. It ran twice: once with the base and development dependencies only, and once adding the CPU builds of PyTorch and torchvision, `imagehash` and `ultralytics` that the repository's CI installs. The optional `ml` extra was not installed because it pins CUDA wheels. The author counted the excluded tests per marker with `pytest --collect-only`, and ran the three blocking static gates from `.github/workflows/ci.yml`: the import contract, `mypy --strict` on the typed allowlist, and `ruff` on the blocking module set.

Reported. Results that need infrastructure, a GPU or data the author did not run are quoted from the repository at the commit shown in each evidence row: the verification pass in `docs/VERIFICATION.md`, the recorded baseline in `tests/KNOWN_FAILURES.md`, the detector table and LiDAR checks in `docs/ENGINEERING_LOG.md`, and the install figures in the README. Results measured on the lab's own driving corpus are quoted from the per-class

precision reports of 26 and 27 August 2026 (`reports/class_precision_*.json`), `docs/MEASUREMENT.md`, `docs/AUTONOMY.md`, `models/README.md` and the README. The corpus itself is not published. Version 1.1 of this report (11 November 2026) adds these results, which version 1.0 withheld.

4 Validation

Reproduced unit tier

With the CPU wheels installed, the unit tier gave `2045 tests` passed, `1 tests` failed and `11 tests` skipped, in a wall time of `64.76 s`. The marker filter excluded `1470 tests` that need Postgres, MinIO, Redis or another service, and `5 tests` that need a CUDA device or downloaded weights. The skips are named in the log: four need CUDA, four need the `transformers` package from the `ml` extra, and one each needs a dashcam sample not in the repository, the depth model stack, or licence-plate weights that are fetched separately. Without PyTorch, the same tier gave `2035 tests` passed and `21 tests` skipped, with the same single failure.

The failure is `tests/test_privacy.py::TestDistillation::test_an_unregistered_teacher_is_refused`. The test opens a Postgres session through `get_sessionmaker()` but carries no `db` marker, so the unit tier selects it and it fails with a refused connection to the local Postgres port. The repository's guard, `tests/test_db_markers.py`, checks markers per file, and `test_privacy.py` carries a `db` marker on a different class, so the guard passes. The README describes the unit tier as green on a fresh clone; at this tag it is not.

`docs/TESTING.md` documents the unit tier as `1449 tests` passed. That figure and the lab's reproduced count differ. The documentation is dated 2026-08-18, before the release. The difference most likely reflects tests added after the document was written rather than a change in outcome.

Quantity	Value	Conditions	Basis
Unit-tier tests passed	`2045 tests`	of 2057 tests selected from 3532 collected; result line '1 failed, 2045 passed, 11 skipped, 1475 deselected'; 1475 deselected by marker (1462 db + 8 infra = requires services; 5 gpu = requires hardware) Differs from another source; both are listed.	Reproduced 2026-11-10
Unit-tier tests failed	`1 tests`	of 2057 tests selected from 3532 collected; result line '1 failed, 2045 passed, 11 skipped, 1475 deselected'; the failure is tests/test_privacy.py::TestDistillation::test_an_unregistered_teacher_is_refused, which opens a Postgres session but carries no db marker (ConnectionRefusedError to 127.0.0.1:5432)	Reproduced 2026-11-10
Unit-tier tests skipped	`11 tests`	of 2057 tests selected from 3532 collected; result line '1 failed, 2045 passed, 11 skipped, 1475 deselected'; 4 need CUDA, 4 need transformers/SigLIP 2 (ml extra), 1 needs a dashcam sample not in the repo, 1 needs the depth model stack, 1 needs plate weights (make pii-models, not fetched)	Reproduced 2026-11-10
Tests excluded as requiring services	`1470 tests`	1462 marked db (Postgres, usually MinIO) + 8 marked infra (Redis or another service), of 3532 collected; counted with pytest --collect-only -m <marker>	Reproduced 2026-11-10
Tests excluded as requiring hardware	`5 tests`	marked gpu (CUDA GPU or downloaded model weights), of 3532 collected	Reproduced 2026-11-10
Unit-tier suite wall time	`64.76 s`	wall time of the whole command incl. interpreter start; pytest itself reported 59.25s; 2057 tests selected	Reproduced 2026-11-10
Unit-tier tests passed, without torch installed	`2035 tests`	of 2057 selected; result line '1 failed, 2035 passed, 21 skipped, 1475 deselected in 97.24s'; same single failure (test_privacy.py, unmarked Postgres use)	Reproduced 2026-11-10
Unit-tier tests skipped, without torch installed	`21 tests`	of 2057 selected; result line '1 failed, 2035 passed, 21 skipped, 1475 deselected in 97.24s'; 18 skip on missing torch/torchvision, 3 on missing sample data, depth stack or plate weights	Reproduced 2026-11-10
Unit-tier tests passed (as documented)	`1449 tests`	make test-unit (-m 'not db and not gpu and not infra') with Postgres stopped; 'about 60 seconds'; documented 2026-08-18, older than v0.1.1 Differs from another source; both are listed.	Reported 2026-08-18

Table 1. Unit tier (make test-unit marker filter) at v0.1.1 on a CPU-only workstation with no services running, with and without the CPU PyTorch wheels, and the count the testing guide documents; reproduced by the author except the last row, reported in the repository. "Reproduced" rows were run by the author; "Reported" rows are quoted from the repository at the commit shown.

Static gates

The import contract that keeps the engine core from importing a concrete domain pack held (contracts kept: `1 contracts`; none broken). `mypy --strict` reported `0 issues` on the eight allowlisted modules, and `ruff` reported `0 findings` on the blocking set. These gates cover the security and measurement modules that CI treats as blocking; the rest of the tree is checked by CI only informationally and was not run here.

Quantity	Value	Conditions	Basis
Import-linter contracts kept	1 contracts	'Contracts: 1 kept, 0 broken'; 770 files, 3076 dependencies analysed	Reproduced 2026-11-10
mypy --strict findings on the CI allowlist	0 issues	'Success: no issues found in 8 source files' (the blocking strict allowlist in .github/workflows/ci.yml)	Reproduced 2026-11-10
Ruff findings on the CI blocking module set	0 findings	'All checks passed!' on the blocking set listed in .github/workflows/ci.yml	Reproduced 2026-11-10

Table 2. Blocking static gates from the CI workflow, run at v0.1.1 with Python 3.11; reproduced by the author. “Reproduced” rows were run by the author; “Reported” rows are quoted from the repository at the commit shown.

Full-infrastructure verification

docs/VERIFICATION.md records a pass run on 2026-09-12 against the live system with its infrastructure up (the evidence notes that make test needs Postgres, MinIO, Redis and a GPU): 3506 tests passed, 0 tests failed, 6 tests skipped and 4 tests expected failures, with 703 tests passed in the web suite. An earlier baseline in tests/KNOWN_FAILURES.md, measured 2026-08-18, records 2488 tests passed. The two runs are from different dates and are not in conflict. The hardware for neither is stated, and the lab did not reproduce either.

Quantity	Value	Conditions	Basis
Python tests passed (full suite, full infra)	3506 tests	full Python suite run on 2026-09-12 against the live system with infrastructure up; result line: 3,506 passed, 6 skipped, 4 xfailed, 0 failed	Reported 2026-09-12
Python tests failed (full suite, full infra)	0 tests	full Python suite run on 2026-09-12 against the live system with infrastructure up; result line: 3,506 passed, 6 skipped, 4 xfailed, 0 failed	Reported 2026-09-12
Python tests skipped (full suite, full infra)	6 tests	full Python suite run on 2026-09-12 against the live system with infrastructure up; result line: 3,506 passed, 6 skipped, 4 xfailed, 0 failed	Reported 2026-09-12
Python tests xfailed (full suite, full infra)	4 tests	full Python suite run on 2026-09-12 against the live system with infrastructure up; result line: 3,506 passed, 6 skipped, 4 xfailed, 0 failed	Reported 2026-09-12
Web (frontend) tests passed	703 tests	across 62 files, verification pass of 2026-09-12	Reported 2026-09-12
Python tests passed (recorded full-infra baseline)	2488 tests	measured 2026-08-18 against full infra (make up) on the test database; 3 skipped, 4 xfailed	Reported 2026-08-18

Table 3. Full Python and web suites with infrastructure up (make up and make test; npm test in web/), on the dates shown; hardware not stated; reported in the repository. “Reproduced” rows were run by the author; “Reported” rows are quoted from the repository at the commit shown.

Detectors on the India Driving Dataset

The engineering log reports two detectors from the model registry, trained on the public India Driving Dataset [2] on a single consumer GPU. The larger YOLO11l model reached 0.44 mAP@50, with precision 0.67 and recall 0.39. The smaller YOLO11n reached 0.34, with precision 0.67 and recall 0.3. The documentation does not state the evaluation split, the IoU threshold for precision and recall, or the confidence threshold, and gives no command to reproduce the figures. The log describes them as weak on the rare long tail.



Figure 3. mAP@50 of the two IDD-trained detectors; evaluation split and thresholds not stated in the source; reported in the repository.

Quantity	Value	Conditions	Basis
mAP@50, idd-yolo11n (YOLO11n) on IDD	0.34	India Driving Dataset (public); evaluation split and IoU/confidence settings not stated in the log	Reported 2026-09-15
mAP@50, idd-yolo11l (YOLO11l) on IDD	0.44	India Driving Dataset (public); evaluation split and IoU/confidence settings not stated in the log	Reported 2026-09-15
Precision, idd-yolo11n (YOLO11n) on IDD	0.67	India Driving Dataset (public); evaluation split and IoU/confidence settings not stated in the log	Reported 2026-09-15
Precision, idd-yolo11l (YOLO11l) on IDD	0.67	India Driving Dataset (public); evaluation split and IoU/confidence settings not stated in the log	Reported 2026-09-15
Recall, idd-yolo11n (YOLO11n) on IDD	0.3	India Driving Dataset (public); evaluation split and IoU/confidence settings not stated in the log	Reported 2026-09-15
Recall, idd-yolo11l (YOLO11l) on IDD	0.39	India Driving Dataset (public); evaluation split and IoU/confidence settings not stated in the log	Reported 2026-09-15

Table 4. Detectors trained and evaluated on the public India Driving Dataset; split, IoU and confidence settings not stated; reported in the repository. “Reproduced” rows were run by the author; “Reported” rows are quoted from the repository at the commit shown.

LiDAR ground plane and HD map on KITTI

Three Velodyne scans from the public KITTI object set [4] were passed through the LiDAR path. RANSAC ground segmentation placed the road plane `1.71 m`, `1.68 m` and `1.69 m` below the sensor, close to the mounting height KITTI documents for its scanner [3]. This checks the geometry of ingest, ground segmentation and rendering; it does not train or evaluate a 3D detector.

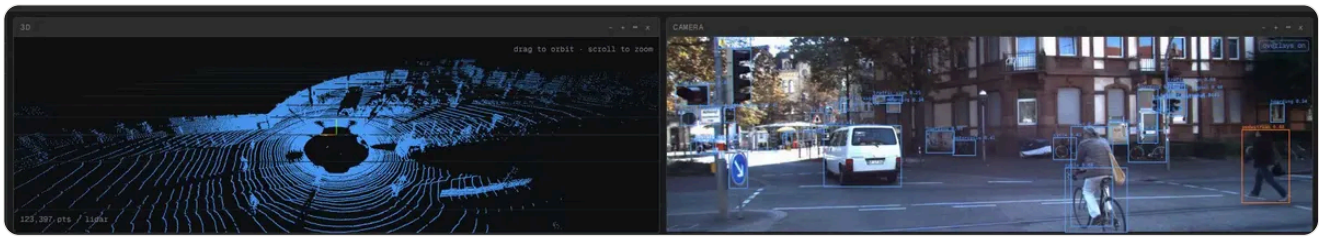


Figure 4. The session inspector on a KITTI raw drive: the scanner's point cloud beside the synchronised camera frame with label boxes. Frame from the repository's verification recording (docs/demo/labeloxav-verification.mp4), cropped; KITTI data, CC BY-NC-SA 3.0.

On KITTI raw drive 0005 [3], which carries a position fix on every frame, the HD map pipeline ran end to end for the first time: lanes were proposed, georeferenced, fused, paired and exported. The Lanelet2 export contained `29 relations` and the OpenDRIVE export `58 lanes`, where both previously contained none. The same document states that the geometry is not yet trustworthy (see Limitations), so these counts show that the exports are now populated and routable, not that the map is correct.

Quantity	Value	Conditions	Basis
RANSAC ground plane depth below sensor, KITTI scan 1 of 3	1.71 m	three Velodyne HDL-64E scans from the public KITTI object set; KITTI mounts the sensor at roughly 1.73 m	Reported 2026-09-15
RANSAC ground plane depth below sensor, KITTI scan 2 of 3	1.68 m	three Velodyne HDL-64E scans from the public KITTI object set; KITTI mounts the sensor at roughly 1.73 m	Reported 2026-09-15
RANSAC ground plane depth below sensor, KITTI scan 3 of 3	1.69 m	three Velodyne HDL-64E scans from the public KITTI object set; KITTI mounts the sensor at roughly 1.73 m	Reported 2026-09-15
Lanelet2 export lanelet relations from one KITTI raw drive	29 relations	KITTI raw drive 0005 (public); the doc states the map geometry is not yet trustworthy (one fused boundary 8,439 m long; median paired lane 2.54 m wide)	Reported 2026-09-12
OpenDRIVE export driving lanes from one KITTI raw drive	58 lanes	KITTI raw drive 0005 (public); 48 roads, 1 road link; the doc states the map geometry is not yet trustworthy	Reported 2026-09-12

Table 5. RANSAC ground plane on three public KITTI object scans, and HD map exports from one public KITTI raw drive; reported in the repository. “Reproduced” rows were run by the author; “Reported” rows are quoted from the repository at the commit shown.

On the lab’s own driving corpus

The corpus is the lab’s own driving footage, machine-labelled by the engine. As of 27 August 2026, `1577 objects` carried a human verdict; the row gives the share of the corpus. Most quality figures below are therefore judged by a machine or measured against an incomplete key, and the text says which.

Label precision by class. A vision-language judge rated a random sample of 80 machine labels per class, leaving the crops it was unsure of undecided. On 27 August 2026, after a track-interpolation pass was reverted, judged precision ranged from `0.0526` for traffic signals and `0` for the fallback class to `0.871` for motorcycles and `0.8667` for pedestrians; sedans scored `0.2368`. These are machine judgements on small samples, and the intervals in the evidence rows are wide. The figures from the day before the revert are in the appendix. The subset the gate accepted automatically was judged `0.93` precise by the same judge, under its strict rule.

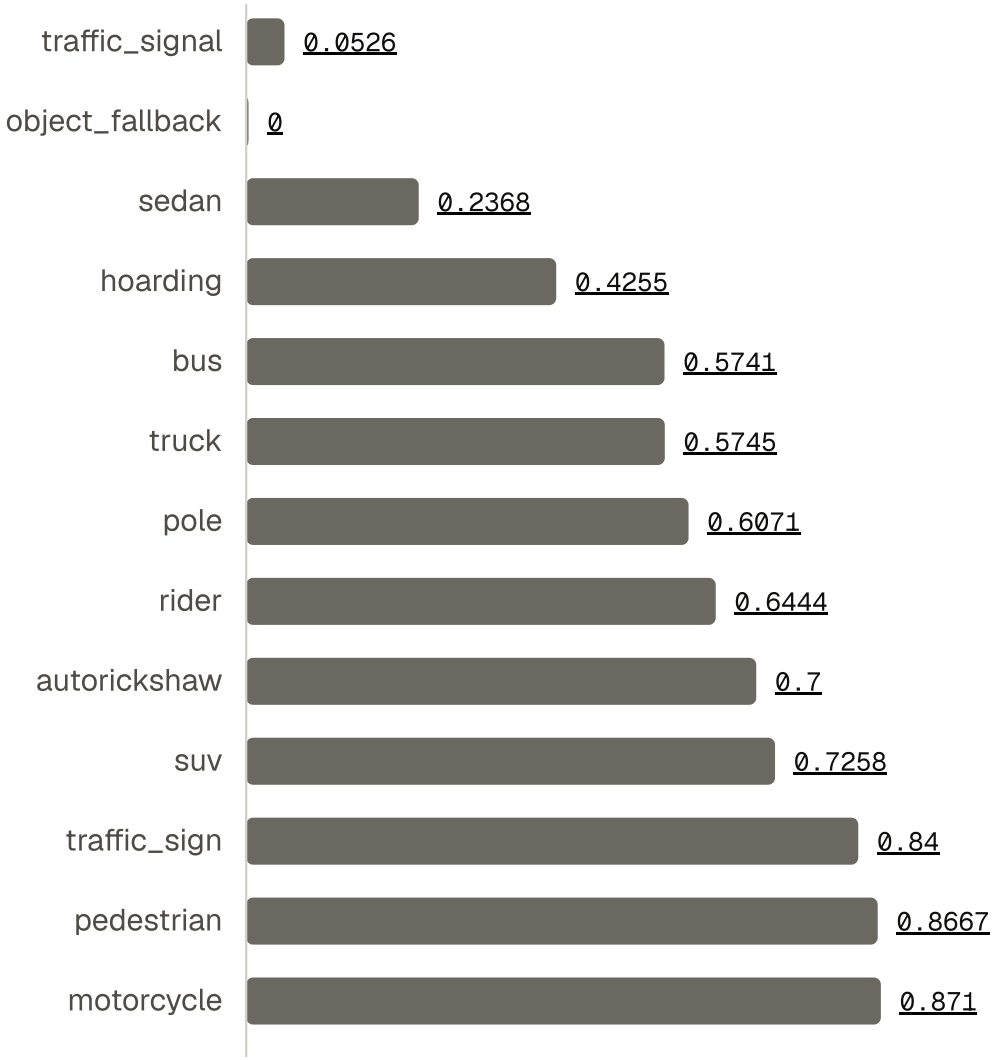


Figure 5. Judged label precision by class on the lab's corpus, 27 August 2026: a vision-language judge on a random sample of 80 machine labels per class, undecided crops excluded; machine-judged, not human-verified. Reported in the repository, reports/class_precision_2026-08-27.json.

Detectors trained on the corpus. Trained on gate-approved detections only, a YOLO11n reached a recall of `0.411` (micro) and `0.282` (macro), and a YOLO11s `0.386` and `0.281`. Trained on every detection a person did not refuse, a YOLO11n reached `0.77` and `0.623`. Recall is measured against an incomplete key, so the repository states precision and mAP as not measurable on this corpus; none of these candidates was promoted.

The champion detector on a gold set. Through the prediction plane, the current champion, an IDD-trained YOLO11l, scores AP@50 `0.083`, AP@50:95 `0.068`, precision `0.164` and recall `0.146` on a 400-object human-verified gold slice, but 351 of its 400 objects have unrecoverable provenance. The model registry records mAP@50 `0.4407` and recall `0.3931` on a sealed gold set of 202 frames. The documentation notes that the registry’s column holds two different yardsticks, and does not state whether the two gold sets are the same. The two pairs of figures differ; both are listed.

Segmentation agreement. Masks from SAM 2 agree with a SAM 1 re-prompt at a median IoU of `0.893 IoU`; about a fifth fall below 0.8, concentrated on riders, motorcycles and autorickshaws.

5 Limitations

The figures measured on the lab's own corpus are machine-judged or measured against an incomplete key. Label precision comes from a vision-language judge, not from people, on small samples with many undecided crops; few labels in the corpus carry a human verdict; and the champion detector's two gold-set results conflict, with most objects of the gold slice lacking recoverable provenance. No evidence item tests whether the confidence gate's thresholds achieve a particular precision; the gate's own documentation states that, outside per-class fits, they are unmeasured constants.

The HD map is not yet trustworthy, as the release notes and `docs/VERIFICATION.md` say: fusion can emit a boundary far longer than the drive it came from, and paired lane widths run narrower than a real lane, which indicates the pairing still matches elements that are not the two sides of one lane. The end-to-end map run covers one public drive.

The unit tier is not green on a clean CPU machine at this tag. One test opens a database session without a `db` marker, and the per-file marker guard does not catch it. The `1470 tests` and `5 tests` excluded from that tier, the GPU auto-labelling paths, the locally served verifier and the web suite were not run by the author. The full-infrastructure figures are reported, with hardware unstated, and were recorded three days before the release tag.

The detector results are reported without split, thresholds or a reproduction command, and the training hardware is described only as a single consumer GPU. The KITTI ground-plane check covers three scans from a public sample. No evidence item measures the accuracy of LiDAR cuboid lifting, the monocular depth prior or the tracker. The engineering log states that the multi-camera rig workspace has been driven only against synthetic sessions, since no multi-camera recording has been ingested.

On footprint, the release notes for v0.1.0 call four backing services not a light footprint. For v0.1.1 the README reports a first install of `6.5 minutes` leaving `4.5 GB` of images on a clean Ubuntu 24.04 without a GPU, timed once on an otherwise unstated machine. Release v0.1.0 could not complete its install and is superseded; this report describes v0.1.1 only.

6 Reproducibility

Version: LabeloxAV v0.1.1, commit `9f2fc6f929e51233e0d773987649683d910e7210`, Apache-2.0. Runtime for all reproduced rows: Python 3.11.15 on Linux 7.0.0-42 (Ubuntu), AMD Ryzen 9 9900X, 30 GB RAM, CPU only, with no Postgres, MinIO, Redis or Docker running. The unit-tier run set `YOLO_CONFIG_DIR` to a writable scratch directory.

```
git clone https://github.com/Sherin-SEF-AI/LabeloxAV && cd LabeloxAV && git checkout v0.1.1
uv venv --python 3.11 && uv pip install -e '.[dev]'

# static gates
.venv/bin/lint-imports
.venv/bin/mypy --strict --ignore-missing-imports core/accel/ap.py core/accel/recapture.py core/version.py \
  core/timebase.py services/api/auth_token.py services/review_policy.py services/anonymize/coverage.py \
  services/training/lineage.py
.venv/bin/ruff check core/accel core/version.py services/verdyx services/llm services/api/main.py \
  services/api/deps.py services/api/auth_token.py services/api/routers/predictions.py services/analytics/evaluation.py

# tests excluded by marker
.venv/bin/python -m pytest --collect-only -q -m db     # likewise -m infra, -m gpu

# unit tier, base dependencies only
.venv/bin/python -m pytest -q -m 'not db and not gpu and not infra' -p no:cacheprovider -rfEsxX

# unit tier with the CPU wheels CI installs
uv pip install torch==2.7.1 torchvision==0.22.1 --index-url https://download.pytorch.org/whl/cpu
uv pip install imagehash==4.3.1 ultralytics==8.3.0
.venv/bin/python -m pytest -q -m 'not db and not gpu and not infra' -p no:cacheprovider -rfEsxX
```

Not run by the author, and needing services or hardware: `make up && make test` (Postgres, MinIO, Redis, Redpanda and lakeFS, plus a CUDA GPU for the `gpu` tests), `cd web && npm test`, the auto-labelling, embedding and training paths (CUDA), the vision-language verifier (a local model server), and the installer `./scripts/install.sh` (Docker). The IDD training and evaluation, the KITTI ground-plane run and the KITTI HD map run have no documented command, and the KITTI raw drives are not part of the repository and carry their own licence (CC BY-NC-SA 3.0).

References

1. Sherin Joseph Roy. LabeloxAV, release v0.1.1 (commit 9f2fc6f). Source code, Apache-2.0. github.com/Sherin-SEF-AI/LabeloxAV/tree/v0.1.1
2. G. Varma, A. Subramanian, A. Namboodiri, M. Chandraker, C. V. Jawahar. IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments. WACV 2019. doi:10.1109/WACV.2019.00190

3. A. Geiger, P. Lenz, C. Stiller, R. Urtasun. Vision meets robotics: The KITTI dataset. International Journal of Robotics Research 32(11), 2013. doi:10.1177/0278364913491297
4. A. Geiger, P. Lenz, R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. CVPR 2012. doi:10.1109/CVPR.2012.6248074
5. F. Poggenhans et al. Lanelet2: A high-definition map framework for the future of automated driving. ITSC 2018. doi:10.1109/ITSC.2018.8569929
6. ASAM e.V. ASAM OpenDRIVE. Standard for the description of road networks. www.asam.net/standards/detail/opendrive/
7. ASAM e.V. ASAM OpenLABEL. Standard for the annotation format of multi-sensor data. www.asam.net/standards/detail/openlabel/
8. H. Caesar et al. nuScenes: A multimodal dataset for autonomous driving. CVPR 2020. doi:10.1109/CVPR42600.2020.01164
9. T.-Y. Lin et al. Microsoft COCO: Common objects in context. ECCV 2014. doi:10.1007/978-3-319-10602-1_48
10. G. Jocher, J. Qiu. Ultralytics YOLO11. Source code, 2024. github.com/ultralytics/ultralytics
11. N. Ravi et al. SAM 2: Segment anything in images and videos. arXiv:2408.00714, 2024. arxiv.org/abs/2408.00714
12. N. Aharon, R. Orfaig, B.-Z. Bobrovsky. BoT-SORT: Robust associations multi-pedestrian tracking. arXiv:2206.14651, 2022. arxiv.org/abs/2206.14651
13. M. A. Fischler, R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24(6), 1981. doi:10.1145/358669.358692
14. B. Zadrozny, C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. KDD 2002. doi:10.1145/775047.775151

Evidence

Quantity	Value	Conditions	Basis	Source and repr
Unit-tier tests passed	2045 tests	of 2057 tests selected from 3532 collected; result line '1 failed, 2045 passed, 11 skipped, 1475 deselected'; 1475 deselected by marker (1462 db + 8 infra = requires services; 5 gpu = requires hardware) Differs from another source; both are listed.	Reproduced 2026-11-10	command: pytho db and not gpu a no:cacheprovide test-unit tier) commit 9f2fc6f (9900X, 30 GB RA (Ubuntu), Python `git clone https://github. AI/LabeloxAV && checkout v0.1.1 3.11 && uv pip && uv pip insta torchvision==0. https://downloa u && uv pip ins imagehash==4.3. && .venv/bin/py 'not db and not -p no:cacheprov`
Unit-tier tests failed	1 tests	of 2057 tests selected from 3532 collected; result line '1 failed, 2045 passed, 11 skipped, 1475 deselected'; the failure is tests/test_privacy.py::TestDistillation::test_an_unregistered_teacher_is_refused, which opens a Postgres session but carries no db marker (ConnectionRefusedError to 127.0.0.1:5432)	Reproduced 2026-11-10	command: pytho db and not gpu a no:cacheprovide test-unit tier) commit 9f2fc6f (9900X, 30 GB RA (Ubuntu), Python `git clone https://github. AI/LabeloxAV && checkout v0.1.1 3.11 && uv pip && uv pip insta torchvision==0. https://downloa u && uv pip ins imagehash==4.3. && .venv/bin/py 'not db and not -p no:cacheprov`
Unit-tier tests skipped	11 tests	of 2057 tests selected from 3532 collected; result line '1 failed, 2045 passed, 11 skipped, 1475 deselected'; 4 need CUDA, 4 need transformers/SigLIP 2 (ml extra), 1 needs a dashcam sample not in the repo, 1 needs the depth model stack, 1 needs plate weights (make pii-models, not fetched)	Reproduced 2026-11-10	command: pytho db and not gpu a no:cacheprovide test-unit tier) commit 9f2fc6f (9900X, 30 GB RA (Ubuntu), Python `git clone https://github. AI/LabeloxAV && checkout v0.1.1 3.11 && uv pip && uv pip insta torchvision==0. https://downloa u && uv pip ins imagehash==4.3. && .venv/bin/py 'not db and not -p no:cacheprov`
Tests excluded as requiring services	1470 tests	1462 marked db (Postgres, usually MinIO) + 8 marked infra (Redis or another service), of 3532 collected; counted with pytest --collect-only -m <marker>	Reproduced 2026-11-10	command: pytho only -q -m <db|in commit 9f2fc6f (9900X, 30 GB RA

Quantity	Value	Conditions	Basis	Source and repr
				(Ubuntu), Python `git clone https://github. AI/LabeloxAV && checkout v0.1.1 3.11 && uv pip && .venv/bin/py collect-only -q`
Tests excluded as requiring hardware	`5 tests`	marked gpu (CUDA GPU or downloaded model weights), of 3532 collected	Reproduced 2026-11-10	command: pytho only -q -m gpu commit 9f2fc6f (9900X, 30 GB RA (Ubuntu), Python `git clone https://github. AI/LabeloxAV && checkout v0.1.1 3.11 && uv pip && .venv/bin/py collect-only -q`
Unit-tier suite wall time	`64.76 s`	wall time of the whole command incl. interpreter start; pytest itself reported 59.25s; 2057 tests selected	Reproduced 2026-11-10	command: pytho db and not gpu a no:cacheprovide test-unit tier) commit 9f2fc6f (9900X, 30 GB RA (Ubuntu), Python `git clone https://github. AI/LabeloxAV && checkout v0.1.1 3.11 && uv pip && uv pip insta torchvision==0. https://downloa u && uv pip ins imagehash==4.3. && .venv/bin/py 'not db and not -p no:cacheprov`
Unit-tier tests passed, without torch installed	`2035 tests`	of 2057 selected; result line '1 failed, 2035 passed, 21 skipped, 1475 deselected in 97.24s'; same single failure (test_privacy.py, unmarked Postgres use)	Reproduced 2026-11-10	command: pytho db and not gpu a no:cacheprovide test-unit tier) commit 9f2fc6f (9900X, 30 GB RA (Ubuntu), Python `git clone https://github. AI/LabeloxAV && checkout v0.1.1 3.11 && uv pip && .venv/bin/py 'not db and not -p no:cacheprov`
Unit-tier tests skipped, without torch installed	`21 tests`	of 2057 selected; result line '1 failed, 2035 passed, 21 skipped, 1475 deselected in 97.24s'; 18 skip on missing torch/torchvision, 3 on missing sample data, depth stack or plate weights	Reproduced 2026-11-10	command: pytho db and not gpu a no:cacheprovide test-unit tier) commit 9f2fc6f (9900X, 30 GB RA (Ubuntu), Python `git clone https://github. AI/LabeloxAV && checkout v0.1.1`

Quantity	Value	Conditions	Basis	Source and repr
				3.11 && uv pip && .venv/bin/py 'not db and not -p no:cacheprov
Unit-tier tests passed (as documented)	1449 tests	make test-unit (-m 'not db and not gpu and not infra') with Postgres stopped; 'about 60 seconds'; documented 2026-08-18, older than v0.1.1 Differs from another source; both are listed.	Reported 2026-08-18	docs/TESTING.m commit 1265f41 (docs/TESTING.m 9f2fc6f) · not sta make test-unit
Import-linter contracts kept	1 contracts	'Contracts: 1 kept, 0 broken'; 770 files, 3076 dependencies analysed	Reproduced 2026-11-10	command: lint-im commit 9f2fc6f (9900X, 30 GB R/ (Ubuntu), Python git clone https://github. AI/LabeloxAV && checkout v0.1.1 3.11 && uv pip && .venv/bin/li
mypy --strict findings on the CI allowlist	0 issues	'Success: no issues found in 8 source files' (the blocking strict allowlist in .github/workflows/ci.yml)	Reproduced 2026-11-10	command: mypy missing-imports core/accel/recap core/version.py c services/api/auth services/review_ services/anonym services/training commit 9f2fc6f (9900X, 30 GB R/ (Ubuntu), Python git clone https://github. AI/LabeloxAV && checkout v0.1.1 3.11 && uv pip && .venv/bin/my ignore-missing- core/accel/ap.p core/accel/reca core/version.py services/api/au services/review services/anonym services/traini
Ruff findings on the CI blocking module set	0 findings	'All checks passed!' on the blocking set listed in .github/workflows/ci.yml	Reproduced 2026-11-10	command: ruff ch core/version.py s services/llm serv services/api/dep services/api/auth services/api/rout services/analytic commit 9f2fc6f (9900X, 30 GB R/ (Ubuntu), Python git clone https://github. AI/LabeloxAV && checkout v0.1.1 3.11 && uv pip && .venv/bin/ru core/version.py services/llm se services/api/de services/api/au

Quantity	Value	Conditions	Basis	Source and repr
				services/api/ro services/analyt
Python tests passed (full suite, full infra)	3506 tests	full Python suite run on 2026-09-12 against the live system with infrastructure up; result line: 3,506 passed, 6 skipped, 4 xfailed, 0 failed	Reported 2026-09-12	docs/VERIFICATI commit 1197b06 docs/VERIFICATI v0.1.1 9f2fc6f) · n make up && make
Python tests failed (full suite, full infra)	0 tests	full Python suite run on 2026-09-12 against the live system with infrastructure up; result line: 3,506 passed, 6 skipped, 4 xfailed, 0 failed	Reported 2026-09-12	docs/VERIFICATI commit 1197b06 docs/VERIFICATI v0.1.1 9f2fc6f) · n make up && make
Python tests skipped (full suite, full infra)	6 tests	full Python suite run on 2026-09-12 against the live system with infrastructure up; result line: 3,506 passed, 6 skipped, 4 xfailed, 0 failed	Reported 2026-09-12	docs/VERIFICATI commit 1197b06 docs/VERIFICATI v0.1.1 9f2fc6f) · n make up && make
Python tests xfailed (full suite, full infra)	4 tests	full Python suite run on 2026-09-12 against the live system with infrastructure up; result line: 3,506 passed, 6 skipped, 4 xfailed, 0 failed	Reported 2026-09-12	docs/VERIFICATI commit 1197b06 docs/VERIFICATI v0.1.1 9f2fc6f) · n make up && make
Web (frontend) tests passed	703 tests	across 62 files, verification pass of 2026-09-12	Reported 2026-09-12	docs/VERIFICATI commit 1197b06 docs/VERIFICATI v0.1.1 9f2fc6f) · n cd web && npm t
Python tests passed (recorded full-infra baseline)	2488 tests	measured 2026-08-18 against full infra (make up) on the test database; 3 skipped, 4 xfailed	Reported 2026-08-18	tests/KNOWN_F commit 2379d7c tests/KNOWN_F at v0.1.1 9f2fc6f) make up && make
mAP@50, idd-yolo11n (YOLO11n) on IDD	0.34	India Driving Dataset (public); evaluation split and IoU/confidence settings not stated in the log	Reported 2026-09-15	docs/ENGINEERI commit 01b9d11 (docs/ENGINEERI at v0.1.1 9f2fc6f) on 'a single cons not documented
mAP@50, idd-yolo11l (YOLO11l) on IDD	0.44	India Driving Dataset (public); evaluation split and IoU/confidence settings not stated in the log	Reported 2026-09-15	docs/ENGINEERI commit 01b9d11 (docs/ENGINEERI at v0.1.1 9f2fc6f) on 'a single cons not documented
Precision, idd-yolo11n (YOLO11n) on IDD	0.67	India Driving Dataset (public); evaluation split and IoU/confidence settings not stated in the log	Reported 2026-09-15	docs/ENGINEERI commit 01b9d11 (docs/ENGINEERI at v0.1.1 9f2fc6f) on 'a single cons not documented
Precision, idd-yolo11l (YOLO11l) on IDD	0.67	India Driving Dataset (public); evaluation split and IoU/confidence settings not stated in the log	Reported 2026-09-15	docs/ENGINEERI commit 01b9d11 (docs/ENGINEERI

Quantity	Value	Conditions	Basis	Source and repr
				at v0.1.1 9f2fc6f) on 'a single cons not documented
Recall, idd-yolo11n (YOLO11n) on IDD	0.3	India Driving Dataset (public); evaluation split and IoU/confidence settings not stated in the log	Reported 2026-09-15	docs/ENGINEERI commit 01b9d11 (docs/ENGINEERI at v0.1.1 9f2fc6f) on 'a single cons not documented
Recall, idd-yolo11l (YOLO11l) on IDD	0.39	India Driving Dataset (public); evaluation split and IoU/confidence settings not stated in the log	Reported 2026-09-15	docs/ENGINEERI commit 01b9d11 (docs/ENGINEERI at v0.1.1 9f2fc6f) on 'a single cons not documented
RANSAC ground plane depth below sensor, KITTI scan 1 of 3	1.71 m	three Velodyne HDL-64E scans from the public KITTI object set; KITTI mounts the sensor at roughly 1.73 m	Reported 2026-09-15	docs/ENGINEERI commit 01b9d11 (docs/ENGINEERI at v0.1.1 9f2fc6f) not documented
RANSAC ground plane depth below sensor, KITTI scan 2 of 3	1.68 m	three Velodyne HDL-64E scans from the public KITTI object set; KITTI mounts the sensor at roughly 1.73 m	Reported 2026-09-15	docs/ENGINEERI commit 01b9d11 (docs/ENGINEERI at v0.1.1 9f2fc6f) not documented
RANSAC ground plane depth below sensor, KITTI scan 3 of 3	1.69 m	three Velodyne HDL-64E scans from the public KITTI object set; KITTI mounts the sensor at roughly 1.73 m	Reported 2026-09-15	docs/ENGINEERI commit 01b9d11 (docs/ENGINEERI at v0.1.1 9f2fc6f) not documented
Lanelet2 export lanelet relations from one KITTI raw drive	29 relations	KITTI raw drive 0005 (public); the doc states the map geometry is not yet trustworthy (one fused boundary 8,439 m long; median paired lane 2.54 m wide)	Reported 2026-09-12	docs/VERIFICATI commit 1197b06 (docs/VERIFICATI v0.1.1 9f2fc6f) · n not documented
OpenDRIVE export driving lanes from one KITTI raw drive	58 lanes	KITTI raw drive 0005 (public); 48 roads, 1 road link; the doc states the map geometry is not yet trustworthy	Reported 2026-09-12	docs/VERIFICATI commit 1197b06 (docs/VERIFICATI v0.1.1 9f2fc6f) · n not documented
First install time from an empty image cache	6.5 minutes	stated as 'about six and a half minutes'; one-command Docker install; no GPU available	Reported 2026-09-15	README.md commit 9f2fc6f (README.md; pre clean Ubuntu 24. otherwise not sta ./scripts/insta
Docker image footprint after first install	4.5 GB	stated as 'roughly 4.5 GB' (api 2.7 GB, web 1.05 GB, database 794 MB)	Reported 2026-09-15	README.md commit 9f2fc6f (README.md; pre clean Ubuntu 24. otherwise not sta ./scripts/insta

Quantity	Value	Conditions	Basis	Source and repr
Judged precision of the pooled auto-accepted subset (strict)	0.93	machine-judged by the calibrated VLM judge, not measured against humans; as of 2026-08-27	Reported 2026-09-15	README.md commit 9f2fc6f (README.md; pre not stated not reproducibl repo: sealed go operator's flee
Champion detector AP@50:95 on sealed gold slice (prediction plane)	0.068	current champion (IDD-trained YOLO11l), 400-object human-verified gold slice; 351 of 400 gold objects have unrecoverable provenance	Reported 2026-08-20	docs/MEASUREM commit 7504034 docs/MEASUREM v0.1.1 9f2fc6f) · n not reproducibl repo: sealed go operator's flee
Champion detector AP@50 on sealed gold slice (prediction plane)	0.083	current champion (IDD-trained YOLO11l), 400-object human-verified gold slice; 351 of 400 gold objects have unrecoverable provenance Differs from another source; both are listed.	Reported 2026-08-20	docs/MEASUREM commit 7504034 docs/MEASUREM v0.1.1 9f2fc6f) · n not reproducibl repo: sealed go operator's flee
Champion detector Precision on sealed gold slice (prediction plane)	0.164	current champion (IDD-trained YOLO11l), 400-object human-verified gold slice; at confidence 0.25; 351 of 400 gold objects have unrecoverable provenance	Reported 2026-08-20	docs/MEASUREM commit 7504034 docs/MEASUREM v0.1.1 9f2fc6f) · n not reproducibl repo: sealed go operator's flee
Champion detector Recall on sealed gold slice (prediction plane)	0.146	current champion (IDD-trained YOLO11l), 400-object human-verified gold slice; at confidence 0.25; 351 of 400 gold objects have unrecoverable provenance Differs from another source; both are listed.	Reported 2026-08-20	docs/MEASUREM commit 7504034 docs/MEASUREM v0.1.1 9f2fc6f) · n not reproducibl repo: sealed go operator's flee
Champion detector mAP@50 on sealed gold set (registry gold metrics)	0.4407	champion mr-idd-yolo11l-local on a sealed gold set of 202 frames / 399 instances; the doc notes the registry's gold_metrics column holds two different yardsticks; whether this is the same gold slice as the 400-object prediction-plane figure is not stated Differs from another source; both are listed.	Reported 2026-09-11	docs/AUTONOM commit 2c37cdb docs/AUTONOM 9f2fc6f) · not sta not reproducibl repo: sealed go operator's flee
Champion detector recall on sealed gold set (registry gold metrics)	0.3931	champion mr-idd-yolo11l-local on a sealed gold set of 202 frames / 399 instances Differs from another source; both are listed.	Reported 2026-09-11	docs/AUTONOM commit 2c37cdb docs/AUTONOM 9f2fc6f) · not sta not reproducibl repo: sealed go operator's flee
Human-verified objects in the corpus	1577 objects	of 578k objects (0.27%), as of 2026-08-27	Reported 2026-09-15	README.md commit 9f2fc6f (README.md; pre not stated not reproducibl repo: sealed go operator's flee
Judged label precision, class	0.6522	before the 2026-08-27 revert of the track-gap interpolation pass; 23 decided crops of an 80-crop random sample (57 judged unsure); interval [0.4489, 0.8119];	Reported 2026-08-	reports/class_pr 26.json

Quantity	Value	Conditions	Basis	Source and repr
'autorickshaw' (raw, leaf-class exact)		superclass-level precision 1.0; judge-error-corrected estimate 1.0 (clamped: true); machine-judged, not human-verified	26	commit b340b18 reports/class_pr 26.json; present stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'autorickshaw' (raw, leaf-class exact)	0.7	after the 2026-08-27 revert of the track-gap interpolation pass; 20 decided crops of an 80-crop random sample (60 judged unsure); interval [0.481, 0.8545]; superclass-level precision 1.0; judge-error-corrected estimate 1.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-27	reports/class_pr 27.json commit a5e3703 reports/class_pr 27.json; present a stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'bus' (raw, leaf-class exact)	0.5273	before the 2026-08-27 revert of the track-gap interpolation pass; 55 decided crops of an 80-crop random sample (25 judged unsure); interval [0.3979, 0.6531]; superclass-level precision 0.8364; judge-error-corrected estimate 1.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-26	reports/class_pr 26.json commit b340b18 reports/class_pr 26.json; present stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'bus' (raw, leaf-class exact)	0.5741	after the 2026-08-27 revert of the track-gap interpolation pass; 54 decided crops of an 80-crop random sample (26 judged unsure); interval [0.4416, 0.6967]; superclass-level precision 0.8519; judge-error-corrected estimate 1.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-27	reports/class_pr 27.json commit a5e3703 reports/class_pr 27.json; present a stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'hoarding' (raw, leaf-class exact)	0.4186	before the 2026-08-27 revert of the track-gap interpolation pass; 43 decided crops of an 80-crop random sample (37 judged unsure); interval [0.2838, 0.5667]; superclass-level precision 0.907; judge-error-corrected estimate 0.7774 (clamped: true); machine-judged, not human-verified	Reported 2026-08-26	reports/class_pr 26.json commit b340b18 reports/class_pr 26.json; present stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla

Quantity	Value	Conditions	Basis	Source and repr
Judged label precision, class 'hoarding' (raw, leaf-class exact)	0.4255	after the 2026-08-27 revert of the track-gap interpolation pass; 47 decided crops of an 80-crop random sample (34 judged unsure); interval [0.2951, 0.5672]; superclass-level precision 0.9362; judge-error-corrected estimate 0.7943 (clamped: true); machine-judged, not human-verified	Reported 2026-08-27	reports/class_pr 27.json commit a5e3703 reports/class_pr 27.json; present a stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'motorcycle' (raw, leaf-class exact)	0.8065	before the 2026-08-27 revert of the track-gap interpolation pass; 31 decided crops of an 80-crop random sample (49 judged unsure); interval [0.6372, 0.9081]; superclass-level precision 0.9032; judge-error-corrected estimate 1.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-26	reports/class_pr 26.json commit b340b18 reports/class_pr 26.json; present stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'motorcycle' (raw, leaf-class exact)	0.871	after the 2026-08-27 revert of the track-gap interpolation pass; 31 decided crops of an 80-crop random sample (49 judged unsure); interval [0.7115, 0.9487]; superclass-level precision 0.9677; judge-error-corrected estimate 1.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-27	reports/class_pr 27.json commit a5e3703 reports/class_pr 27.json; present a stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'mpv' (raw, leaf-class exact)	0.7941	before the 2026-08-27 revert of the track-gap interpolation pass; 68 decided crops of an 80-crop random sample (12 judged unsure); interval [0.6836, 0.8732]; superclass-level precision 1.0; judge-error-corrected estimate 1.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-26	reports/class_pr 26.json commit b340b18 reports/class_pr 26.json; present stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'object_fallback' (raw, leaf-class exact)	0.3714	before the 2026-08-27 revert of the track-gap interpolation pass; 35 decided crops of an 80-crop random sample (45 judged unsure); interval [0.2317, 0.5366]; superclass-level precision 1.0; judge-error-corrected estimate 0.6613 (clamped: false); machine-judged, not human-verified	Reported 2026-08-26	reports/class_pr 26.json commit b340b18 reports/class_pr 26.json; present stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla

Quantity	Value	Conditions	Basis	Source and repr
				python -m scripts.run_cla
Judged label precision, class 'object_fallback' (raw, leaf-class exact)	0	after the 2026-08-27 revert of the track-gap interpolation pass; 76 decided crops of an 80-crop random sample (4 judged unsure); interval [0.0, 0.0481]; superclass-level precision 0.1711; judge-error-corrected estimate 0.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-27	reports/class_pr 27.json commit a5e3703 reports/class_pr 27.json; present a stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'pedestrian' (raw, leaf-class exact)	0.8462	before the 2026-08-27 revert of the track-gap interpolation pass; 39 decided crops of an 80-crop random sample (41 judged unsure); interval [0.7027, 0.9275]; superclass-level precision 0.9744; judge-error-corrected estimate 1.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-26	reports/class_pr 26.json commit b340b18 reports/class_pr 26.json; present stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'pedestrian' (raw, leaf-class exact)	0.8667	after the 2026-08-27 revert of the track-gap interpolation pass; 45 decided crops of an 80-crop random sample (35 judged unsure); interval [0.7382, 0.9374]; superclass-level precision 1.0; judge-error-corrected estimate 1.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-27	reports/class_pr 27.json commit a5e3703 reports/class_pr 27.json; present a stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'pole' (raw, leaf-class exact)	0.5556	before the 2026-08-27 revert of the track-gap interpolation pass; 36 decided crops of an 80-crop random sample (44 judged unsure); interval [0.3958, 0.7046]; superclass-level precision 0.8056; judge-error-corrected estimate 1.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-26	reports/class_pr 26.json commit b340b18 reports/class_pr 26.json; present stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'pole' (raw, leaf-class exact)	0.6071	after the 2026-08-27 revert of the track-gap interpolation pass; 56 decided crops of an 80-crop random sample (24 judged unsure); interval [0.4763, 0.7242]; superclass-level precision 1.0; judge-error-corrected estimate 1.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-27	reports/class_pr 27.json commit a5e3703 reports/class_pr 27.json; present a stated not reproducibl repo: requires Postgres corpus judge on GPU; c

Quantity	Value	Conditions	Basis	Source and repr
				scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'rider' (raw, leaf-class exact)	0.587	before the 2026-08-27 revert of the track-gap interpolation pass; 46 decided crops of an 80-crop random sample (34 judged unsure); interval [0.4434, 0.7171]; superclass-level precision 0.913; judge-error-corrected estimate 1.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-26	reports/class_pr 26.json commit b340b18 reports/class_pr 26.json; present stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'rider' (raw, leaf-class exact)	0.6444	after the 2026-08-27 revert of the track-gap interpolation pass; 45 decided crops of an 80-crop random sample (35 judged unsure); interval [0.4984, 0.7678]; superclass-level precision 0.9778; judge-error-corrected estimate 1.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-27	reports/class_pr 27.json commit a5e3703 reports/class_pr 27.json; present stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'sedan' (raw, leaf-class exact)	0.2195	before the 2026-08-27 revert of the track-gap interpolation pass; 41 decided crops of an 80-crop random sample (39 judged unsure); interval [0.12, 0.3671]; superclass-level precision 0.878; judge-error-corrected estimate 0.2876 (clamped: true); machine-judged, not human-verified	Reported 2026-08-26	reports/class_pr 26.json commit b340b18 reports/class_pr 26.json; present stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'sedan' (raw, leaf-class exact)	0.2368	after the 2026-08-27 revert of the track-gap interpolation pass; 38 decided crops of an 80-crop random sample (42 judged unsure); interval [0.1299, 0.3921]; superclass-level precision 0.9737; judge-error-corrected estimate 0.3301 (clamped: false); machine-judged, not human-verified	Reported 2026-08-27	reports/class_pr 27.json commit a5e3703 reports/class_pr 27.json; present stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'street_light' (raw, leaf-class exact)	0.0357	before the 2026-08-27 revert of the track-gap interpolation pass; 28 decided crops of an 80-crop random sample (52 judged unsure); interval [0.0063, 0.1771]; superclass-level precision 0.6429; judge-error-corrected estimate 0.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-26	reports/class_pr 26.json commit b340b18 reports/class_pr 26.json; present stated not reproducibl repo: requires Postgres corpus

Quantity	Value	Conditions	Basis	Source and repr
				judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'suv' (raw, leaf-class exact)	0.7419	before the 2026-08-27 revert of the track-gap interpolation pass; 62 decided crops of an 80-crop random sample (18 judged unsure); interval [0.6212, 0.8345]; superclass-level precision 0.9839; judge-error-corrected estimate 1.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-26	reports/class_pr 26.json commit b340b18 reports/class_pr 26.json; present stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'suv' (raw, leaf-class exact)	0.7258	after the 2026-08-27 revert of the track-gap interpolation pass; 62 decided crops of an 80-crop random sample (18 judged unsure); interval [0.6041, 0.8212]; superclass-level precision 0.9839; judge-error-corrected estimate 1.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-27	reports/class_pr 27.json commit a5e3703 reports/class_pr 27.json; present a stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'traffic_sign' (raw, leaf-class exact)	0.6939	before the 2026-08-27 revert of the track-gap interpolation pass; 49 decided crops of an 80-crop random sample (31 judged unsure); interval [0.5547, 0.8048]; superclass-level precision 0.9184; judge-error-corrected estimate 1.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-26	reports/class_pr 26.json commit b340b18 reports/class_pr 26.json; present stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'traffic_sign' (raw, leaf-class exact)	0.84	after the 2026-08-27 revert of the track-gap interpolation pass; 50 decided crops of an 80-crop random sample (30 judged unsure); interval [0.7149, 0.9166]; superclass-level precision 1.0; judge-error-corrected estimate 1.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-27	reports/class_pr 27.json commit a5e3703 reports/class_pr 27.json; present a stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'traffic_signal' (raw, leaf-class exact)	0.0175	before the 2026-08-27 revert of the track-gap interpolation pass; 57 decided crops of an 80-crop random sample (23 judged unsure); interval [0.0031, 0.0929]; superclass-level precision 0.9649; judge-error-corrected estimate 0.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-26	reports/class_pr 26.json commit b340b18 reports/class_pr 26.json; present stated not reproducibl repo: requires

Quantity	Value	Conditions	Basis	Source and repr
				Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'traffic_signal' (raw, leaf-class exact)	0.0526	after the 2026-08-27 revert of the track-gap interpolation pass; 57 decided crops of an 80-crop random sample (23 judged unsure); interval [0.0181, 0.1437]; superclass-level precision 1.0; judge-error-corrected estimate 0.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-27	reports/class_pr 27.json commit a5e3703 reports/class_pr 27.json; present a stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'truck' (raw, leaf-class exact)	0.5111	before the 2026-08-27 revert of the track-gap interpolation pass; 45 decided crops of an 80-crop random sample (35 judged unsure); interval [0.37, 0.6504]; superclass-level precision 0.7333; judge-error-corrected estimate 1.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-26	reports/class_pr 26.json commit b340b18 reports/class_pr 26.json; present stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Judged label precision, class 'truck' (raw, leaf-class exact)	0.5745	after the 2026-08-27 revert of the track-gap interpolation pass; 47 decided crops of an 80-crop random sample (33 judged unsure); interval [0.4328, 0.7049]; superclass-level precision 0.7872; judge-error-corrected estimate 1.0 (clamped: true); machine-judged, not human-verified	Reported 2026-08-27	reports/class_pr 27.json commit a5e3703 reports/class_pr 27.json; present a stated not reproducibl repo: requires Postgres corpus judge on GPU; c scripts/run_cla python -m scripts.run_cla
Detector macro recall, real-v1-nano (YOLO11n)	0.282	trained on gate-approved detections only; 10 classes; validation split at the session boundary; recall against an incomplete key (precision and mAP stated as not measurable on this corpus); candidate, not promoted	Reported 2026-08-02	models/README commit 02e6e0e models/README 9f2fc6f) · not sta not reproducibl repo: validatio operator's flee published)
Detector micro recall, real-v1-nano (YOLO11n)	0.411	trained on gate-approved detections only; 10 classes; validation split at the session boundary; recall against an incomplete key (precision and mAP stated as not measurable on this corpus); candidate, not promoted; confidence threshold 0.25	Reported 2026-08-02	models/README commit 02e6e0e models/README 9f2fc6f) · not sta not reproducibl repo: validatio operator's flee published)

Quantity	Value	Conditions	Basis	Source and repr
Detector macro recall, real-v1-small (YOLO11s)	0.281	trained on gate-approved detections only; 10 classes; validation split at the session boundary; recall against an incomplete key (precision and mAP stated as not measurable on this corpus); candidate, not promoted	Reported 2026-08-02	models/README commit 02e6e0e models/README 9f2fc6f) · not sta not reproducibl repo: validatio operator's flee published)
Detector micro recall, real-v1-small (YOLO11s)	0.386	trained on gate-approved detections only; 10 classes; validation split at the session boundary; recall against an incomplete key (precision and mAP stated as not measurable on this corpus); candidate, not promoted; confidence threshold 0.25	Reported 2026-08-02	models/README commit 02e6e0e models/README 9f2fc6f) · not sta not reproducibl repo: validatio operator's flee published)
Detector macro recall, real-v2-nano (YOLO11n)	0.623	trained on every detection a human did not refuse; 10 classes; validation split at the session boundary; recall against an incomplete key (precision and mAP stated as not measurable on this corpus); candidate, not promoted	Reported 2026-08-02	models/README commit 02e6e0e models/README 9f2fc6f) · not sta not reproducibl repo: validatio operator's flee published)
Detector micro recall, real-v2-nano (YOLO11n)	0.77	trained on every detection a human did not refuse; 10 classes; validation split at the session boundary; recall against an incomplete key (precision and mAP stated as not measurable on this corpus); candidate, not promoted; confidence threshold 0.25	Reported 2026-08-02	models/README commit 02e6e0e models/README 9f2fc6f) · not sta not reproducibl repo: validatio operator's flee published)
Median mask IoU between SAM 2 masks and SAM 1 re-prompt (seg_verify)	0.893 IoU	about a fifth fall below 0.8, concentrated on riders, motorcycles and autorickshaws; about 195 ms per masked object	Reported 2026-09-15	README.md commit 9f2fc6f (README.md; pre not stated not reproducibl repo: sealed go operator's flee

Table 6. Every measured quantity in this report, with its source, commit, hardware and the command that reproduces it. “Reproduced” rows were run by the author; “Reported” rows are quoted from the repository at the commit shown.

Cite this report

APA

Roy, S. J. (2026). A confidence-gated annotation and data engine for driving data: LabeloxAV v0.1.1 and its public evidence (Technical Report IF-TR-2026-04, Version 1.1). IsoFrames. https://isoframes.io/reports/if-tr-2026-04/

BibTeX

```
@techreport{isoframes_if_tr_2026_04,
  author      = {Roy, Sherin Joseph},
  title       = {A confidence-gated annotation and data engine for driving data: LabeloxAV v0.1.1 and its public
evidence},
  institution = {IsoFrames, BluRabbit Mobility},
  number      = {IF-TR-2026-04},
  type        = {Technical Report},
  year        = {2026},
  month       = nov,
```

```
  note         = {Version 1.1},
  url          = {https://isoframes.io/reports/if-tr-2026-04/}
}
```