Creating Impactful Autonomous Driving Datasets: A Strategic Guide from Research Gap to Benchmark

Richard Schwarzkopf[†,*], Jonas Merkert, Frank Bieder[†],
Annika Bätz, Alexander Blumberg, Carlos Fernandez, Felix Hauser[†], Fabian Immel[†],
Christian Kinzig, Hendrik Königshof[†], Fabian Konstantinidis, Martin Lauer, Willi Poh,
Nils Rack, Kevin Rösch[†], Yinzhe Shen, Marlon Steiner, Gleb Stepanov, Dominik Strutz,
Ömer Şahin Taş, Julian Truetsch[†], Kaiwen Wang, Royden Wagner,
Jan-Hendrik Pauls, Christoph Stiller[†]

Abstract:

Well-designed autonomous driving datasets have fundamentally shaped research progress, yet existing literature primarily describes *what* datasets contain rather than *how* to strategically design impactful ones. This is especially limiting for small and medium-sized labs and startups that cannot afford to misallocate scarce resources. We argue that impactful dataset creation begins with a diagnosis: whether a research question is blocked by a *data problem* or an *evaluation problem*, and proceeds by selecting the *minimal data operator(s)* that closes the resulting gap, recording new data only when no cheaper operator(s) suffices. We analyze the evolution of major autonomous driving (AD) datasets through this lens and distill a strategic framework spanning gap identification, operator choice, sensor suite design, and annotation strategy. We ground the framework in a running case study of our KITScenes dataset family. The datasets are available at: https://kitscenes.com/.

Keywords: Autonomous Driving, Dataset Design, Benchmarks, Research Gap Identification

1 Introduction

In autonomous driving (AD) research, public datasets play a crucial role for developing and evaluating new methods, enabling fair comparison of methods. However, existing literature often only describes *what* a dataset contains, providing an ex-post-facto account of the sensor modalities, the geographic diversity, or the provided annotations. Far fewer works address *how* to strategically design an impactful dataset or *why* certain decisions were made, offering little guidance on systematically identifying high-impact research gaps, navigating cost trade-offs, and determining the best approach to advance the field in both the short and long term.

We argue that impactful dataset creation should be driven by three questions asked *before* any data is collected. First, a diagnosis: What is the main research gap the team is most interested in, and where are they confident they can close this gap significantly?

[*]all authors are associated with Karlsruhe Institute of Technology (KIT), all dataset authors are ordered alphabetically, [†]also associated with the FZI Research Center for Information Technology, Karlsruhe. E-Mail: (firstname.lastname@kit.edu).

Second, is the targeted research question blocked because a method *cannot be developed* for lack of training data (a **data problem**), or because its performance *cannot be conclusively measured and compared* (an **evaluation problem**)? Third, a choice of means: among the available *data operators*, from re-labeling and combining existing datasets, through synthesis (simulation or generation), to recording new real-world data, which is the *minimal* one that closes the diagnosed gap?

To make this concrete, we present a running case study of our **KITScenes** dataset family, which deliberately spans the two extremes of the cost spectrum: *KITScenes Multimodal* [1], a high-fidelity recorded dataset with production-grade HD maps, where recording was the justified operator; and *KITScenes LongTail* [2], a lean long-tail evaluation benchmark that pairs low-cost recording with high-value human reasoning traces. Notably, KITScenes Multimodal had to defend itself against the option to *extend* Argoverse 2 [3] with new map labels, a far cheaper operator, and we use that abandoned decision to illustrate when recording does, and does not, pay off. Our contributions are:

- An analysis of major AD datasets through the lens of the research gap each addressed and the *data operator* used to create it (Section 2).
- A strategic framework that guides a researcher from diagnosing a data vs. evaluation problem, through choosing the minimal data operator, to sensor, annotation, and scaling decisions (Section 3).
- A running case study of the KITScenes family that grounds each step in concrete decisions, including the record-vs-reuse trade-off, and distills the lessons we learned.

2 The Evolution of AD Datasets

The role of datasets in robotics and autonomous driving has shifted fundamentally over the past decade. Early benchmarks were conceived primarily as instruments to rigorously and reproducibly *evaluate* systems and methods. With the rise of powerful machine learning, the emphasis drifted from data as an evaluation basis toward data as the engine of data-driven development. This shift kept growing in scale and application diversity, and has recently produced a divergence between the data used for *training* and the data used for *evaluation*: internet-scale pretrained models are adapted via in-context learning or fine-tuned on large-scale data such as the NVIDIA Physical AI dataset [4] or nuPlan [5], while being evaluated on curated, long-tail benchmarks such as the Waymo Open Dataset for End-to-End Driving [6]. This split mirrors our central distinction between the *data problem* (training) and the *evaluation problem* (benchmarking), developed in Section 3. This section surveys how AD datasets evolved along this trajectory: Table 1 summarizes the key datasets and, crucially, the *data operator* each used, while Figure 1 illustrates the resulting separation into large-scale training data and high-fidelity evaluation data. We organize this evolution into clusters by era, data operator (the way the data was produced, defined in Section 3.3), and annotation regime (Table 1), describing each by the research it enabled rather than cataloguing datasets. The **foundation** cluster (2012-2017) established that progress needs standardized, reproducible evaluation: KITTI [7] introduced recorded, human-verified multi-sensor data with held-out splits, enabling core detection,

tracking, flow, depth, and odometry research. In parallel, a **synthetic** cluster used simulation for cheap pixel-perfect labels and controllable domains, from photo-realistic clones [8] to engine-based closed-loop testing. The **multimodal robotaxi** cluster (2018-2022) then prioritized scale and sensor fusion: nuScenes [9] first delivered a full 360° suite with HD maps, and Argoverse 2 [3] and the Waymo Open Dataset [10, 11] scaled multimodal detection, tracking, forecasting, and online/changed-map research. A distinct **crowdsourced collection** cluster instead mined commodity-device imagery (dashcams, phones) [12, 13], trading fidelity for geographic and appearance diversity to drive robustness and domain generalization. More recently, **derived and unified** benchmarks [14, 15, 16] reused existing data to define new tasks and reduce fragmentation without new recording. Finally, the field split into two regimes (Figure 1): massive auto-labeled and generative data for planning and world models [5, 4, 17], versus small, high-fidelity sets for conclusive long-tail evaluation [6, 2] critical for Level 4 autonomy.

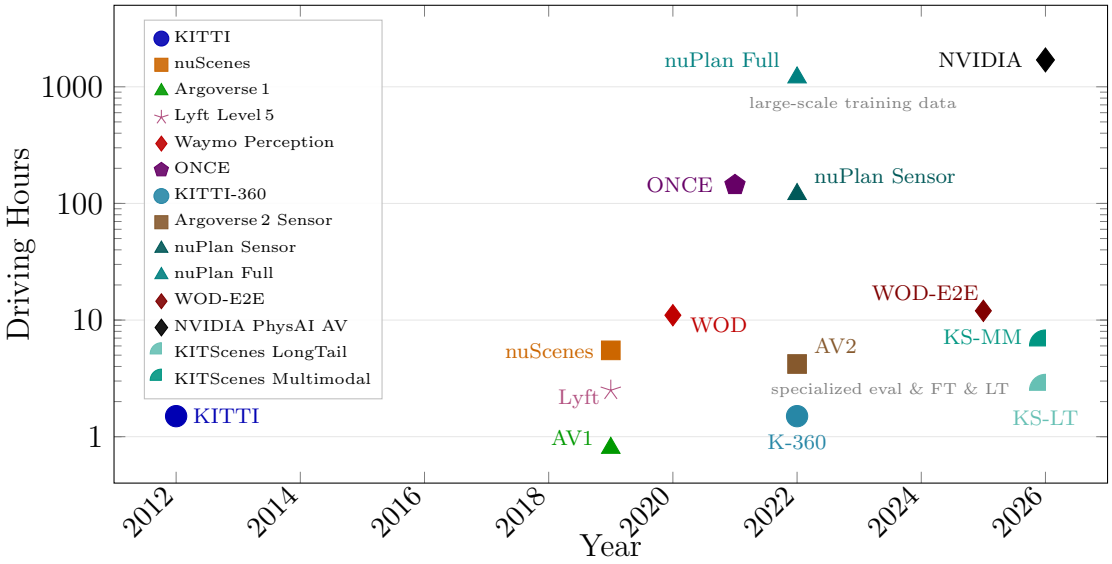

Figure 1: Publicly available autonomous driving dataset with sensor data and labels (no KITTI raw data recording, NVIDIA labels not public yet but promised, WOD with current size). A scale separation trend can be observed for datasets designed for scale with autolabeling and datasets designed for high fidelity and specialized evaluation, fine-tuning, or long-tail distribution sampling.

Emerging needs: Despite the wealth of available data, this evolution has revealed requirements that current benchmarks only partially address. First, the training/evaluation divergence raises the bar for curated evaluation sets: as models are increasingly trained on large-scale auto-labeled data, conclusive benchmarking demands the highest sensor, calibration, synchronization, and annotation fidelity and data sampled deliberately for long-tail coverage. Second, a persistent gap exists between research data and the real-world software stacks in which AV methods are ultimately deployed, both in data formats and annotation richness. A concrete instance is the lack of HD maps in standardized, open-source formats such as Lanelet2 [18], with deployment-ready completeness including explicit logical associations between 3D traffic lights and the lanes they govern, which are essential for planning and standard in production maps, yet largely absent from public research datasets. These emerging needs are the design target of our case study (Section 3).

The KITScenes family is built to close them on underrepresented European roads, and we use it throughout the remainder of this paper to ground the strategic framework.

Table 1: Evolution of autonomous driving datasets, grouped into *clusters* by era, the *data operator(s)* used and their dominant *annotation regime*. Each row names only the most seminal dataset of the cluster; further examples are cited compactly. Operators (**R**, **C**, **U**, **E**, **S**) are defined in Table 2 (Section 3.3). Annotation: **H** = human-in-the-loop, **A** = auto-labeling.

Cluster	**Era**	**Op.**	**Ann.**	**Seminal (+ refs)**	**Gap → Enabled**
Foundation perception benchmarks	2012–17	**R**	**H**	KITTI [7, 19, 20, 21, 22]	First standardized real-world benchmarks → detection, tracking, flow, depth, odometry, segmentation.
Synthetic & simulation	2016–	**S**	**A**	Virtual KITTI [8, 23, 24]	Cheap pixel-perfect labels → controllable domains, closed-loop training/eval.
Multimodal robotaxi fleets	2018–22	**R**	**H**	nuScenes [9, 10, 11, 3, 25, 26, 27, 28, 29, 30]	360° suites + HD maps at scale → 3D detection/tracking, forecasting, online/changed maps.
Crowdsourced collection	2017–	**C**	**H+A**	BDD100K [12, 13, 31]	Commodity-device diversity at low cost → robustness, domain generalization.
Derived & unified benchmarks	2023–	**U/E**	**H**	OpenLane-V2 [14, 15, 16]	New tasks/formats from existing data → topology, map verification, cross-dataset training.
Large-scale planning & E2E / generative	2024–	**R/S**	**A**	nuPlan [5, 4, 17]	Massive auto-labeled & generative data → learning-based planning, world models.
Long-tail & closed-loop evaluation	2024–	**R/E/S**	**H**	Waymo E2E [6, 32, 33]	Conclusive eval lagged training scale → long-tail, closed-loop E2E evaluation.
KITScenes Multimodal	2026	**R**	**H**	KITScenes MM	High-fidelity eval → stack-ready Lanelet2 HD maps with traffic-light-to-lane links.
KITScenes LongTail	2026	**R**	**H**	KITScenes LT	Long-tail 360° eval → E2E logs with reasoning traces, ranking score.

3 Strategic Framework for Dataset and Evaluation Design

We now turn the observations of Section 2 into an actionable process. The framework is a sequence of decisions, each of which narrows the design space and constrains the next: (i) *diagnose* whether the research question is blocked by a data or an evaluation problem;

(ii) *identify* the precise gap to close; (iii) *choose the minimal data operator* that closes it; and only then (iv) commit to the concrete sensor, annotation, and scaling decisions that the chosen operator entails. The guiding principle throughout is impact-per-effort: each step is framed to spend scarce resources where they create the most leverage.

We illustrate every step with the running KITScenes case study, set apart in the green callout boxes so that the general guidance remains distinguishable from examples that are specific to our dataset family.

3.1 Diagnose: Data Problem vs. Evaluation Problem

First, a researcher should be precise about *why* progress is currently blocked. We find it clarifying to separate two failure modes. A **data problem** exists when a capable method *cannot be developed or trained* because the data does not exist at sufficient scale or diversity; the bottleneck is on the supply side. An **evaluation problem** exists when methods can be built, but their performance *cannot be conclusively measured, compared, or trusted*, e.g. because no benchmark captures the relevant conditions, the metrics are uninformative, or the test data is too easy, too small, or contaminated by leakage. The two are not mutually exclusive, but naming the dominant one is decisive, because it routes every subsequent choice: a data problem typically calls for scale or diversity and is often best served by cheaper operators (Section 3.3), whereas an evaluation problem calls for fidelity, careful curation, and metric design, and may justify the most expensive operator and annotation strategy.

This distinction also explains the field-wide divergence noted in Section 2: large auto-labeled corpora answer the data problem of training, while smaller, high-fidelity, long-tail sets answer the evaluation problem of trustworthy benchmarking. A common and costly mistake is to attack an evaluation problem with a scale-oriented data effort, or vice versa.

KITScenes Case Study

KITScenes datasets target both *data* (data from the EU with deployment-grade HD maps) and an *evaluation* problem. Modern perception and end-to-end models can already be *trained* on abundant data, yet they cannot be *conclusively evaluated* for application-relevant capabilities: long-range perception, complete and stack-ready HD maps, and competent behavior in rare events. KITScenes Multimodal addresses the fidelity side (the highest sensor and map fidelity for conclusive benchmarking on irregular road layouts in dense European urban environments), while KITScenes LongTail addresses the coverage side (deliberate long-tail sampling).

3.2 Identifying the Research Gap and Target Benchmarks

High-impact datasets of the past identified their gap through close study of state-of-the-art methods and their lack of rigorous, large-scale evaluation.

We advocate approaching a dataset project with a specific target benchmark and task in mind from the outset, since this choice dictates the required sensors, annotations, and metrics. A candidate gap can be located along several complementary dimensions: the *saturation* of an existing task, the definition of an *entirely new task*, and the *application alignment* of an established task. We diagnosed in KITScenes Multimodal [1] a concrete

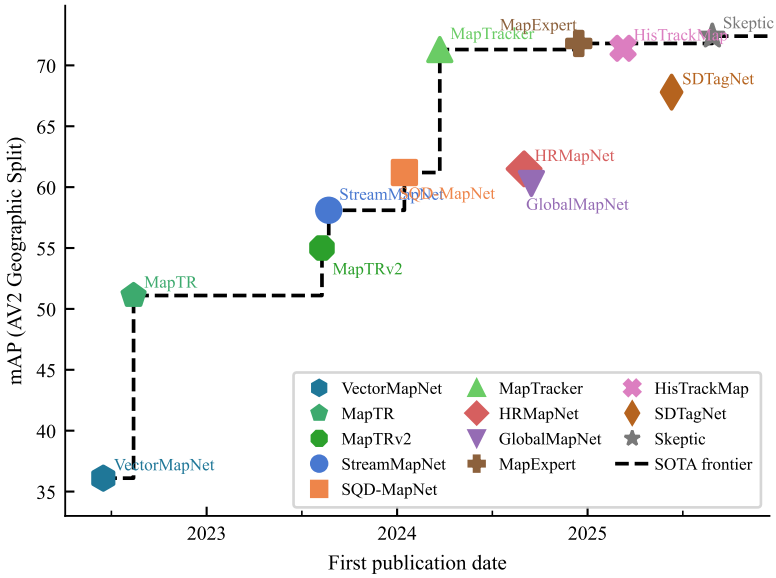

Figure 2: Historical SOTA progression of representative online HD map construction methods [34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44] on AV2 [3]. Progress saturates after MapTracker [38], signalling a gap of evaluation conclusiveness rather than of incremental accuracy. Figure referenced by [1].

gap in online HD map construction. On existing benchmarks the task has visibly saturated as depicted in Figure 2, while model outputs remain far from complete, missing most of the relational 3D map elements that deployment actually requires. This is a gap of conclusiveness and application alignment rather than of further incremental accuracy.

KITScenes Case Study

KITScenes addresses two initial research gaps, one per dataset. *KITScenes Multimodal* targets: no curated, high-fidelity dataset that pairs surround-view, long-range sensing with open-format, stack-ready HD maps that carry the logical associations (e.g. traffic-light-to-lane) required for planning. Closing it enables a more complete online HD map construction that recovers map elements *and* their relations, and its conclusiveness rests on the highest sensor, calibration, and annotation fidelity together with relational map-construction metrics that keep detection and topology separable, unlike the aggregated OpenLane-V2 score [14]. While this research gap is the initially targeted one, more research is enabled by its high-fidelity data and labels. *KITScenes LongTail* targets: decision-making in long-tail scenarios cannot be conclusively measured, because matching a single expert trajectory by displacement error ignores that multiple maneuvers are valid. It frames reasoned, instruction-grounded long-tail driving as a new task and scores it with the Multi-Maneuver Score [2] and human reasoning traces rather than displacement error alone. Together the two datasets span three design dimensions at once: *new tasks*, *application alignment*, and *evaluation conclusiveness*.

3.3 Choosing the Minimal Data Operator

Once the main research gap is defined, the central engineering decision is *how* to produce the data. We find it useful to think in terms of a small set of *data operators* (Table 2) that can be combined almost arbitrarily: **Exploitation** (adapting, re-labeling, extending, or resampling existing data), **Collection** (scraping or mining existing sources), **Unification** (unifying datasets and APIs), **Synthesis** (novel data from engine-based simulators or generative models), and **Recording** (new real-world capture). They differ in cost and output: only recording and synthesis yield genuinely new data, and only recording produces new *real-world* data, at the price of hardware, software, and labeling effort that no other operator incurs.

Our central recommendation follows directly: identify the *minimal* operator, or composition of operators, that solves the diagnosed problem, and treat recording as the operator of last resort. Many impactful datasets did exactly this without any new recording: the synthetic and derived/unified clusters of Table 1, including 123D [16] unified existing datasets behind a single API, while ArgoTweak [15] exploited AV2 map labels for a controlled map-verification benchmark; yet both progressed the field substantially. The cost/benefit of recording is highly context-dependent: it hinges on the infrastructure already available, on the potential for future reuse of the novel real-world data and how broad the enabled research could potentially be, even when focusing mainly on a specific research gap at first.

	Exploitation	**Collection**	**Unification**	**Synthesis**	**Recording**
New (sensor) data	✗	✗	✗	✓	✓
Real-world data	✓	✓	✓	✗	✓
labeling/scenario design required	✗	✓	✗	✓	✓
Modify / extend labels	✓	◐	◐	◐	◐
Curation / mining required	✗	✓	◐	✗	✓
High compute requirements	✗	✗	✗	✓	◐
Real HW & recording SW	✗	✗	✗	✗	✓
Notes	Relabel or modify labels	Scrape/mine or crowdsource	Unify datasets & APIs	Sim. envs or gen. AI; scenario/label design	Sensors & SW; full labeling

Table 2: Data Operators vs. their properties. ✓ = yes, ✗ = no, ◐ = partial.

KITScenes Case Study

Closing our research gap with recording KITScenes Multimodal was not the only option we discussed. A minimal *Exploitation* plan existed as well: extend Argoverse 2 [3] with complete Lanelet2-style 3D relational maps required for deployment-grade HD maps. We abandoned this for three reasons specific to our gap: (i) the complex European urban road layout we targeted is absent from AV2; (ii) conclusive long-range evaluation needs sensor and calibration fidelity that AV2's released data cannot provide; and (iii) reprojection-accurate map labels presuppose the very high-fidelity raw sensor data we lacked. Recording became the minimal viable operator, but was *only* feasible because substantial hardware, recording software, and mapping expertise already existed in-house.

3.4 From Operator to a New Dataset: Sensors, Annotation, Scaling and Benchmarks

The remaining decisions are largely contingent on the chosen operator. We group them as remaining topics to keep the framework compact; the considerations below apply most directly when recording, but the annotation and scaling guidance generalises to other data operators as well.

Sensor suite design. When recording, the sensor suite determines whether new tasks can be enabled or whether a target application can be tackled. While overlap to existing setups, e.g. a central top lidar, allows researchers to compare across domains, novel tasks or saturated benchmarks require changes w.r.t. established approaches. Including both classical and novel systems enables a direct evaluation of the two and supports methods that leverage their unification. One often overlooked part of the setup is its calibration and synchronization quality, which often determine whether the data can support demanding downstream tasks at all.

KITScenes Case Study

Our sensor suite combines high-resolution global-shutter cameras, multiple long-range lidars (effective range beyond 400 m), 4D imaging radar, and redundant GNSS/INS, all hardware-synchronized (Figure 3). Choosing a roof-box arrangement over integrated sensors mirrored the L4 robotaxi domain and prioritized fidelity and calibration accuracy over L2 application alignment, reflecting our evaluation-first goal.

Annotation strategy and quality. Annotations can be produced by *human-in-the-loop* labeling (annotators working from guidelines and quality control, aided by automated tools) or by *fully automatic* offline auto-labeling. Human annotation is more costly but higher quality, while at fleet scale the cost of human labeling becomes prohibitive and auto-labeling is the only viable option. While large-scale auto-labeled corpora enable previously unprecedented scaling, curated datasets with the highest calibration, synchronization, sensor, and annotation quality become *more* valuable, precisely because they provide the trustworthy yardstick against which fleet-trained models are measured.

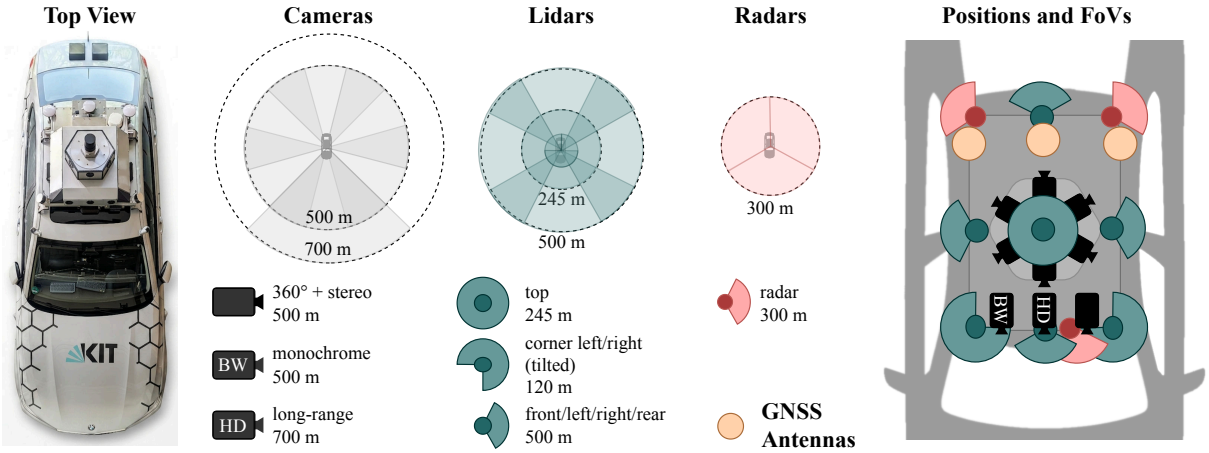

Figure 3: KITScenes Multimodal Sensor Setup. Our sensor rack (left) is depicted along with sensing range (center), as well as sensor positions and their field of view (right).

KITScenes Case Study

To achieve the necessary annotation quality required for evaluation, we follow the human-in-the-loop paradigm, but heavily utilize tooling to simplify the process. For instance, traffic signs or 3D landmarks obtain an initial auto-label [45, 46], which is then verified by humans. In contrast, HD map lane elements and their topology, including traffic-element-to-lane associations, are labeled manually with tooling support; reliable auto-labeling for maps remains an open problem. We found that quality-testing already-created HD maps for defects and mislabels is effective to automate with heuristics and classification models, leading to a more efficient manual annotation process and a higher quality of the annotations, when combined with a good visualization and UI.

Getting started and scaling out. We strongly recommend defining and producing a small but representatively diverse end-to-end slice of the data and labels as early as possible: a handful of fully processed samples forces the labeling, processing, and evaluation pipelines to exist and exposes the real goal as well as processing issues or hidden assumptions concretely.

KITScenes Case Study

Even using a small number of samples across cities, sensor setups, calibration versions, or annotation epochs allowed us to reveal a great number of hidden assumptions.

Benchmarks, baselines, metrics, and leaderboards. Three decisions govern whether evaluation is conclusive. First, splits must be free of leakage: in driving data, naive random or temporal splits let geographically adjacent observations bleed across train and test, so separation is best enforced by geographic distance if possible. Second, metrics should be well designed, conclusive, and highly correlated with the application goal. Third, a fully held-out test set served via a leaderboard needs special care in the release, since privileged information (maps, future frames, geo-referenced poses) is easily leaked. Base-

lines should be easy to reproduce and well established if possible to allow comparison to previous work.

KITScenes Case Study

KITScenes ships with five initial benchmarks: online HD map construction, long-range monocular depth, novel view synthesis, and end-to-end driving on KITScenes Multimodal, plus long-tail end-to-end driving on KITScenes LongTail. To keep evaluation conclusive, the Multimodal test set withholds all map data (the first held-out test set in online HD map construction) and separates splits by geographic distance rather than time. For application alignment we favor conclusive, map-grounded metrics, e.g. the Multi-Maneuver Score [2] for driving and relational map-construction metrics that keep detection and topology separable, unlike the aggregated OpenLane-V2 score [14]. We host a public leaderboard for the LongTail benchmark and run a Workshop Challenge to drive fast adoption.

4 Outlook and Conclusion

The future of AD research data. The training/evaluation divergence of Section 2 will likely sharpen. On the training side, fleet-scale auto-labeled and generative corpora will keep growing, making the *cheaper* data operators (exploitation, unification, and synthesis) increasingly powerful for teams that cannot record at scale. On the evaluation side, the premium will shift to small, high-fidelity, long-tail sets whose value lies in curation and metric design rather than volume. We expect the most impactful academic contributions to come not from out-recording industry, but from sharply-scoped operators applied to well-diagnosed gaps: new tasks, application-aligned metrics, and conclusive evaluation. Two needs stand out. First, integration with open-source stacks such as Autoware [47] and Apollo, which requires HD maps in standardized formats (Lanelet2 [18], OpenDRIVE [48]) and the relational annotations, notably 3D traffic lights with explicit lane associations, that rule-based and hybrid planners depend on. Second, the rise of vision-language models opens evaluation of *reasoning* about driving, where detailed HD maps can serve a dual role as geometric priors and as structured grounding for spatial language.

The future of KITScenes. Because KITScenes was recorded once at high fidelity, it is naturally a substrate for cheaper downstream operators: its recordings can be *combined* into unified resources such as 123D [16], *exploited* to extend benchmarks in the spirit of OpenLane-V2 [14], and used to *generate* novel views and scenarios for pseudo-closed-loop evaluation. This is the long-term payoff that helped justify recording in the first place: a single high-fidelity capture amortized across many future, low-cost derivatives.

Conclusion. Datasets remain foundational to autonomous driving research, but their impact is decided long before any data is collected. We argued that creators should first *diagnose* whether they face a data or an evaluation problem, then *identify* a sharply-defined gap, and only then choose the *minimal data operator* that closes it, treating recording as the operator of last resort, justified by existing infrastructure and downstream reuse.

References

[1] R. Schwarzkopf, F. Immel, A. Blumberg, J. Merkert, N. Rack, K. Wang, A. Bätz, C. Fernandez, F. Hauser, F. Konstantinidis, W. Poh, K. Rösch, Y. Shen, M. Steiner, G. Stepanov, D. Strutz, J. Truetsch, J. Villa, R. Wagner, F. Bieder, O. c. Taş, H. Caesar, J.-H. Pauls, and C. Stiller, "The road ahead in autonomous driving: The kitscenes multimodal dataset," *arXiv preprint arXiv:*, 2025. [Online]. Available: https://arxiv.org/

[2] R. Wagner, O. S. Tas, J. Villa, F. Hauser, Y. Shen, M. Steiner, D. Strutz, C. Fernandez, C. Kinzig, G. S. Guitierrez-Cabello, H. Königshof, F. Immel, R. Schwarzkopf, N. A. Rack, K. Rösch, K. Wang, J.-H. Pauls, M. Lauer, I. Gilitschenski, H. Caesar, and C. Stiller, "Longtail driving scenarios with reasoning traces: The kitscenes longtail dataset," 2026. [Online]. Available: https://arxiv.org/abs/2603.23607

[3] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, D. Ramanan, P. Carr, and J. Hays, "Argoverse 2: Next generation datasets for self-driving perception and forecasting," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021.

[4] PhysicalAI Autonomous Vehicles, "NVIDIA Autonomous Vehicle Dataset," 2025, accessed: 2026-01-30. [Online]. Available: https://huggingface.co/datasets/nvidia/PhysicalAI-Autonomous-Vehicles

[5] N. Karnchanachari, D. Geromichalos, K. S. Tan, N. Li, C. Eriksen, S. Yaghoubi, N. Mehdipour, G. Bernasconi, W. K. Fong, Y. Guo, and H. Caesar, "Towards learning-based planning: The nuplan benchmark for real-world autonomous driving," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 629–636.

[6] Waymo Open Dataset, "Vision-based End-to-End Driving Challenge 2025," 2025, accessed: 2025-11-01. [Online]. Available: https://waymo.com/open/challenges/2025/e2e-driving

[7] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

[8] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *CVPR*, 2016.

[9] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[10] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[11] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou, Z. Yang, A. Chouard, P. Sun, J. Ngiam, V. Vasudevan, A. McCauley, J. Shlens, and D. Anguelov, "Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 9710–9719.

[12] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[13] G. Neuhold, T. Ollmann, S. Rota Bulò, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5000–5009.

[14] H. Wang, T. Li, Y. Li, L. Chen, C. Sima, Z. Liu, B. Wang, P. Jia, Y. Wang, S. Jiang *et al.*, "Openlane-v2: A topology reasoning benchmark for unified 3d hd mapping," in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[15] L. Wild, R. Valencia, and P. Jensfelt, "Argotweak: Towards self-updating hd maps through structured priors," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[16] D. Dauner, V. Charraut, B. Berle, T. Li, L. Nguyen, J. Wang, C. Jing, M. Igl, H. Caesar, B. Ivanovic, A. Geiger, and K. Chitta, "123d: Unifying multi-modal autonomous driving data at scale," *arXiv preprint arXiv:2605.08084*, 2026.

[17] X. Ren, Y. Lu, T. Cao, R. Gao, S. Huang, A. Sabour, T. Shen, T. Pfaff, J. Z. Wu, R. Chen, S. W. Kim, J. Gao, L. Leal-Taixe, M. Chen, S. Fidler, and H. Ling, "Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models," 2025. [Online]. Available: https://arxiv.org/abs/2506.09042

[18] F. Poggenhans, J.-H. Pauls, J. Janosovits, S. Orf, M. Naumann, F. Kuhnt, and M. Mayr, "Lanelet2: A high-definition map framework for the future of automated driving," in *Proc. IEEE Intell. Trans. Syst. Conf.*, Hawaii, USA, November 2018. [Online]. Available: http://www.mrt.kit.edu/z/publ/download/2018/Poggenhans2018Lanelet2.pdf

[19] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[20] Y. Liao, J. Xie, and A. Geiger, "KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d," *Pattern Analysis and Machine Intelligence (PAMI)*, 2022.

[21] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[22] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, " The ApolloScape Open Dataset for Autonomous Driving and Its Application ," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 42, no. 10, pp. 2702–2719, Oct. 2020. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/TPAMI.2019.2926463

[23] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, vol. 78. PMLR, 2017, pp. 1–16.

[24] X. Jia, Z. Yang, Q. Li, Z. Zhang, and J. Yan, "Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving," in *NeurIPS 2024 Datasets and Benchmarks Track*, 2024.

[25] M.-F. Chang, J. W. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, and J. Hays, "Argoverse: 3d tracking and forecasting with rich maps," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[26] J. Houston, G. Zuidhof, L. Bergamini, Y. Ye, L. Chen, A. Jain, S. Omari, V. Iglovikov, and P. Ondruska, "One thousand and one hours: Self-driving motion prediction dataset," in *Proceedings of the 2020 Conference on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, vol. 155. PMLR, 2021, pp. 409–418.

[27] J. Mao, N. Minzhe, C. Jiang, h. liang, J. Chen, X. Liang, Y. Li, C. Ye, W. Zhang, Z. Li, J. Yu, C. XU, and H. Xu, "One million scenes for autonomous driving: Once dataset," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung, Eds., vol. 1, 2021. [Online]. Available: https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/67c6a1e7ce56d3d6fa748ab6d9af3fd7-Paper-round1.pdf

[28] M. Alibeigi, W. Ljungbergh, A. Tonderski, G. Hess, A. Lilja, C. Lindstrom, D. Motorniuk, J. Fu, J. Widahl, and C. Petersson, "Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[29] P. Xiao, Z. Shao, S. Hao, Z. Zhang, X. Chai, J. Jiao, Z. Li, J. Wu, K. Sun, K. Jiang, Y. Wang, and D. Yang, "PandaSet: Advanced sensor suite dataset for autonomous driving," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 2021, pp. 3095–3101.

[30] J. Geyer, Y. Kassahun, M. Mahmudi, X. Ricou, R. Durgesh, A. S. Chung, L. Hauswald, V. H. Pham, M. Mühlegg, S. Dorn, T. Fernandez, M. Jänicke, S. Mirashi, C. Savani, M. Sturm, O. Vorobiov, M. Oelker, S. Garreis, and P. Schuberth, "A2D2: Audi autonomous driving dataset," 2020. [Online]. Available: https://www.a2d2.audi

[31] H. Schafer, E. Santana, A. Haden, and R. Biasini, "A commute in data: The comma2k19 dataset," 2018.

[32] D. Dauner, M. Hallgarten, T. Li, X. Weng, Z. Huang, Z. Yang, H. Li, I. Gilitschenski, B. Ivanovic, M. Pavone, A. Geiger, and K. Chitta, "NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[33] W. Ljungbergh, A. Tonderski, J. Johnander, H. Caesar, K. Åström, M. Felsberg, and C. Petersson, "NeuroNCAP: Photorealistic closed-loop safety testing for autonomous driving," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2024, pp. 161–177.

[34] B. Liao, S. Chen, X. Wang, T. Cheng, Q. Zhang, W. Liu, and C. Huang, "Maptr: Structured modeling and learning for online vectorized hd map construction," in *The Eleventh International Conference on Learning Representations*, 2022.

[35] B. Liao, S. Chen, Y. Zhang, B. Jiang, Q. Zhang, W. Liu, C. Huang, and X. Wang, "Maptrv2: An end-to-end framework for online vectorized hd map construction," *International Journal of Computer Vision*, Oct 2024. [Online]. Available: https://doi.org/10.1007/s11263-024-02235-z

[36] T. Yuan, Y. Liu, Y. Wang, Y. Wang, and H. Zhao, "Streammapnet: Streaming mapping network for vectorized online hd map construction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7356–7365.

[37] S. Wang, F. Jia, W. Mao, Y. Liu, Y. Zhao, Z. Chen, T. Wang, C. Zhang, X. Zhang, and F. Zhao, "Stream query denoising for vectorized hd-map construction," in *European Conference on Computer Vision*. Springer, 2024, pp. 203–220.

[38] J. Chen, Y. Wu, J. Tan, H. Ma, and Y. Furukawa, "Maptracker: Tracking with strided memory fusion for consistent vector hd mapping," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, pp. 90–107.

[39] X. Zhang, G. Liu, Z. Liu, N. Xu, Y. Liu, and J. Zhao, "Enhancing vectorized map perception with historical rasterized maps," in *European Conference on Computer Vision*. Springer, 2024, pp. 422–439.

[40] A. Shi, Y. Cai, X. Chen, J. Pu, Z. Fu, and H. Lu, "Globalmapnet: An online framework for vectorized global hd map construction," *arXiv preprint arXiv:2409.10063*, 2024.

[41] D. Zhang, D. Chen, P. Zhi, Y. Chen, Z. Yuan, C. Li, R. Zhou, Q. Zhou *et al.*, "Mapexpert: Online hd map construction with simple and efficient sparse map element expert," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 14, 2025, pp. 14 745–14 753.

[42] J. Yang, S. Yang, X. Tan, and H. Wang, "Histrackmap: Global vectorized high-definition map construction via history map tracking," *arXiv preprint arXiv:2503.07168*, 2025.

[43] F. Immel, J.-H. Pauls, R. Fehler, F. Bieder, J. Merkert, and C. Stiller, "SDTagnet: Leveraging text-annotated navigation maps for online HD map construction," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. [Online]. Available: https://openreview.net/forum?id=N3E1cU8Cv3

[44] F. Erdoğan, M. R. Barın, and F. Güney, "Mapping like a skeptic: Probabilistic bev projection for online hd mapping," *arXiv preprint arXiv:2508.21689*, 2025.

[45] M. L. Carnot, E. Fastermann, J. Kunze, E. Peukert, A. Ludwig, and B. Franczyk, "Gtsign-220: A crowd-sourced, stvo-aligned benchmark for fine-grained german traffic sign recognition," in *Intelligent Vehicles Symposium (IV)*, 2026.

[46] J.-H. Pauls, B. Schmidt, and C. Stiller, "Automatic mapping of tailored landmark representations for automated driving and map learning," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 6725–6731.

[47] Autoware Foundation, "Autoware," https://github.com/autowarefoundation/autoware, accessed: 2026-05-02.

[48] ASAM e.V., *ASAM OpenDRIVE 1.8.0 Specification*, November 2023, published November 22, 2023. [Online]. Available: https://www.asam.net/standards/detail/opendrive/