

ArgoTweak: Towards Self-Updating HD Maps through Structured Priors

Lena Wild^{1,2} Rafael Valencia² Patric Jensfelt¹
¹KTH Royal Institute of Technology ²TRATON

{lwild, patric}@kth.se rafael.valencia.carreno@se.traton.com

Abstract

Reliable integration of prior information is crucial for self-verifying and self-updating HD maps. However, no public dataset includes the required triplet of prior maps, current maps, and sensor data. As a result, existing methods must rely on synthetic priors, which create inconsistencies and lead to a significant sim2real gap. To address this, we introduce ArgoTweak, the first dataset to complete the triplet with realistic map priors. At its core, ArgoTweak employs a bijective mapping framework, breaking down large-scale modifications into fine-grained atomic changes at the map element level, thus ensuring interpretability. This paradigm shift enables accurate change detection and integration while preserving unchanged elements with high fidelity. Experiments show that training models on ArgoTweak significantly reduces the sim2real gap compared to synthetic priors. Extensive ablations further highlight the impact of structured priors and detailed change annotations. By establishing a benchmark for explainable, prior-aided HD mapping, ArgoTweak advances scalable, self-improving mapping solutions. The dataset, baselines, map modification toolbox, and further resources are available at <https://KTH-RPL.github.io/ArgoTweak/>.

1. Introduction

High-definition (HD) maps are essential for autonomous driving, offering precise lane-level information for long-horizon predictions and occlusions handling [6]. Traditionally, these HD maps have been created through offline manual annotation – a process that, while accurate, is both labor-intensive and geographically constrained. This inherent lack of scalability has spurred a shift toward automated, end-to-end HD map generation, where bird’s-eye-view (BEV) feature backbones predict map structures directly from sensor data [13, 16, 21].

While these automated methods reduce manual effort, challenges remain. Occlusions, sensor noise, and environmental variations impact reliability, and online-generated maps often lack the semantic richness and accuracy of

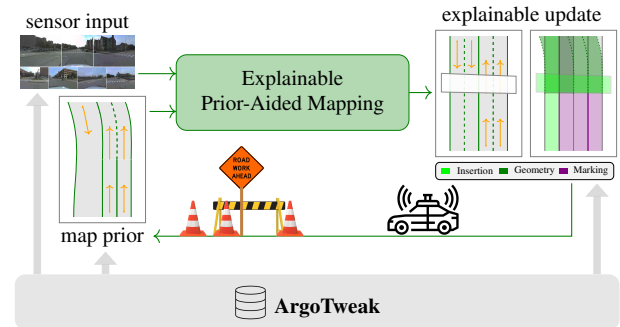


Figure 1. Overview of the ArgoTweak dataset and framework: Prior maps and current sensor input are integrated to enable self-updating HD mapping. Our annotations systematically decompose large-scale modifications into element-level changes, ensuring precise and explainable updates.

their offline counterparts [1]. To bridge this gap, recent approaches integrate priors, such as outdated or less-accurate pre-existing maps, into the generation pipeline, as these lower-quality maps are typically readily available [1, 9, 23, 24, 27, 34]. Rather than constructing maps from scratch, these models can refine inconsistencies in the priors, leading to promising improvements in map quality.

Looking ahead, the role of map priors could extend beyond improving generation quality. In vehicles, HD maps are treated as a continuous source of information, but blindly trusting them is risky. Instead, autonomous vehicle fleets need to dynamically verify and update maps against online sensor data to ensure reliability. We argue that rather than treating map generation, change detection, and map updating as separate tasks, future approaches must unify these processes: By leveraging priors, enforcing consistency, and providing structured, explainable modifications at scale, HD maps could evolve into self-improving, continuously updating road representations.

Despite these exciting perspectives, research on integrating map priors remains constrained by data limitations. No public dataset currently provides the requisite triplet: a map prior, current sensor data, and an up-to-date ground-truth map. To compensate, existing approaches generate syn-

thetic priors through scripted alterations, warping, or selective dropout of map elements [1, 9, 12, 24, 27, 34]. This raises three key challenges: (1) The diversity of synthetic priors complicates method comparison, as each embeds different amounts of ground-truth information, with unclear effects on performance metrics. (2) Current evaluation metrics fail to differentiate between performance on unchanged and newly updated map regions, making it unclear whether a model merely preserves existing structures or effectively updates the map. (3) Synthetic perturbation methods do not capture the structured, semantically correlated nature of real-world changes. As studies suggest [1, 12], this mismatch creates a significant simulation-to-reality (sim2real) gap, where models trained on synthetic priors struggle to generalize to real-world scenarios.

To overcome these limitations and advance the vision of self-improving, continuously updating road representations, we introduce ArgoTweak – the first hand-curated dataset featuring realistic map priors to align with up-to-date sensor data and ground-truth maps from the Argoverse 2 Map Change Dataset [12] (Fig. 1). Additionally, we re-annotate real-world changes from [12] to adhere to modern HD mapping standards, enabling the first evaluation of map prior integration approaches in real-world scenarios.

At the core of ArgoTweak is a bijective mapping framework that decomposes large-scale map modifications into atomic changes (*e.g.*, insertions, deletions, and lane attribute updates) providing fine-grained, explainable labels that capture the nature and extent of modifications at the element level. This allows us to introduce a comprehensive metric that assesses the updated map in changed and unchanged regions separately, exposing the shortcomings of existing metrics in capturing map adaptation capabilities.

To demonstrate the value of our dataset, we train a baseline model, achieving a significantly reduced sim2real gap compared to existing approaches. Furthermore, we leverage the explainability of both our dataset and model to identify key challenges in integrating prior maps into modern HD mapping pipelines through detailed ablation studies. In summary, our main contributions are the following:

- We introduce ArgoTweak, the first hand-curated dataset to complete the triplet of up-to-date sensor data and maps of [12] with realistic map priors and refined real-world changes, enabling standardized training and evaluation of prior integration.
- We define a novel, comprehensive benchmark for update efficacy through our explainable, atomic change annotations, distinguishing pre-existing from updated regions.
- We propose a flexible baseline architecture for explainable prior-aided mapping and show a significantly reduced sim2real gap for ArgoTweak-trained models, while highlighting key challenges in prior integration through extensive ablation experiments.

2. Related work

2.1. Online mapping with and without prior

HD map generation was initially framed as a semantic segmentation problem in the bird’s-eye view [37]. However, the need for structured outputs led to vectorized approaches like HDMapNet [13], which combined rasterized segmentation with vectorized representations. VectorMapNet [21] improved this with a two-stage pipeline modeling spatial relationships without heuristic post-processing. A major breakthrough came with MapTR [16] and MapTR-v2 [17], which reformulated map decoding as a one-stage DETR-like [3] process, enhancing speed and accuracy. While MapTR-v2 remains a strong baseline, recent works advanced along three directions: (1) refining single-frame generation with improved queries and decoding [5, 8, 20, 22, 31, 32, 36, 38], (2) enhancing temporal consistency via multi-frame fusion and global map stitching [4, 11, 26, 30, 33, 35], and (3) enriching maps with topological and semantic details [14, 15, 29].

All these methods rely solely on sensor data. However, as [24] highlights, using outdated or lower-quality maps as a prior can improve results. Building on prior works incorporating standard definition maps [10, 15, 23, 34], [24] and [1] propose hybrid queries combining existing HD map elements with learnable ones. PriorDrive [34] employs hybrid prior embedding and dual encoding for SD and lower-quality HD maps, while M3TR [9] adapts to varying map priors. In our previous work, ExelMap [27], we introduced an explainable, element-based approach for identifying and updating changed map elements.

2.2. Datasets and challenges

The availability of datasets varies across research areas. For prior-less online mapping, public datasets such as nuScenes [2], Argoverse 2 [28], and OpenLane-v2 [25] provide key resources, despite concerns over geographically overlapping splits [18, 33]. These datasets, both in their original and re-grouped forms, remain widely used (*cf.* Tab. 1).

In contrast, prior-aided HD mapping lacks public datasets that include map priors, current sensor data, and up-to-date ground-truth maps. To address this, researchers synthesize priors through discrete modifications (*e.g.*, element dropout, duplication) [1, 9, 24, 34], continuous transformations (*e.g.*, noise injection, warping) [1, 24], or rule-based approaches mimicking the semantically correlated nature of real-world changes [12, 24, 27].

However, existing synthetic prior generation techniques exhibit major limitations. Since studies selectively modify map elements in arbitrary ways, results from performance evaluations are not directly comparable. Furthermore, common evaluation metrics like mean average precision fail to

distinguish between preserving existing structures and correctly updating outdated regions, making it unclear whether models genuinely adapt to real-world changes or simply maintain known elements [9, 27]. A third issue is related to the reliance on synthetic perturbations, which introduces a sim2real gap where models trained on artificial modifications struggle to generalize to real-world updates [1].

3. Motivation

In real-world environments, changes are typically localized and relatively rare, yet their effects can be highly consequential for safety [12]. Hence, although most of the map remains stable over time, even minor undetected modifications can lead to critical failures. A robust mapping approach must balance two competing objectives: accurately detecting and integrating changes, and preserving unchanged elements with high fidelity.

We argue that achieving this balance requires more than just the triplet of standardized priors, sensor data, and up-to-date ground truth – it necessitates a paradigm shift encompassing data representation, training methodologies, and evaluation protocols. To reliably integrate priors into mapping, we hence introduce the **ArgoTweak dataset**, the first to provide explicitly annotated priors using a novel **bijection change mapping framework** (Sec. 4). On the model side, we propose an **explainable prior-aided mapping network** (Sec. 5) that makes changes interpretable. To ensure rigorous evaluation, we introduce a **fine-grained metric** that quantifies both stability in unchanged regions and responsiveness to updates (Sec. 6). Finally, we integrate our dataset, metric, and model (Sec. 7) to validate our claims and establish a **new benchmark for explainable prior-aided HD mapping**.

4. The ArgoTweak Dataset

4.1. Bijective change mapping

As our first contribution, we introduce **bijection change mapping**, a novel framework to systematically provide change annotations for each map element. While not a strict mathematical bijection, it is a design principle that guides how we relate high-level *structural updates* – complex modifications to road geometry or semantics – to fine-grained *atomic changes* applied at the element level (i.e., individual lane segments or pedestrian crossings in the widely used Argoverse 2 HD map format [28]).

Our motivation is twofold. First, we want any local road change, no matter how complex, to be decomposable into a unique and traceable set of edits on individual elements, to facilitate map learning. Second, to reduce annotation ambiguity, road layout changes must be annotated in a consistent manner, avoiding arbitrary distinctions between full element replacement and incremental modification. Enforc-

ing this consistency is key for interpretability and model supervision.

Bijectivity requirements. We formally define atomic changes as a_i with set $\mathbf{A} = \{a_1, a_2, \dots, a_n\}$ and structural updates as y_i with set $\mathbf{Y} = \{y_1, y_2, \dots, y_n\}$, where \mathbf{Y} comprises all large-scale modifications that can affect an HD map. Finally, x_i is an example set of specific atomic changes that lead to the structural update y_i , and $\mathbf{X} = P(\mathbf{A})$ the set of all such permutationally invariant subsets of \mathbf{A} . The bijective design principle satisfies two conditions:

- **Surjectivity:** Every structural map update should be explainable in terms of a combination of atomic changes. This ensures coverage.

$$\forall y_i \in \mathbf{Y} \exists x_j \in \mathbf{X} : f(x_j) = y_i. \quad (1)$$

- **Injectivity:** The same structural update should not be represented in two fundamentally different ways. This ensures uniqueness.

$$\forall x_i, x_j \in \mathbf{X}, \forall y_i, y_j \in \mathbf{Y} : y_j = y_i \Rightarrow x_j = x_i. \quad (2)$$

Atomic changes and macro-modifications. To operationalize this framework, we first define the atomic changes $\mathbf{A} = \{a_1, \dots, a_n\}$ as the set of low-level element-wise edit operations on map elements in the Argoverse 2 format [28]:

- **geometry:** modification of element boundaries,
- **markings:** changes to lane markings (type, color),
- **type:** changes to lane semantics (e.g., bus-only),
- **connectivity:** changes to predecessors/successors,
- **insertion / deletion:** creation / removal of map elements.

Next, we constrain structural changes \mathbf{Y} to a closed set of interpretable *macro-modifications* $\hat{\mathbf{Y}}$. This design choice is essential: in principle, real-world road layouts can change in unbounded and combinatorially complex ways. Without a limited vocabulary of update types, it would be impossible to maintain a consistent mapping between atomic edits and structural updates. To address this problem, we ask: *how does the local road change, functionally and structurally?* We find that most changes can be meaningfully abstracted into five categories:

- **shape**, e.g., the local road has been widened,
- **appearance**, e.g., the local road has updated markings,
- **function**, e.g., a local road is now reserved for buses,
- **lane graph**, e.g., a new merge has been added,
- **lane number**, e.g., the total lane count in the local road increased/decreased.

By defining this interpretable macro space $\hat{\mathbf{Y}}$, we enable a **soft bijection**: a practical, near-injective mapping from macro-modifications to unique compositions of atomic changes \mathbf{A} . While this is an approximation of the full bijectivity between \mathbf{A} and \mathbf{Y} , our experiments later demonstrate its consistency and practical usefulness.

Dataset	Split	Scenes	Avg. duration	HD map		Change annotations		
				prior	gt	global	frame	element
nuScenes [2]	train/val/test	700/150/150	20s	✗	✓	n.a.	n.a.	n.a.
Argoverse 2 Sensor [28]	train/val/test	750/150/100	15s	✗	✓	n.a.	n.a.	n.a.
Argoverse 2 TbV [12]	train	799	54s	✗	✓	n.a.	n.a.	n.a.
	val	111		✓	✗	✓	~	✗
	test	133		✓	✗	✗	✗	✗
ArgoTweak (ours)	train/val/test	697/102/111	56s	✓	✓	✓	✓	✓

Table 1. Comparison of public datasets used in HD mapping, highlighting the availability of priors for training and testing. The last columns detail the presence of change annotations at scenario, *i.e.*, global level, frame-level or element-level. ArgoTweak is the first dataset to complement ground-truth HD maps and sensor data with realistic and real-world map priors, and element-wise annotations.

Constructing the soft bijection. When constructing the mapping between our atomic change vocabulary \mathbf{A} and constrained macro space $\hat{\mathbf{Y}}$, a central challenge is that many macro-modifications can be represented in multiple ways. For example, a shape change could be expressed via geometry edits to existing elements, or modeled as deleting old segments and inserting new ones. This ambiguity undermines injectivity: the same structural update could correspond to multiple disjoint atomic edit sets. Moreover, connectivity changes create ambiguities – if map elements are treated as lane graph vertices, adding a new predecessor, for instance, does not necessarily change the role of the edited lane segment itself.

To resolve this, we introduce a disambiguation rule based on the **underlying road graph**. We represent map elements as *edges* in a non-directional lane graph, where junctions form vertices. Insertions and deletions are used *only* when a change also alters this topology – e.g., adding a new connection or splitting a lane. If the topology remains unchanged, we express the update as an in-place edit using atomic geometry, marking, or type changes.

This principle guides the design of our mapping matrix \mathcal{C} , which encodes which atomic change types contribute to each macro-modification:

$$\mathcal{C} = \begin{array}{c|ccccc} & \text{geo} & \text{mark} & \text{type} & \text{ins} & \text{del} \\ \hline \text{shape} & \checkmark & \times & \times & \checkmark^* & \checkmark^* \\ \text{appearance} & \times & \checkmark & \times & \checkmark^* & \checkmark^* \\ \text{function} & \times & \times & \checkmark & \checkmark^* & \checkmark^* \\ \text{lane graph} & \times & \times & \times & \checkmark & \checkmark \\ \text{lane number} & 0 & 0 & 0 & +1 & -1 \end{array} \quad (3)$$

Here, $\mathcal{C}_{ij} = \checkmark$ indicates that the macro-modification \hat{y}_i is produced by the atomic change a_j . For lane number, $\mathcal{C}_{ij} = \pm 1$ signals an increase or decrease in the total lane number. The starred entries indicate that insert/delete operations are only used when the lane graph changes too. Details regarding the practical application of the framework and illustrative figures can be found in [Appendix A](#) of the supplementary material.

4.2. Building the dataset

Equipped with our bijective change mapping, we now select the dataset to which we apply our framework. As summarized in Tab. 1, existing public datasets for *HD mapping* pair sensor data with ground-truth HD maps but lack outdated priors preceding real-world changes. Such priors are present in [12], a dataset for *HD map change detection*. While [12] includes some real-world outdated priors and current sensor data, it does not provide ground truth maps for changed regions, making it unsuitable for developing and evaluating map updating methods.

Another limitation is that real-world stale maps are rare. As a result, real-world priors are only available for validation and testing in [12] (*cf.* Tab. 1), whereas for training, the authors propose a rule-based synthetic map modification approach. However, they report a substantial sim2real gap when models trained on these synthetic priors are evaluated on real-world changes. This challenge is expected to persist when generating updated maps [1].

To address these limitations, we construct a high-quality training set by introducing realistic structural modifications to ground-truth maps within our bijective change mapping framework. This approach ensures full control over the nature and distribution of changes, eliminating the need to mine rare real-world priors while maintaining high realism. For testing, we leverage the real-world priors and up-to-date sensor data from the validation split of [12], and annotate the complementing ground truth maps. We use the former validation as our test set, as global and frame-wise annotations are not publicly available for the original test set, making change localization challenging.

Finally, to systematically analyze a potential sim2real gap, we reorganize the original training split. Maps from Washington, D.C., are designated exclusively for validation to prevent geographical data leakage [18, 33]. This setup allows us to measure the sim2real gap by comparing performance on realistic but manually created validation maps against the real-world test set.

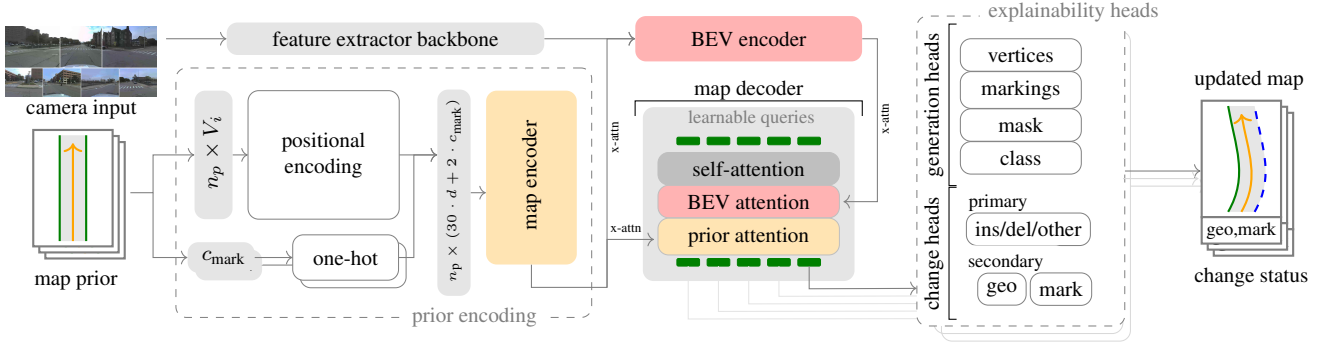


Figure 2. Network architecture overview. A BEV encoder and a prior encoder extract features from camera input and map prior, while the map decoder predicts both updated map elements and their change status (e.g., insertion, deletion, geometry/marking edits) through explainable multi-head training. See Sec. 5.1 for implementation details.

4.3. Additional dataset refinements

We refine the base dataset [12] with several improvements. Following [15], we apply an OpenLane-V2-inspired [25] merging process to prior and updated maps, resolving unnecessary breakpoints in lane segments. The search boundaries for element merging include lane graph changes, changes to the lane properties and the tracked change status in terms of atomic changes. Additionally, we unify pedestrian crossing edges, previously defined in both clockwise and counterclockwise orientations. While our priors and ground truth maps are in 2D, z-coordinates can be sampled from the base dataset’s ground height annotations as needed. Detailed dataset statistics can be found in Appendix B of the supplementary material.

5. Explainable prior-aided mapping

Task definition. At timestamp t , given current sensor data and prior map M_{prior} , the goal is to estimate whether M_{prior} is in agreement with current sensor data, detect changed elements and update them accordingly to produce the ground truth map M_{gt} . This includes detecting changes in lane markings and types, insertions, deletions, and geometric modifications at an element level.

5.1. Network architecture

To serve as a baseline for future research and a vehicle for our evaluation, we propose a flexible map updating scheme that can operate at different levels of explainability: without explicit change assessment (*i.e.*, not explainable), with a binary change detection head (*i.e.* changed/unchanged), or with a full explainability module that attributes specific atomic changes to individual map elements. We adopt LaneSegNet [15] as our backbone due to its lane-segment-based formulation, which aligns well with our bijective mapping framework. We extend the backbone by adding a map prior encoder and explainability heads (Fig. 2).

Prior encoding. We use the map prior encoding scheme proposed in [23], which we previously demonstrated to be compatible with LaneSegNet [15] in our earlier work [27]. Our prior includes 10 2D-points for left and right boundary and centerline, class labels (pedestrian crossing or lane), and semantic attributes (left/right lane line markings). We extract the geometric representation V for each of the n_{prior} elements in the prior map,

$$V = \{V_{left}, V_{right}, V_{center}\} = \{(x_i, y_i)\}_{i=1}^{30}.$$

Lane line markings are encoded using a one-hot scheme for all 7 marking types c_{mark} . Lane marking color information is not used in the present configuration. The same holds for lane type, as none of the current map generation networks detect bus or bike lane separately. The encoded coordinates and boundary types are concatenated into a polyline sequence of shape $n_{prior} \times (30 \cdot d + 2 \cdot c_{mark})$, where d is the positional embedding dimension.

Once encoded, the prior can be incorporated into the map generation pipeline in two ways. The first approach, used in, *e.g.*, [23, 27], treats the prior as an additional modality in cross-attention alongside BEV-features. However, bandwidth limitations potentially hinder full integration [1]. The second approach replaces fixed hierarchical queries with prior tokens, refining them into a posterior map through decoder layers [1, 9, 24]. While strategy 2 can improve stability, we found strategy 1 to be more flexible and effective for integrating larger changes.

Explainability heads. In our previous work [27], we proposed the idea of using multi-task learning with change assessment heads. Since that approach was limited to insertions and deletions, we redesign the heads to handle a more complex setting with diverse atomic changes.

Our key insight is that some change categories are mutually exclusive, while others can co-occur. Inserted or deleted elements cannot meaningfully undergo further

changes, whereas geometry and lane marking modifications can happen simultaneously on a single element. Thus, we introduce a two-stage assessment: a primary multi-class classification head determines the element’s status from mutually exclusive labels {No Change, Insertion, Deletion, Other}. If classified as "Other," secondary binary heads can be considered for geometric and lane marking modifications. This modular design allows swift integration of additional change categories in the future.

With these three heads, our final loss is

$$\begin{aligned} \mathcal{L} = & \lambda_{\text{vec}}\mathcal{L}_{\text{vec}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{type}}\mathcal{L}_{\text{type}} \\ & + \lambda_{\text{cd,prim.}}\mathcal{L}_{\text{cd,prim.}} + \sum_{i \in [\text{geo,mark}]} \lambda_{\text{cd,sec.}}^i \mathcal{L}_{\text{cd,sec.}}^i, \end{aligned} \quad (4)$$

where $\mathcal{L}_{\text{seg}} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}}$ is the segmentation loss from [15], and the loss weights are defined as $\lambda_{\text{vec}} = 0.025$, $\lambda_{\text{seg}} = 3.0$, $\lambda_{\text{ce}} = 1.0$, $\lambda_{\text{dice}} = 1.0$, $\lambda_{\text{cls}} = 1.5$, $\lambda_{\text{type}} = 0.01$, $\lambda_{\text{cd,primary}} = \lambda_{\text{cd,secondary}}^i = 0.5$. We use cross-entropy loss for the primary and Focal Loss [19] for the secondary heads.

6. Metric

Complementing our dataset and explainable model, we introduce a fine-grained evaluation protocol as the third pillar to systematically assess both stability in unchanged regions and responsiveness to updates. Inspired by [9] and [27], we propose a change aware dual-metric framework that comprises a coarse detection accuracy (mACC) and fine-grained map generation average precision (mAPC).

Fine-grained mAPC. Given the set of change categories \mathcal{C} of size $|\mathcal{C}| = C$, we evaluate each predicted map element \hat{V} against its ground-truth counterpart V *only if* their predicted and true change status match, $\hat{c}_V = c_V$. The change-aware lane segment distance is defined by adapting [15] to

$$\begin{aligned} D_{ls}^c(V, \hat{V}) = & \frac{1}{2} \left[\text{Chamfer}([V_{\text{left}}, V_{\text{right}}], [\hat{V}_{\text{left}}, \hat{V}_{\text{right}}]) \right. \\ & \left. + \text{Fréchet}(V_{\text{center}}, \hat{V}_{\text{center}}) \right] \otimes \mathbf{1}\{\hat{c}_V = c_V\}. \end{aligned} \quad (5)$$

We compute the average precision per class (AP_c) at distance thresholds {1.0, 2.0, 3.0}m. For non-directional pedestrian crossings, we use Chamfer distance at {0.5, 1.0, 1.5}m. The class-wise mAP_c combines lane segments and pedestrian crossings,

$$\text{mAP}_c = \frac{1}{2}(\text{AP}_c^{\text{ls}} + \text{AP}_c^{\text{pc}}). \quad (6)$$

Finally, the overall class-aware mean average precision is

$$\text{mAPC} = \frac{1}{C} \sum_{c \in \mathcal{C}} \text{mAP}_c. \quad (7)$$

This ensures equal weighting across object and change types, preventing dominance by over-represented classes.

Coarse mACC. Extending [12] and our prior work [27], we introduce a coarse change detection metric. For each frame, a binary ground-truth label $y_c \in \{0, 1\}$ indicates whether at least one map element changed for change type $c \in \mathcal{C}$. The model’s prediction $\hat{y}_c \in \{0, 1\}$ follows the same criterion. Class-wise precision and recall are captured via

$$\text{Acc}_c^{+(-)} = \frac{\sum_{i=1}^N \mathbf{1}\{\hat{y}_c = y_c\} \cdot \mathbf{1}\{y_c = 1(0)\}}{\sum_{i=1}^N \mathbf{1}\{y_c = 1(0)\}}. \quad (8)$$

The final accuracy metric is:

$$\text{mACC} = \frac{1}{C} \sum_{c \in \mathcal{C}} \text{mAcc}_c, \quad (9)$$

with

$$\text{mAcc}_c = \frac{1}{2}(\text{Acc}_c^+ + \text{Acc}_c^-). \quad (10)$$

Evaluating both mACC and mAPC offers several advantages. The coarse mACC offers an initial assessment, with low scores signaling a risk of missing updates, while fine-grained mAPC evaluates element accuracy by change type. High mACC but low mAPC suggests detecting changes without precise localization, whereas poor mACC may indicate an overly conservative model. This multi-tiered evaluation framework facilitates model refinement and establishes a foundation for nuanced comparison.

7. Experiments

We train our network for 10 epochs with a batch size of 8 and AdamW optimizer on 8 NVIDIA A10G Tensor Core GPUs. For feature extraction, we employ a pretrained ResNet-50 [7]. We use camera as the only sensor input and crop our map size to $50 \times 50\text{m}^2$ [27]. While not speed-optimized, our model runs at ~ 4 FPS on a single NVIDIA A10G GPU.

7.1. Map updating without change modelling

In this experiment, we examine how different priors influence model performance to highlight the inadequacy of mAP for evaluating prior-aided mapping. We train our network on ArgoTweak, and on synthetically generated priors inspired by established prior-generation methods [1, 9, 12, 24, 27, 34]. We consider three types of priors:

- **Continuous modifications:** We add Gaussian noise ($\mu = 0$, $\sigma = 0.5$) to all vertices of the ground truth map.
- **Discrete modifications:** We randomly delete or shift entire map elements, with a Gaussian drift ($\mu = 0$, $\sigma = 0.5$) and probabilities $p_{\text{del}} = p_{\text{shift}} = 0.2$.
- **Rule-based modifications:** Using the approach in [12], rule-based, scripted edits such as pedestrian crossing insertions and deletions, bike lane additions, and lane marking changes are generated (*cf.* Appendix C).

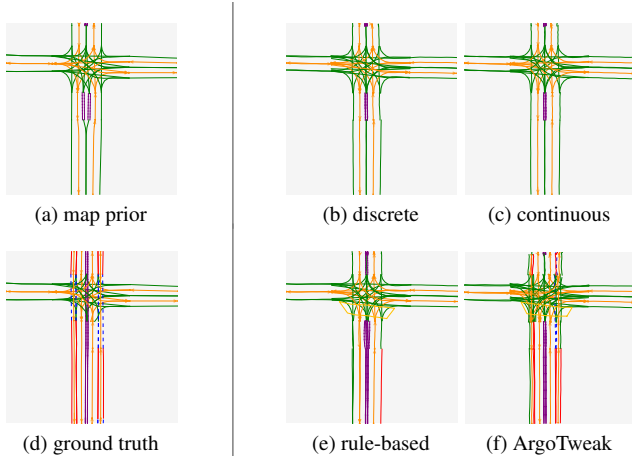


Figure 3. Qualitative comparison of model outputs with different priors. Only the ArgoTweak-trained network captures complex road updates, while models trained on synthetic priors produce limited or overfit edits. However, these differences are not reflected in mAP scores (Tab. 2), illustrating a limitation of standard metrics.

Map prior	AP _{ls}	AP _{pc}	mAP
no prior (baseline)	32.9	45.9	39.4
continuous modifications	71.6	75.5	73.5
discrete modifications	71.0	71.9	71.5
rule-based editing [12]	<u>74.2</u>	<u>79.3</u>	<u>76.7</u>
ArgoTweak	75.8	79.6	77.7

Table 2. Quantitative comparison of models trained on different priors. Despite strong qualitative variation across outputs (Fig. 3), this is not reflected in the similar mAP scores, highlighting the need for change-aware evaluation.

To ensure fair comparison, we use neither change annotations nor change assessment heads, because synthetic priors – especially those generated through noise perturbations – cannot meaningfully be described by atomic changes. While the mAP for our ArgoTweak-trained model is only slightly higher than when trained on synthetic priors (a 1% gain, Tab. 2), this result is by design and central to our argument. The purpose of this experiment is not to showcase large mAP improvements, but to demonstrate that mAP fails to reflect meaningful differences in model behavior. Despite similar mAP values across all priors, manual inspection reveals stark qualitative differences (Fig. 3). Specifically, the ArgoTweak-trained model captures complex map updates; models trained on rule-based priors tend to overfit to lane marking changes; and noise-based priors only support minor geometric corrections. Notably, models trained on synthetic priors show greater stability in unchanged regions.

These observations underscore a critical limitation:

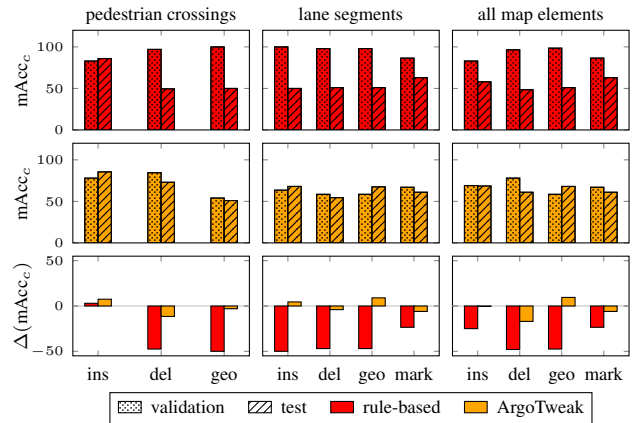


Figure 4. When training on rule-based priors (red), the sim2real gap (third row) is substantially larger than when trained on ArgoTweak (yellow).

Without a notion of change, existing metrics like mAP collapse very different behaviors into nearly identical scores. This masks key tradeoffs in stability vs. adaptability, and offers no actionable guidance for improving model responsiveness to updates. The present findings reinforce that beyond manual inspection and with only mAP as a metric we are effectively blind to meaningful model differences.

7.2. Sim2real gap assessment

While the first experiment demonstrated why change-aware modeling is necessary in the first place, we now show that even change-aware models fail to generalize when trained on scripted priors – highlighting the need for ArgoTweak’s realism to close the sim2real gap. In this experiment, we train two versions of our network with primary and secondary change assessment heads. Given that noise-based synthetic priors do not align with meaningful change categories (see Sec. 7.1), we conduct a comparative analysis between training on ArgoTweak and on rule-based priors [12]. Since both datasets describe realistic changes, we apply atomic change reasoning.

To identify sim2real gaps, we evaluate both networks on the real-world test split, as well as on the ArgoTweak validation set and the rule-based validation set, respectively. We compute the absolute difference in mAcc_c for $c \in \{\text{insertion, deletion, geometry change, mark change}\}$. The results, presented in Fig. 4, indicate that the model trained on rule-based priors exhibits significantly larger sim2real gaps across all metrics. For the combined metric of mACC_c, we observe $\Delta\text{mACC} = -36.0$ for the model trained on a rule-based prior, whereas our ArgoTweak dataset reduces this gap by more than a factor of 10 to $\Delta\text{mACC} = -3.5$. Additional results computed on our fine metric mAPC are shown in Appendix D.

Changes	mAP _c		mAP				mAPC	mAcc _c					mACC	
	$\neg c$	c	ins	del	geo	mark		c	ins	del	geo	mark		
(1) none	–	–	–	–	–	–	77.7	–	–	–	–	–	–	–
(2) $c/\neg c^\dagger$	79.0	14.7	–	–	–	–	77.5	46.9	66.5	–	–	–	–	66.5
(3) $c/\neg c$	79.0	10.4	–	–	–	–	77.7	44.7	70.5	–	–	–	–	70.5
(4) full [†]	80.5	19.7	9.4	15.7	–	5.1 [†]	79.4	10.1 [†]	71.0	67.0	60.5	–	65.5 [†]	64.3
(5) full [‡]	77.4	16.0	9.0	16.8	–	5.1 [‡]	78.3	10.3 [‡]	71.5	67.0	60.5	–	65.5 [‡]	64.3
(6) full	78.5	14.0	7.6	16.3	4.6	6.3	78.8	8.7	71.5	68.5	61.0	68.0	61.0	64.6

Table 3. Performance comparison for ArgoTweak-trained models with different levels of annotation detail: no change annotation (*none*), binary change annotation ($c/\neg c$) and atomic change annotation (*full*). [†] models trained on ArgoTweak without annotation of geometry changes. [‡] primary change detection head only. Notably, mAPC and mACC are computed over different sets of \mathcal{C} , with $\mathcal{C} = \{c, \neg c\}$ for models 2 and 3, and $\mathcal{C} = \{\text{ins}, \text{del}, \text{geo}^{(\dagger)}, \text{mark}\}$ for models 4-6.

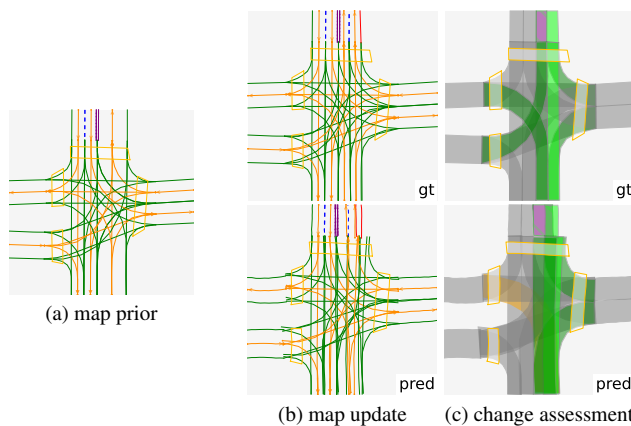


Figure 5. Example of our ArgoTweak-trained model. For change assessment, purple denotes marking changes, light green insertions, dark green geometry edits. Yellow elements were classified as "Other", but no secondary head was triggered.

7.3. Ablation studies

We investigate how varying levels of annotation detail affect performance. Notably, models trained under different annotation granularities cannot be directly compared: As detailed in Sec. 4.3, search boundaries for element merging depend on the change status, meaning annotation granularity alters lane segment partitions.

For models trained without change annotations (Tab. 3, 1), the only available performance indicator is mAP (cf. Sec. 7.1). This value is higher for networks trained with full change-aware approaches (Tab. 3, 4-6). Within the constraints imposed by different merging conditions, this suggests that change annotation not only facilitates detailed evaluation but also aids the network during training.

Next, we compare two networks trained with binary change labels and the primary change assessment head to distinguish between changed and unchanged elements (Tab. 3, 2 and 3). Here, model 2 excludes geometric

changes from the annotations. This exclusion is motivated by the observation that current map generation methods often lack the geometric precision necessary to reliably distinguish subtle road shape modifications from noise. Configuration 2 achieves comparable mAP and mAP _{$\neg c$} but outperforms model 3 in terms of mAPC and mAP_c. This suggests that omitting geometric changes enables the network to better predict the shape of changed elements. Interestingly, the model's mean accuracy (mACC) decreases, indicating that overall change detection is negatively impacted when deviations between prior and predicted maps are not annotated.

Finally, we evaluate models trained with fine-grained atomic change annotations (Tab. 3, 4-6 and Fig. 5). In this experiment, we remove the secondary change assessment head (model 5), treating marking and geometric changes as a single category. While differences in map element merging prevent direct numerical comparisons, we do not observe significant performance degradation across key metrics. This suggests that our bijective mapping framework effectively captures the diversity of real-world map changes, ensuring that the annotation strategy does not introduce confusion, even with the full set of change categories.

8. Conclusion

We introduced ArgoTweak, the first dataset to pair realistic map priors with current sensor data and up-to-date ground-truth maps. With our bijective mapping framework, we structured map updates into an explainable process, enabling fine-grained annotations that distinguish real-world changes from stable elements. Through extensive experiments, we demonstrated that models trained with ArgoTweak significantly reduce the sim2real gap, while our fine-grained metrics, mAPC and mACC, unlocked deep insights into the balance between map stability and adaptability. By setting a new benchmark for self-updating HD maps, ArgoTweak advances scalable, explainable, and continuously improving mapping solutions.

Acknowledgements

The research work was funded by the Swedish Foundation for Strategic Research (SSF) under the project DeltaMap (ID22-0045). This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. We thank Mohammad Nazari for his feedback on this work and the technical support.

References

- [1] Samuel M. Bateman, Ning Xu, H. Charles Zhao, Yael Ben Shalom, Vince Gong, Greg Long, and Will Maddern. Exploring Real World Map Change Generalization of Prior-Informed HD Map Prediction Models. In *CVPR Workshops*, 2024. 1, 2, 3, 4, 5, 6
- [2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. NuScenes: A multi-modal dataset for autonomous driving. In *CVPR*, 2020. 2, 4
- [3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *ECCV*, 2020. 2
- [4] Jiacheng Chen, Yuefan Wu, Jiaqi Tan, Hang Ma, and Yasutaka Furukawa. MapTracker: Tracking with Strided Memory Fusion for Consistent Vector HD Mapping. In *ECCV*, 2024. 2
- [5] Sehwan Choi, Jungho Kim, Hongjae Shin, and Jungwook Choi. Mask2Map: Vectorized HD Map Construction Using Bird’s Eye View Segmentation Masks. In *ECCV*, 2024. 2
- [6] Gamal Elghazaly, Raphaël Frank, Scott Harvey, and Stefan Saffko. High-Definition Maps: Comprehensive Survey, Challenges, and Future Perspectives. *ITS*, 4:527–550, 2023. 1
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 6
- [8] Haotian Hu, Fanyi Wang, Yaonong Wang, Laifeng Hu, Jingwei Xu, and Zhiwang Zhang. ADMap: Anti-disturbance framework for reconstructing online vectorized HD map. In *ECCV*, 2024. 2
- [9] Fabian Immel, Richard Fehler, Frank Bieder, Jan-Hendrik Pauls, and Christoph Stiller. M3TR: Generalist HD Map Construction with Variable Map Priors. *arXiv preprint 2411.10316*, 2024. 1, 2, 3, 5, 6
- [10] Zhou Jiang, Zhenxin Zhu, Pengfei Li, Huan ang Gao, Tianyuan Yuan, Yongliang Shi, Hang Zhao, and Hao Zhao. P-MapNet: Far-seeing Map Generator Enhanced by both SDMap and HDMap Priors. *arXiv preprint 2403.10521*, 2024. 2
- [11] Nayeon Kim, Hongje Seong, Daehyun Ji, and Sujin Jang. Unveiling the Hidden: Online Vectorized HD Map Construction with Clip-Level Token Interaction and Propagation. In *NeurIPS*, 2024. 2
- [12] John W. Lambert and James Hays. Trust, but Verify: Cross-modality fusion for hd map change detection. In *NeurIPS Datasets and Benchmarks*, 2021. 2, 3, 4, 5, 6, 7, 15
- [13] Qi Li, Yue Wang, Yilun Wang, and Hang Zhao. HDMapNet: An Online HD Map Construction and Evaluation Framework. In *ICRA*, 2022. 1, 2
- [14] Tianyu Li, Li Chen, Huijie Wang, Yang Li, Jiazhi Yang, Xiangwei Geng, Shengyin Jiang, Yuting Wang, Hang Xu, Chunjing Xu, Junchi Yan, Ping Luo, and Hongyang Li. Graph-based Topology Reasoning for Driving Scenes. *arXiv preprint 2304.05277*, 2023. 2
- [15] Tianyu Li, Peijin Jia, Bangjun Wang, Li Chen, Kun Jiang, Junchi Yan, and Hongyang Li. LaneSegNet: Map Learning with Lane Segment Perception for Autonomous Driving. In *ICLR*, 2024. 2, 5, 6, 11
- [16] Bencheng Liao, Shaoyu Chen, Xinggong Wang, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Chang Huang. MapTR: Structured Modeling and Learning for Online Vectorized HD Map Construction. In *ICRL*, 2023. 1, 2
- [17] Bencheng Liao, Shaoyu Chen, Yunchi Zhang, Bo Jiang, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggong Wang. MapTRv2: An End-to-End Framework for Online Vectorized HD Map Construction. *IJCV*, 133, 2025. 2
- [18] Adam Lilja, Junsheng Fu, Erik Stenborg, and Lars Hammarstrand. Localization Is All You Evaluate: Data Leakage in Online Mapping Datasets and How to Fix It. In *CVPR*, 2024. 2, 4
- [19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. *PAMI*, 42(2):318–327, 2020. 6
- [20] Xiaolu Liu, Song Wang, Wentong Li, Ruizi Yang, Junbo Chen, and Jianke Zhu. MGMap: Mask-Guided Learning for Online Vectorized HD Map Construction. In *CVPR*, 2024. 2
- [21] Yicheng Liu, Yuan Yuantian, Yue Wang, Yilun Wang, and Hang Zhao. VectorMapNet: End-to-end Vectorized HD Map Learning. In *ICML*, 2023. 1, 2
- [22] Zihao Liu, Xiaoyu Zhang, Guangwei Liu, Ji Zhao, and Ningyi Xu. Leveraging Enhanced Queries of Point Sets for Vectorized Map Construction. In *ECCV*, 2024. 2
- [23] Katie Z Luo, Xinshuo Weng, Yan Wang, Shuang Wu, Jie Li, Kilian Q Weinberger, Yue Wang, and Marco Pavone. Augmenting Lane Perception and Topology Understanding with Standard Definition Navigation Maps. In *ICRA*, 2024. 1, 2, 5
- [24] Rémy Sun, Li Yang, Diane Lingrand, and Frédéric Precioso. Mind the map! Accounting for existing map information when estimating online HDMaps from sensor data. In *WACV*, 2025. 1, 2, 5, 6
- [25] Huijie Wang, Tianyu Li, Yang Li, Li Chen, Chonghao Sima, Zhenbo Liu, Bangjun Wang, Peijin Jia, Yuting Wang, Shengyin Jiang, Feng Wen, Hang Xu, Ping Luo, Junchi Yan, Wei Zhang, and Hongyang Li. Openlane-v2: A topology reasoning benchmark for unified 3d hd mapping. In *NeurIPS*, 2023. 2, 5
- [26] Shuo Wang, Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Zehui Chen, Tiancai Wang, Chi Zhang, Xiangyu Zhang, and Feng Zhao. Stream Query Denoising for Vectorized HD-Map Construction. In *ECCV*, 2024. 2
- [27] Lena Wild, Ludvig Ericson, Rafael Valencia, and Patric Jensfelt. ExelMap: Explainable Element-based HD-Map Change Detection and Update, 2024. 1, 2, 3, 5, 6

- [28] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next Generation Datasets for Self-Driving Perception and Forecasting. In *NeurIPS Datasets and Benchmarks*, 2021. [2](#), [3](#), [4](#), [11](#)
- [29] Dongming Wu, Jiahao Chang, Fan Jia, Yingfei Liu, Tiancai Wang, and Jianbing Shen. TopoMLP: An Simple yet Strong Pipeline for Driving Topology Reasoning. *ICLR*, 2024. [2](#)
- [30] Xuan Xiong, Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, and Hang Zhao. Neural Map Prior for Autonomous Driving. In *CVPR*, 2023. [2](#)
- [31] Zhenhua Xu, Kwan-Yee. K. Wong, and Hengshuang Zhao. InsMapper: Exploring Inner-instance Information for Vectorized HD Mapping. In *ECCV*, 2024. [2](#)
- [32] Jing Yang, Minyue Jiang, Sen Yang, Xiao Tan, Yingying Li, Errui Ding, Jingdong Wang, and Hanli Wang. MGMapNet: Multi-Granularity Representation Learning for End-to-End Vectorized HD Map Construction. In *ICLR*, 2025. [2](#)
- [33] Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, and Hang Zhao. StreamMapNet: Streaming Mapping Network for Vectorized Online HD Map Construction. In *WACV*, 2024. [2](#), [4](#)
- [34] Shuang Zeng, Xinyuan Chang, Xinran Liu, Zheng Pan, and Xing Wei. Driving with Prior Maps: Unified Vector Prior Encoding for Autonomous Vehicle Mapping. *arXiv preprint arXiv:2409.05352*, 2024. [1](#), [2](#), [6](#)
- [35] Xiaoyu Zhang, Guangwei Liu, Zihao Liu, Ningyi Xu, Yunhui Liu, and Ji Zhao. Enhancing Vectorized Map Perception with Historical Rasterized Maps. In *ECCV*, 2024. [2](#)
- [36] Zhixin Zhang, Yiyuan Zhang, Xiaohan Ding, Fusheng Jin, and Xiangyu Yue. Online Vectorized HD Map Construction using Geometry. In *ECCV*, 2024. [2](#)
- [37] Brady Zhou and Philipp Krähenbühl. Cross-view Transformers for real-time Map-view Semantic Segmentation. In *CVPR*, 2022. [2](#)
- [38] Yi Zhou, Hui Zhang, Jiaqian Yu, Yifan Yang, Sangil Jung, Seung-In Park, and ByungIn Yoo. HIMap: Hybrid Representation Learning for End-to-end Vectorized HD Map Construction. In *CVPR*, 2024. [2](#)