Comparing Waymo's 'Surprise'-Based Metrics with Classic Surrogate Safety Metrics

A Scenario-Based Comparative Study

Master's thesis in Mobility Engineering

XINRONG MAO

DEPARTMENT OF MECHANICS AND MARITIME SCIENCES

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's Thesis in Mobility Engineering

Comparing Waymo's 'Surprise'-Based Metrics with Classic Surrogate Safety Metrics

A Scenario-Based Comparative Study

XINRONG MAO

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mechanics and Maritime Sciences
Division of Vehicle Safety
Chalmers University of Technology
Gothenburg, Sweden 2026

Comparing Waymo's 'Surprise'-Based Metrics with Classic Surrogate Safety Metrics
A Scenario-Based Comparative Study
XINRONG MAO

© XINRONG MAO, 2026.

Supervisor: Minxiang Zhao, Department of Mechanics and Maritime Sciences
Examiner: Jonas Bärgman, Department of Mechanics and Maritime Sciences

Master's Thesis 2026
Department of Mechanics and Maritime Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2026

Comparing Waymo's 'Surprise'-Based Metrics with Classic Surrogate Safety Metrics
A Scenario-Based Comparative Study
XINRONG MAO
Department of Mechanics and Maritime Sciences
Chalmers University of Technology

Abstract

Traffic safety evaluation for automated driving requires methods that can identify potentially critical scenarios before crashes occur. Classical surrogate safety metrics, such as Time Headway (THW), Time-to-Collision (TTC), and Deceleration Rate to Avoid a Crash (DRAC), characterize physical conflict mainly in terms of distance, relative speed, and required braking. However, risk can also arise when surrounding road users behave in ways that deviate from the ego vehicle's expectations. This thesis developed a computational framework for quantifying "surprise" in road traffic environments and investigated whether surprise-based metrics could provide complementary information for identifying risky highway events, with a particular focus on cut-ins.

To quantify surprise, a kinematics-based belief-generation model was developed in an ego-centered coordinate system to estimate the expected future motion of surrounding vehicles. Under a constant-velocity assumption, the model generated a Gaussian belief distribution over future vehicle states. Surprise was then computed by comparing the predicted belief with observed vehicle behavior using four surprise-based metrics: Shannon Surprisal, Residual Information, Bayesian Surprise, and Antithesis. The method was evaluated using naturalistic highway trajectory data from the highD dataset. Surprise thresholds were derived from normal-driving baseline data from highD, and the resulting surprise-based detections were compared against detections obtained from a THW-based classical screening method and manually labelled risky cut-in events.

The results showed that the THW-based classical baseline and the surprise-based approach were strongly related but not identical. The overlap analysis identified 941 shared events out of 1,225 unique detected events, corresponding to a Jaccard overlap of 0.7682. This indicated that many risky cut-ins contained both physical closeness and unexpected motion. The precision–recall comparison showed that the THW-based score achieved the highest average precision, with an AP of 0.982. Among the surprise-based metrics, Antithesis performed best with an AP of 0.949, followed by Bayesian Surprise with 0.912 and Residual Information with 0.781. The sensitivity analysis further showed that the history window had a clearer influence on the surprise-based results than the lookahead time.

Overall, the findings confirm that surprise-based metrics can be useful for describing unexpected vehicle behaviour, but they should not be interpreted as direct replacements for the THW-based classical baseline. Instead, they could serve as complementary indicators to specifically capture surprise, but from the results it is clear (and expected) that the surprise metrics do not capture situational urgency.

A main limitation of the work was that the belief model used in this thesis was based on a simple constant-velocity assumption. Future work should investigate more advanced belief-generation models, such as machine-learning-based or interaction-aware prediction models, to evaluate whether improved prediction can lead to more reliable surprise-based safety assessment.

Keywords: traffic safety, surprise metrics, surrogate safety, cut-in, highD, belief model, time headway, trajectory prediction.

Preface

This report presents the outcome of our master's thesis project carried out at the Department of Mechanics and Maritime Sciences at Chalmers University of Technology during the spring of 2026. The thesis investigates the use of surprise-based traffic safety metrics for identifying risky cut-in events in highway driving scenarios. These metrics are compared with a THW-based classical baseline in order to examine whether surprise-based approaches can provide complementary information for traffic safety evaluation.

The study was conducted using naturalistic highway trajectory data from the highD dataset. The work included the development of a computational framework for trajectory prediction, belief generation, surprise calculation, event detection, and performance evaluation. The analysis focused on risky cut-in events, normal-driving baseline construction, precision–recall evaluation, and the sensitivity of surprise-based detection to temporal parameters.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Minxiang, for his guidance, valuable discussions, and continuous support throughout this thesis work. His suggestions helped me improve the development of the framework, the analysis of the results, and the written presentation of the thesis.

I would also like to thank my examiner, Jonas Bärgman, for his detailed feedback and constructive comments during the thesis process. His comments were very helpful for improving the structure, methodology, and interpretation of the study.

Finally, I would like to thank my family and friends for their continuous support, patience, and encouragement throughout my master's studies.

Xinrong Mao, Gothenburg, June 2026

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis, listed in alphabetical order:

A	Antithesis
AP	Average Precision
BS	Bayesian Surprise
CV	Constant Velocity
DRAC	Deceleration Rate to Avoid a Crash
FSM	Field of Safe Motion
FST	Field of Safe Travel
FV	Following Vehicle
GMM	Gaussian Mixture Model
HD	High Definition
ID	Identifier
KL	Kullback–Leibler
LV	Leading Vehicle
NCV	Nearly Constant Velocity
PR	Precision–Recall
RI	Residual Information
SAE	Society of Automotive Engineers
THW	Time Headway
TTC	Time-to-Collision

Nomenclature

The following nomenclature summarizes the main indices, sets, parameters, variables, and functions used throughout this thesis.

Indices

i	Index for a frame, sample, or observation in a trajectory segment
k	Prediction step index in the discrete-time motion model; also used as mixture-mode index where stated
t	Time or frame index

Sets and Spaces

$\mathcal{B}_{t+k}$	Belief distribution over the predicted vehicle state at future step $t+k$
$\mathcal{C}_t$	Available prediction context at time t
E_S	Set of cut-in events detected by the surprise-based method
E_T	Set of cut-in events detected by the classical THW-based method
$\mathcal{I}$	Interaction context from surrounding traffic participants
$\mathcal{M}$	Map or road-context information used in learning-based belief generation
$\mathcal{X}$	Outcome or hypothesis space used in the formulation of surprise metrics

Parameters and Constants

T	Sampling period between consecutive frames
d	Longitudinal distance or gap between two vehicles
d_{gap}	Net bumper-to-bumper longitudinal gap

H	Prediction horizon
h	History window used to define the prior belief
K	Number of candidate motion modes in a Gaussian mixture model
l_e, l_f	Lengths of the cut-in vehicle and the following vehicle, respectively
N	Number of samples, events, or observations, depending on context
$p95, p99$	95th and 99th percentile thresholds derived from the normal-driving baseline
q_x, q_y	Longitudinal and lateral process-noise intensities
z	Lookahead time used for surprise evaluation
τ	Expectation threshold used in the Antithesis metric

Variables and Functions

$A(P, Q)$	Antithesis surprise between prior belief P and posterior belief Q
a_x, a_y	Longitudinal and lateral acceleration components
BS	Bayesian Surprise
$D_{\mathrm{KL}}(\cdot\|\|\cdot)$	Kullback–Leibler divergence
$DRAC$	Deceleration Rate to Avoid a Crash
$e_{p,0}, e_{v,0}$	Generic initial position and velocity residuals from constant-velocity fitting
$e_{x,0}, e_{y,0}$	Initial longitudinal and lateral position residuals
$e_{v_x,0}, e_{v_y,0}$	Initial longitudinal and lateral velocity residuals
$\mathbf{F}_{\mathrm{2D}}$	State transition matrix of the two-dimensional constant-velocity model
$h_r(x; P)$	Residual Information of observed outcome x under prior belief P
$\boldsymbol{\mu}_{t+k}$	Predicted mean position vector at future step $t + k$
$\boldsymbol{\mu}_{k,t+z}$	Mean of the kth Gaussian mixture component at future time $t + z$
$P(x)$	Prior probability or prior belief density of outcome x
$Q(x)$	Posterior belief density of outcome x
Q_y	Posterior belief distribution after observing y
$\mathbf{P}_0$	Initial covariance matrix for the full state $[x, v_x, y, v_y]^\mathsf{T}$
$P_{x,0}, P_{y,0}$	Longitudinal and lateral initial covariance blocks
$\mathbf{P}_k$	Propagated covariance matrix for the state $[x, v_x, y, v_y]^\mathsf{T}$
$\mathbf{Q}_{\mathrm{2D}}$	Process-noise covariance matrix of the two-dimensional constant-velocity model

$\mathbf{\Sigma}_{k,t+z}$	Covariance of the kth Gaussian mixture component at future time $t+z$
RI	Residual Information
$RI_{p95}, BS_{p95}, A_{p95}$	Baseline 95th percentile thresholds for Residual Information, Bayesian Surprise, and Antithesis
RI_{p99}	Baseline 99th percentile threshold for Residual Information
$S(x; P)$	Shannon Surprisal of outcome x under prior belief P
$\mathbf{x}_{\text{rel}}(t)$	Relative state vector of a surrounding vehicle at frame t
$\mathbf{x}_k$	Vehicle state vector at discrete prediction step k
$\mathbf{x}_t$	Current relative state vector used for prediction
$\bar{\mathbf{x}}_{t+k}$	Predicted mean state at future step $t+k$
$\mathbf{x}_{t+z}$	Predicted or observed vehicle state at lookahead time $t+z$
$\mathbf{X}_{1:t}$	Observed motion history up to time t
$\mathbf{Y}_{t+1:t+H}$	Future trajectory over the prediction horizon
THW	Time Headway
$\text{THW}_{\text{min}}^{\text{post}}$	Minimum finite post-change THW value for a cut-in event
TTC	Time-to-Collision
v_F, v_L	Speeds of the following vehicle and leading vehicle, respectively
v_f	Longitudinal speed of the following vehicle
$v_{x,\text{rel}}, v_{y,\text{rel}}$	Relative longitudinal and lateral velocity components in the ego-centered coordinate system
$v_{x,t}, v_{y,t}$	Current relative longitudinal and lateral velocity components
Δv	Closing speed, defined as $v_F - v_L$
δv_i	Frame-to-frame velocity increment used for process-noise calibration
$\mathbf{w}_k$	Process-noise vector at prediction step k
x	Observed outcome or possible hypothesis state
x'	Most likely outcome under the prior belief distribution
x_e, x_f	Longitudinal positions of the cut-in vehicle and the following vehicle, respectively
$x_{\text{forward}}, y_{\text{left}}$	Ego-centered longitudinal and lateral relative positions
x_t, y_t	Current relative longitudinal and lateral positions
y	New observation used to update the belief distribution
π_k	Probability weight of the kth Gaussian mixture component
$\sigma_x(t+k), \sigma_y(t+k)$	Marginal standard deviations of predicted position at future step $t+k$

Contents

List of Figures

List of Tables

1

Introduction

1.1 Background

As intelligent transportation systems and automated driving technologies continue to develop, traffic safety evaluation has become an increasingly important research topic. Modern vehicles must interpret the surrounding traffic environment, monitor nearby road users, predict their possible future behaviour, and make decisions under uncertainty. This is especially challenging in highway traffic, where vehicles interact at high speed and with limited reaction time. Therefore, safety evaluation methods are needed not only to describe current traffic states, but also to identify situations that may become safety-critical for human drivers, advanced driver assistance systems, and automated driving systems [1, 2]. Such identification is also important for the development and validation of automated driving systems, because rare but safety-critical situations must be analysed even when actual crashes are not available in the data [3].

However, defining what is safety-critical is not always straightforward. A situation may be considered objectively critical when it involves a short time gap, high required deceleration, or increased collision risk. At the same time, a situation may also be perceived as critical if it is unexpected, sudden, or difficult to interpret, even when the physical distance between vehicles is not extremely small [4, 5]. This means that traffic risk (or at least perceived risk) is not only related to physical proximity, but also to how the situation evolves relative to prior expectations.

Traditional traffic safety evaluation methods commonly use surrogate safety metrics such as Time-to-Collision (TTC), Time Headway (THW), and Deceleration Rate to Avoid a Crash (DRAC). These metrics are widely used because they are simple, interpretable, and computationally efficient, and they describe traffic conflicts based on vehicle spacing, relative velocity, and required braking effort [6, 7, 8]. However, they are usually based on simplified kinematic assumptions and predefined thresholds. As a result, they mainly capture the physical relationship between vehicles at a given moment. This may lead to false positives when close but stable interactions are classified as risky, and may also miss situations where an unexpected manoeuvre changes the future interaction before the vehicles become physically close.

To address this limitation, surprise-based traffic safety metrics have recently been proposed. Instead of only measuring distance, time gap, or required deceleration,

these metrics evaluate whether the observed behaviour of a surrounding vehicle is unexpected from the perspective of the ego vehicle. In this framework, the future motion of surrounding vehicles is represented as a probabilistic belief, and the observed motion is compared with this belief. If the observed manoeuvre was previously considered unlikely, or if it strongly changes the prior belief, a higher surprise response is produced. This idea is related to information-theoretic and Bayesian definitions of surprise [9, 10, 11].

This thesis uses naturalistic highway trajectory data from the highD dataset, which provides drone-based vehicle trajectories and has been used for scenario-based validation of highly automated driving systems [12]. Since the data contain many ordinary interactions, different driving directions, and varying traffic conditions, the trajectories are first transformed into an ego-centred coordinate representation. This allows the relative position and motion of surrounding vehicles to be analysed under a consistent sign convention. In addition, normal driving segments are used as a reference distribution for surprise-based metrics, so that unusual behaviour can be defined relative to typical highway interactions rather than by manually fixed thresholds.

Classical surrogate safety metrics and surprise-based metrics therefore assess traffic safety from different perspectives. Classical metrics mainly describe the physical interaction between vehicles, while surprise-based metrics focus on unexpected changes in the surrounding vehicle behaviour. Comparing these two types of metrics can provide insight into how different definitions of safety-criticality affect the identification of risky highway scenarios.

1.2 Research Questions

1. How do surprise-based traffic safety metrics compare with a THW-based classical baseline in identifying risky driving scenarios under the same highway cut-in scenarios?
2. How do the history window h and lookahead time z affect the detection performance of surprise-based metrics for risky cut-in identification?

1.3 Aim and Objectives

This thesis aims to investigate the concept of surprise in traffic safety analysis and to compare surprise-based traffic safety metrics with a classical baseline based on THW. To support this investigation, a kinematics-based probabilistic belief representation is constructed from vehicle trajectory data in the highD dataset to represent expected vehicle motion in highway driving scenarios.

To achieve this aim, the thesis has the following objectives:

1. Develop belief generation models that represent probabilistic expectations of vehicle motion based on vehicle trajectory data from the highD dataset.

2. Design and implement a Python-based toolchain to extract representative highway driving scenarios, such as cut-in maneuvers and hard-braking events.
3. Compute surprise-based traffic safety metrics using the developed belief generation models, and compute a THW-based classical baseline for comparison.
4. Analyze the differences between surprise-based metrics and the THW-based classical baseline in identifying safety-critical driving scenarios, including cases uniquely detected by surprise-based metrics, cases uniquely detected by the THW-based screening method, and cases detected by both methods.
5. Analyze how different settings of surprise metric parameters, particularly the history window and lookahead time, affect the identification results of surprise-based metrics for risky cut-in scenarios.

1.4 Scope

This thesis focuses on highway driving scenarios using the highD dataset. The analysis is limited to vehicle interaction scenarios observed in naturalistic highway traffic, with particular attention to risky cut-in events. Normal driving segments are also used as baseline references for evaluating typical vehicle behaviour. Other traffic environments, such as urban intersections, roundabouts, pedestrian interactions, and mixed traffic settings, are outside the scope of this work.

The study introduces several classical surrogate safety metrics, including TTC, THW, and DRAC, to provide the theoretical background for conventional traffic-conflict assessment. However, the empirical event screening and comparison in this thesis focus on a THW-based classical baseline, using the minimum post-cut-in THW as the operational criterion. TTC and DRAC are therefore discussed as relevant classical metrics, but they are not used in the final event screening or precision–recall comparison. The surprise-based metrics include Shannon Surprisal, Residual Information (RI), Bayesian Surprise (BS), and Antithesis. These metrics identify risk through unexpected changes in the surrounding traffic environment, measured by how strongly the observed vehicle motion deviates from probabilistic beliefs about future motion.

The main purpose of this thesis is to compare how these two perspectives can be used to identify safety-critical events in the same highway driving scenarios. The analysis is based on their different characteristics: the THW-based classical baseline focuses on post-cut-in longitudinal time gap, while surprise-based metrics focus on expectation violation under uncertainty. Based on these interpretations, the thesis examines scenarios detected only by the THW-based screening method, scenarios detected only by surprise-based metrics, and scenarios detected by both. These groups are then evaluated using manually labelled reference data to assess how well each approach identifies risky cut-in events and distinguishes them from normal driving interactions.

In addition, this thesis analyzes how key parameter settings in the surprise-based metrics affect the final identification results. Particular attention is given to the

history window h and the lookahead time z, since these parameters determine how prior and posterior beliefs are constructed and compared. This analysis is used to examine the sensitivity of surprise-based risky cut-in identification to different temporal settings.

The surprise-based metrics are computed from a probabilistic belief representation, which represents expected future vehicle motion as a distribution rather than as a single deterministic trajectory. However, this thesis does not aim to provide a general benchmark of trajectory prediction methods. The belief representation is used to support the computation of surprise-based metrics, rather than being evaluated as a standalone prediction model. Therefore, the conclusions of this thesis are limited to the selected dataset, scenario types, metric definitions, parameter settings, manual labelling procedure, and modelling assumptions adopted in this study.

2

Theory

2.1 Classical Surrogate Safety Metrics

Classical surrogate safety metrics are widely used to assess potential traffic conflicts when crash outcomes are not directly observed. Rather than relying on crash records, these metrics estimate conflict severity from the relative motion states of interacting vehicles, such as spacing, speed, relative speed, and deceleration demand.

This section reviews three classical surrogate safety metrics: THW, TTC, and DRAC. These metrics describe potential traffic conflicts from different perspectives. THW describes the time gap between a following vehicle and its leader, with a lower value indicating closer following. TTC estimates the time remaining until a collision would occur if the vehicles continued with their current motion states. DRAC describes the deceleration required to avoid a crash under the same assumptions, with a higher value indicating greater braking demand. Together, THW, TTC, and DRAC define traffic conflict severity in terms of time gap, collision urgency, and required braking effort.

Despite their widespread application and practical interpretability, classical surrogate safety metrics are limited to observable motion-based relationships and do not explicitly represent probabilistic expectation or uncertainty in future behaviour. In the empirical comparison later in this thesis, the classical baseline is based on post-cut-in THW. TTC and DRAC are retained as relevant theoretical reference metrics, but they are not used in the final event screening or precision–recall comparison.

2.1.1 THW

THW is a commonly used surrogate safety metric for describing the temporal spacing condition between a following vehicle (FV) and a leading vehicle (LV). In this thesis, THW is used to describe the post-change temporal gap between the cut-in vehicle and its new follower. It is computed from the net bumper-to-bumper longitudinal gap and the speed of the following vehicle.

It should be noted that the terms time headway and time gap are not always used consistently in the traffic-safety literature. In a strict SAE definition, time headway refers to the time difference between the same reference point of two vehicles passing a given location, while time gap refers to the clearance time based on the space between the rear of the leading vehicle and the front of the following vehicle

[13]. The metric used in this thesis is based on the net bumper-to-bumper gap and therefore corresponds more closely to time gap. However, since similar gap-based definitions are commonly referred to as THW in surrogate safety studies, and since the implemented pipeline uses this convention, the term THW is retained in this thesis. The exact computation is explicitly given below.

The THW value is defined as

$$\mathrm{THW} = \frac{d}{v_F}, \tag{2.1}$$

where d denotes the longitudinal distance gap between the FV and the LV, and v_F is the speed of the FV.

A smaller THW indicates a shorter time gap and therefore a closer following condition, while a larger THW generally represents a more comfortable longitudinal spacing. In cut-in scenarios, THW can be used to describe how much longitudinal safety margin remains after a lane-changing vehicle enters the lane in front of the ego vehicle. If the inserted gap is small or the ego vehicle speed is high, the THW value may decrease rapidly, indicating a potentially more critical interaction.

The main advantage of THW is its simple physical meaning and low computational cost, since it only requires distance and speed information. However, THW has several limitations. It does not explicitly consider the relative speed between vehicles, and therefore cannot distinguish well between situations with the same spacing but different closing speeds. In addition, THW mainly describes longitudinal spacing and does not directly capture the lateral motion involved in lane-changing or cut-in manoeuvres. For this reason, THW is better interpreted as a simple spacing-based indicator rather than a complete measure of interaction risk. In this thesis, THW is used as the classical baseline metric for characterising the longitudinal severity of post-cut-in situations.

2.1.2 TTC

TTC is a commonly used indicator in traffic safety analysis for assessing the criticality of rear-end interactions [14]. It represents the time remaining until a collision would occur between an FV and an LV, under the assumption that both vehicles maintain their current velocities. The TTC value is given by

$$\mathrm{TTC} = \frac{d}{\Delta v}, \tag{2.2}$$

where d is the longitudinal gap between the two vehicles and $\Delta v = v_F - v_L$ is the closing speed. This expression is meaningful only when the FV is closing in on the LV, that is, when $\Delta v > 0$. If the FV is not closing in, the interaction is usually not treated as an immediate rear-end collision course under the TTC assumption.

2.1.3 DRAC

DRAC is a surrogate safety metric that characterises the severity of a potential rear-end collision in terms of required braking effort [8]. It describes the constant

deceleration that the FV must apply to avoid a collision with the LV, assuming unchanged kinematic conditions. DRAC is computed as

$$\mathrm{DRAC} = \frac{\Delta v^2}{2d}, \tag{2.3}$$

where $\Delta v = v_F - v_L$ is the closing speed and d denotes the longitudinal distance separating the two vehicles. A larger DRAC value indicates that stronger braking would be required to avoid a collision.

2.2 Surprise-Based Metrics

Surprise-based metrics describe traffic situations by comparing observed road user behaviour with probabilistic predictions of expected future motion. Their central idea is that surprise occurs when the realised behaviour deviates from what was previously considered likely [10, 11]. Unlike classical surrogate safety metrics, which focus on kinematic conflict severity, surprise-based metrics focus on expectation violation under uncertainty [11].

To support this analysis, future motion is represented as a probabilistic belief rather than as a single deterministic prediction [15, 10]. On this basis, different surprise metrics quantify the mismatch between prediction and observation in different ways [10]. In this thesis, four surprise-based metrics are considered: Shannon Surprisal, Residual Information, Bayesian Surprise, and Antithesis. The formal definitions and theoretical background of these metrics are introduced in the following subsections.

2.2.1 Shannon Surprisal

Shannon Surprisal originates from information theory and measures the information content, or unexpectedness, of an observed outcome under a given belief model [16]. Let X denote a discrete random variable with outcome space $\mathcal{X}$, and let P represent a prior belief distribution over X. When a specific outcome $x \in \mathcal{X}$ is observed, the Shannon Surprisal is defined as

$$S(x; P) = -\log P(x), \tag{2.4}$$

where $P(x)$ denotes the probability assigned to the observed outcome by the prior belief distribution P.

The notation $S(x; P)$ emphasises that surprisal is evaluated for a realised outcome x with respect to the prior distribution P [16]. The choice of logarithm base determines the unit of measurement: using a logarithm with base 2 gives units of bits, while using the natural logarithm gives units of nats [16].

Shannon Surprisal assigns higher values to outcomes that are less probable under the prior belief distribution [16]. If an outcome is highly expected, that is, if $P(x)$ is large, its surprisal is low. Conversely, outcomes that were assigned low prior probability produce large surprisal values. In the limiting case, an outcome that

is certain under the prior, $P(x) = 1$, yields zero surprisal, while outcomes with vanishing probability lead to unbounded surprisal [16].

This leads to what can be described as a *surprise floor* problem [10]. Intuitively, when the most expected event occurs, humans do not feel surprised. However, Shannon Surprisal assigns a non-zero value even to the most likely outcome. Specifically, let

$$x' = \arg\max_{x \in \mathcal{X}} P(x) \tag{2.5}$$

denote the outcome with the highest prior probability. As long as $P(x') < 1$, its surprisal remains strictly positive:

$$S(x'; P) = -\log P(x') > 0. \tag{2.6}$$

As a result, surprisal remains positive even for outcomes that are maximally expected under the belief model, which does not fully align with the human intuition that fully expected events should induce no surprise [10]. This limitation is addressed by the Residual Information measure, which adjusts the surprisal value relative to the most expected outcome under the same belief distribution [10].

2.2.2 Residual Information

Residual Information is a surprise measure that evaluates how much an observed outcome deviates from the most expected one [10]. It is defined as

$$h_r(x; P) = \log\left(\frac{P(x')}{P(x)}\right), \tag{2.7}$$

where x denotes the observed outcome, and x' denotes the outcome with the highest probability under the belief distribution P.

This formulation directly compares the probability of what actually happened with the probability of what was most strongly expected [10]. If the observed outcome coincides with the most likely prediction, then $P(x) = P(x')$ and the residual information is zero. As the observed outcome deviates from the dominant expectation, the residual information increases.

By construction, the most expected outcome always yields zero surprise. As a result, Residual Information does not suffer from the *surprise floor* problem and aligns well with intuitive interpretations of surprise [10].

2.2.3 Bayesian Surprise

Bayesian Surprise describes surprise as the amount of belief change caused by a new observation [9]. Let P denote the prior belief and let Q_y denote the posterior belief after observing y. Bayesian Surprise is defined as the Kullback–Leibler (KL) divergence between the posterior and the prior belief [9, 17]:

$$\mathrm{BS}(y) = D_{\mathrm{KL}}(Q_y \,\|\, P), \tag{2.8}$$

where the first argument is the updated belief and the second argument is the original belief.

This formulation measures surprise as the difference between the belief before and after observing new evidence, quantifying how much the overall expectation of future behaviour is updated [9, 10].

A small value of Bayesian Surprise indicates that the observation is consistent with prior expectations, while a large value indicates that the observation leads to a significant update of the belief distribution [9]. Unlike Shannon Surprisal, Bayesian Surprise is a global measure that reflects changes in the overall belief structure rather than the improbability of a single outcome [10].

2.2.4 Antithesis

To address this flooring problem, the Antithesis metric was proposed as an extension of Bayesian Surprise [10]. Like Bayesian Surprise, Antithesis is based on the difference between the prior and posterior belief distributions [10]. However, the two metrics differ in how this belief change is evaluated. Bayesian Surprise integrates the belief update over the entire belief space, including regions where the observation is already well supported by the prior belief. As a result, it may still assign a positive surprise value to largely expected observations [10].

Antithesis instead focuses on the part of the belief update that corresponds to genuinely surprising information [10]. Belief changes in regions that are already expected are suppressed, while changes in unexpected regions are retained. In this way, Antithesis provides a more selective surprise measure and reduces the flooring effect observed in Bayesian Surprise [10].

Formally, let x denote a possible outcome in the hypothesis space $\mathcal{X}$, such as a candidate future state or trajectory of the target vehicle. Let $P(x)$ denote the prior belief distribution before observing new evidence, and let $Q(x)$ denote the posterior belief distribution after incorporating the evidence. Antithesis surprise is defined as follows [10]:

$$A(P, Q) = \int_{\{x \in \mathcal{X} \colon P(x) < \tau,\, Q(x) > P(x)\}} Q(x) \, \log\left(\frac{Q(x)}{P(x)}\right) \, \mathrm{d}x, \tag{2.9}$$

where τ is an expectation threshold that separates typical outcomes from atypical outcomes under the prior belief [10].

This definition highlights two necessary conditions for an outcome to contribute to Antithesis surprise [10]. First, the outcome must lie outside the expectation region, meaning that it was assigned a relatively low probability under the prior belief, $P(x) < \tau$. Second, the outcome must exhibit increased belief after observing new evidence, such that its posterior probability exceeds its prior probability, $Q(x) > P(x)$. Only outcomes that satisfy both conditions are considered. For these outcomes, the term $\log(Q(x)/P(x))$ measures the information gain introduced by the belief update, and this gain is weighted by the posterior probability $Q(x)$.

Figure 2.1 illustrates a case where belief updates do not extend beyond the original expectation region, resulting in zero Antithesis surprise.

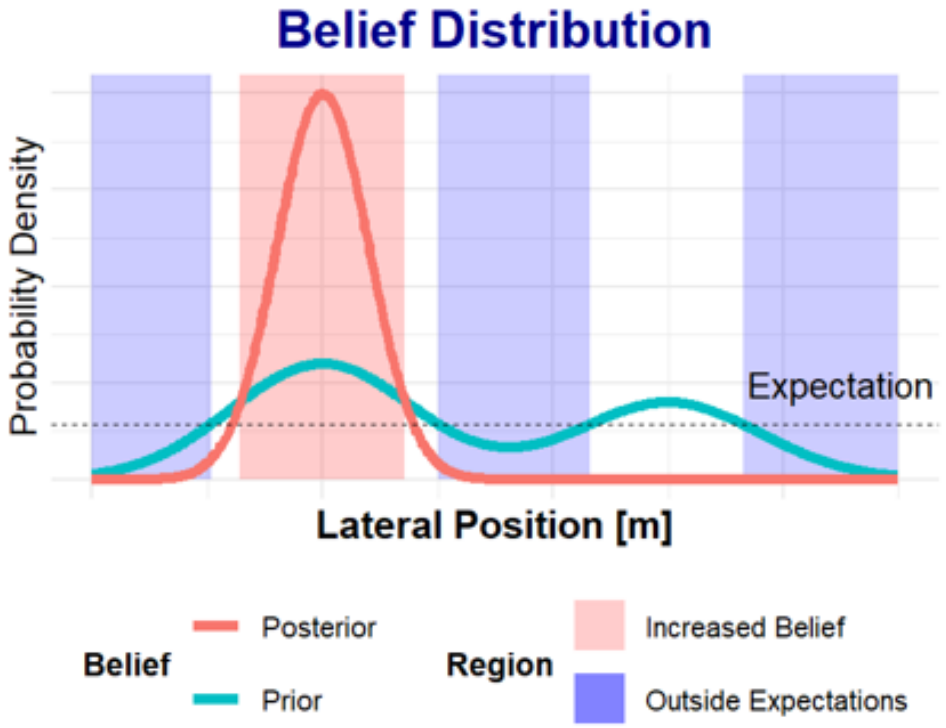

Figure 2.1: Illustration of Antithesis when the regions *Outside Expectation* and *Increased Belief* do not overlap.

The region of surprising information is defined as the overlap of two sub-regions, as illustrated in Figure 2.1 [10]. The first is the *Outside Expectation* region, where the prior belief assigns a low probability to an outcome. In this work, an expectation threshold is defined using the average probability of the prior distribution. Outcomes with prior probabilities below this threshold are shown as the blue shaded areas in the figure.

The second sub-region is the *Increased Belief* region, where the posterior belief after observing new evidence is higher than the prior belief [10]. This region is shown as the red shaded area and represents outcomes that become more likely after the observation.

Antithesis considers surprise only in the overlap of these two regions [10]. In other words, an outcome is considered surprising only if it was originally unexpected and becomes more likely after the belief update. This overlap corresponds to the introduction of new expectations that were not present in the prior belief.

In the example shown in Figure 2.1, the *Outside Expectation* region and the *Increased Belief* region do not overlap. As a result, there is no region over which the Antithesis metric is evaluated, and the Antithesis surprise value is zero [10]. This example shows that Antithesis ignores belief updates that merely reinforce exist-

ing expectations and responds only to belief changes that introduce new predicted behaviour.

2.3 Belief Generation for Vehicle Motion

Belief generation for vehicle motion aims to describe future vehicle behaviour in a probabilistic way. Instead of providing only a single deterministic trajectory, it represents possible future motions together with their likelihoods, so that uncertainty in traffic scenes can be explicitly captured. In surprise-based safety assessment, this belief distribution serves as the expected motion model against which the realised motion is compared. Therefore, the quality of the belief representation directly affects the interpretation of surprise metrics [15].

2.3.1 Belief Generation Based on Kinematic Models

To compute surprise-based safety metrics, a probabilistic belief over future vehicle motion is required. In this thesis, such a belief is generated using a kinematic motion model. The model itself is not sufficient without suitable trajectory data: it requires frame-level motion states, including vehicle position, vehicle velocity, a known sampling time step, and a consistent coordinate representation. These requirements are important because the predicted belief is propagated from the observed current state, and the observed future trajectory is later compared with this belief. The detailed preprocessing of the highD trajectory data and the construction of these model inputs are described in the methodology chapter.

Kinematic models describe motion through low-order derivatives of position and represent deviations from ideal motion by stochastic process noise. They are widely used because they are physically interpretable, computationally efficient, and suitable for short-horizon prediction. In the estimation literature, these models are also referred to as polynomial models, since the nominal motion is obtained by setting a certain derivative of position to zero, while uncertainty is introduced through random disturbances [18].

In this work, the constant velocity (CV) model is adopted as the kinematics-based belief representation used for the surprise calculations. To avoid introducing separate one-dimensional and two-dimensional versions of the same model, the CV model is written directly in the coordinate form used in this thesis. The vehicle state is defined as

$$\mathbf{x}_k = \begin{bmatrix} x_k \\ v_{x,k} \\ y_k \\ v_{y,k} \end{bmatrix}, \tag{2.10}$$

where x_k and y_k are the longitudinal and lateral positions, and $v_{x,k}$ and $v_{y,k}$ are the corresponding velocities at time step k.

Under the nominal CV assumption, acceleration is zero, which means that the object moves with constant velocity in the absence of disturbances. Since real vehicles do

not follow perfectly constant-velocity motion, deviations are represented by Gaussian process noise:

$$\mathbf{x}_{k+1} = \mathbf{F}_{2\mathrm{D}}\mathbf{x}_k + \mathbf{w}_k, \qquad \mathbf{w}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_{2\mathrm{D}}). \tag{2.11}$$

The system matrix and process-noise covariance are then given by one unified expression:

$$\begin{aligned} \mathbf{F}_{2\mathrm{D}} &= \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix}, \\ \mathbf{Q}_{2\mathrm{D}} &= \begin{bmatrix} q_x\frac{T^3}{3} & q_x\frac{T^2}{2} & 0 & 0 \\ q_x\frac{T^2}{2} & q_xT & 0 & 0 \\ 0 & 0 & q_y\frac{T^3}{3} & q_y\frac{T^2}{2} \\ 0 & 0 & q_y\frac{T^2}{2} & q_yT \end{bmatrix}. \end{aligned} \tag{2.12}$$

Here, T is the sampling period, and q_x and q_y are the process noise intensities in the longitudinal and lateral directions, respectively. This formulation presents the implemented CV belief model in a single consistent form while allowing the uncertainty in longitudinal and lateral motion to be modelled separately, which is particularly useful in highway scenarios where car-following and lane-changing behaviours may exhibit different dynamic characteristics [18].

Based on the above model, the belief at a future prediction horizon is obtained by propagating the current state estimate forward in time. As a result, each future step is associated with a Gaussian distribution over the predicted vehicle position. The predicted mean represents the most likely motion under the kinematic assumption, while the covariance describes the uncertainty that grows with prediction time. This probabilistic belief forms the basis for the subsequent computation of surprise-based metrics, since the observed future trajectory can then be evaluated against the predicted belief distribution.

2.3.2 Belief Generation Based on Learning-Based Models

Unlike kinematic models, learning-based models do not rely on predefined motion equations such as constant velocity or constant acceleration. Instead, they learn a mapping from observed traffic context to future motion directly from data. In this framework, the task of belief generation can be formulated as learning a conditional distribution over future trajectories given historical observations, surrounding agents, and road context:

$$p(\mathbf{Y}_{t+1:t+H} \mid \mathbf{X}_{1:t}, \mathcal{M}, \mathcal{I}), \tag{2.13}$$

where $\mathbf{X}_{1:t}$ denotes the observed motion history of the target vehicle, $\mathcal{M}$ denotes map information, and $\mathcal{I}$ denotes the interaction context from surrounding traffic participants. Compared with kinematic models, this formulation has broader data requirements: it needs sufficient history of the target vehicle, information about surrounding agents, and, when used, static road or map context. In addition, a

learning-based model must be trained on representative trajectory data so that it can learn the relationship between traffic context and future motion.

A central advantage of learning-based models is that they can represent the inherently multimodal nature of future driving behaviour. In real traffic, the same observed state may lead to several plausible futures, such as lane keeping, lane changing, braking, or accelerating. Therefore, learning-based belief generation should not be understood as predicting a single deterministic trajectory, but rather as estimating a probability distribution over multiple feasible future motions. This probabilistic interpretation is especially important for surprise-based safety metrics, since surprise is defined relative to an observer's prior belief distribution rather than to one single nominal prediction [10].

Recent trajectory prediction models achieve this by jointly encoding dynamic agent histories and structured scene context. A representative example is VectorNet, which represents both HD map elements and agent trajectories in vectorised form rather than rasterised images. In VectorNet, map entities such as lane boundaries, crosswalks, and traffic signs, as well as observed vehicle trajectories, are approximated as polylines and then decomposed into vectors. Each vector is treated as a node with geometric and semantic attributes, which allows the traffic scene to be represented as a graph [19]. This vectorised representation preserves structural information in a more direct way than image-based rendering and provides an efficient foundation for motion forecasting.

To model scene structure, VectorNet [19] employs a hierarchical graph architecture. At the first level, vectors belonging to the same polyline are aggregated into polyline-level features, capturing local spatial and semantic structure. At the second level, all polyline features are connected through a global interaction graph to model higher-order dependencies among map elements and traffic participants. In this way, the model can capture both local geometric continuity and global interaction effects, which are critical for predicting behaviour in complex traffic scenarios. After encoding, the future trajectory of the target vehicle is decoded from the learned scene representation.

From the perspective of surprise modelling, the key requirement is that the learned predictor should provide a probabilistic belief over future states. Dinparastdjadid et al. describe such a belief as the output of a machine-learned generative model that predicts how a traffic situation may evolve based on static and dynamic context. In their formulation, the model produces a set of weighted candidate trajectories together with likelihood information. At each future time step, the uncertainty around the predicted state along each candidate trajectory can be represented by a Gaussian distribution. By combining the Gaussian predictions from multiple candidate trajectories, the resulting belief at a given future timestamp becomes a Gaussian Mixture Model (GMM) over future position [10]. This provides a natural probabilistic representation for comparing expected and observed motion.

More formally, the learned belief at a future time $t + z$ may be expressed as

$$p(\mathbf{x}_{t+z} \mid \mathcal{C}_t) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}\Big(\mathbf{x}_{t+z}; \boldsymbol{\mu}_{k,t+z}, \boldsymbol{\Sigma}_{k,t+z}\Big), \tag{2.14}$$

where $\mathcal{C}_t$ denotes the available context at time t, K is the number of candidate motion modes, π_k is the probability weight of the kth mode, and $\mathcal{N}(\mathbf{x}_{t+z}; \boldsymbol{\mu}_{k,t+z}, \boldsymbol{\Sigma}_{k,t+z})$ is the Gaussian density describing the predicted state uncertainty for that mode. In practical applications, $\mathcal{C}_t$ must contain enough data to distinguish between plausible future manoeuvres, such as recent motion history, neighbouring vehicle states, and relevant road-context information. The estimation of the mixture weights, means, and covariances also depends on the data used to train or calibrate the prediction model. This formulation explicitly captures two forms of uncertainty: uncertainty over which manoeuvre the vehicle will take, and uncertainty over the exact state along a selected manoeuvre [10].

An additional advantage of this formulation is that it separates the time when a prediction is generated from the future time about which the prediction is made. This distinction is essential for later surprise computation. The prior belief can be generated at an earlier time instant, while the posterior belief is updated using more recent observations, but both beliefs are defined over the same future state. Therefore, learning-based generative models not only provide trajectory forecasts, but also establish the probabilistic belief structure required for probabilistic mismatch and belief mismatch surprise measures [10].

In summary, learning-based belief generation models provide a flexible and expressive alternative to classical kinematic predictors. By learning directly from data, they can integrate road geometry, multi-agent interaction, and multimodal future evolution into a unified probabilistic prediction framework. For this reason, they are well suited for surprise-based traffic safety analysis, where the key quantity of interest is not only the most likely future motion, but the full belief distribution against which observed behaviour is evaluated. In this thesis, however, learning-based models are discussed as theoretical context rather than implemented as part of the evaluation. The implemented surprise calculations use the kinematics-based belief representation described in the previous subsection.

3

Methodology

3.1 Overall Workflow

Figure 3.1 summarizes the overall methodology adopted in this thesis. The analysis begins with the processing of the highD dataset to extract the trajectory, lane, and interaction information required for risky cut-in assessment.

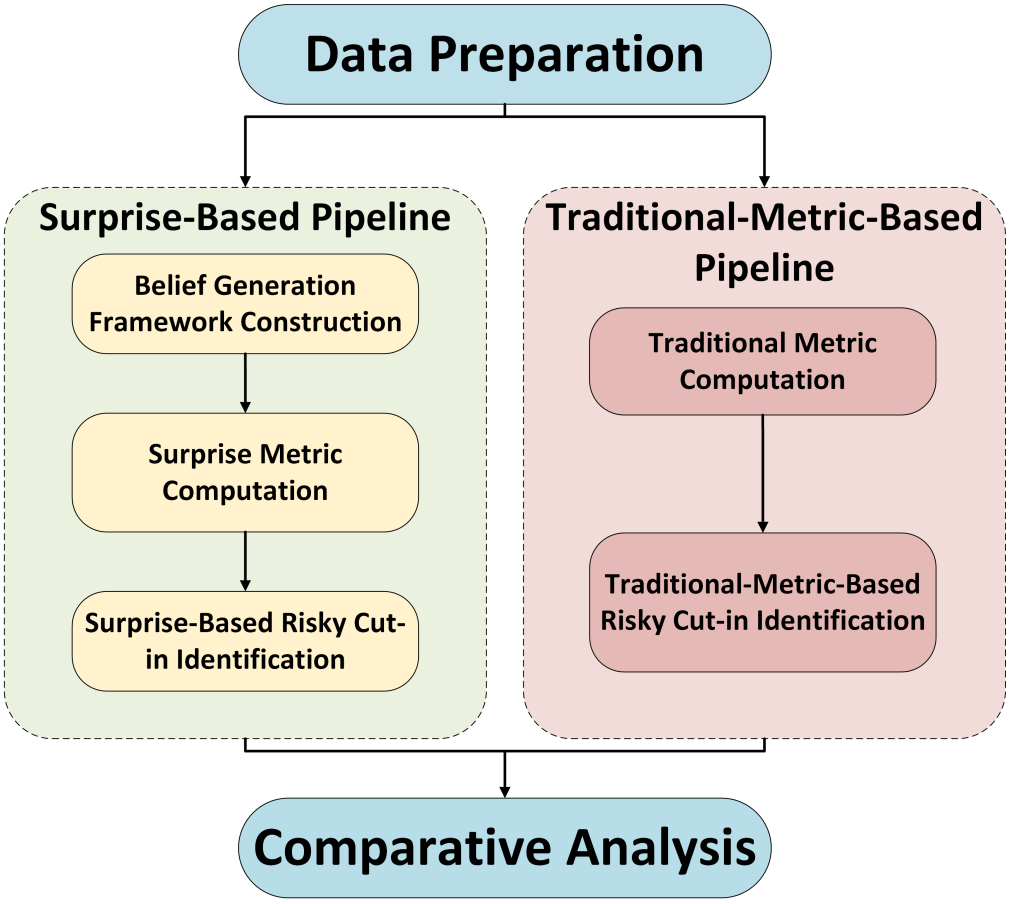

Figure 3.1: Overall workflow for risky cut-in identification and comparison using surprise-based metrics and a THW-based classical baseline.

After data preparation, the study proceeds along two parallel branches. In the

surprise-based branch, a belief generation framework is constructed to represent future vehicle motion in probabilistic form. The predicted belief and the observed future motion are then combined to compute surprise-based safety metrics, and risky cut-in events are identified from cut-in candidates that simultaneously exhibit abnormal surprise responses.

In parallel, the THW-based classical branch identifies risky cut-in events directly from the prepared dataset using the minimum post-cut-in THW threshold and supporting geometric filters. Unlike the surprise-based branch, this pipeline does not require explicit belief generation. Instead, it evaluates hazardous interactions based on post-cut-in temporal spacing and predefined safety criteria.

Finally, the outputs of the two branches are compared systematically. The comparison considers the number of detected events, the overlap between the detected event sets, and the types of risky cut-in behavior emphasized by each method. This workflow therefore provides a structured basis for evaluating the differences and complementarities between belief-based surprise metrics and the THW-based classical baseline in cut-in safety assessment.

3.2 Dataset and Scenario Preparation

3.2.1 Dataset Description

This thesis used the highD dataset, which contains naturalistic highway trajectory data recorded from drone-based observations. The dataset provides frame-level vehicle trajectories and vehicle-level metadata for vehicles observed on German highways. These data are suitable for this thesis because they describe real highway interactions at a high temporal resolution and include manoeuvres such as lane changes, car-following interactions, and cut-in situations. In this thesis, all 60 local highD recordings were used. The recordings were sampled at 25 Hz, corresponding to a time step of 0.04 s.

An overview of the dataset information relevant to this thesis is given in Table 3.1. In addition to the number of recordings and the sampling rate, the table reports exposure-based quantities, including the number of observed vehicles, vehicle hours, and vehicle kilometres. These quantities describe the amount of vehicle-level trajectory data available for the analysis, rather than only the duration of the video recordings.

Table 3.1: Overview of the highD trajectory data used in this thesis.

Dataset characteristic	Value
Number of recordings	60
Recording locations	6 highway locations
Sampling rate	25 Hz
Total recording duration	16.66 h
Number of observed vehicles	110,516
Vehicle hours	441.40 h
Vehicle kilometres	44,475.61 km
Average vehicle speed	100.76 km/h
Lane changes recorded in track metadata	13,949
Vehicle classes	89,139 cars; 21,377 trucks
Main scenario type analysed in this thesis	Risky cut-in events
Baseline reference data	Normal-driving vehicle interactions

Here, vehicle hours refer to the accumulated visible duration of all individual vehicle trajectories, while vehicle kilometres refer to the accumulated travelled distance of all observed vehicles. The average vehicle speed was calculated as the ratio between vehicle kilometres and vehicle hours. The lane-change value refers to the lane changes recorded in the local track metadata and is therefore reported as a dataset descriptor rather than as the final number of risky cut-in events used in the analysis.

Overall, the dataset provided a large amount of vehicle-level trajectory information under naturalistic highway conditions. This was important for two parts of the analysis. First, the large number of observed vehicles and lane-change manoeuvres provided the basis for extracting candidate cut-in events. Second, the continuous frame-level trajectories allowed normal-driving interactions to be selected as baseline references for estimating typical surprise-metric values.

3.2.2 Target Scenario Types

The main target scenario considered in this study is highway cut-in interaction. A cut-in event generally refers to a situation in which a vehicle changes lane and enters the lane space in front of or near another vehicle, thereby altering the local interaction pattern and potentially increasing safety risk. Such scenarios are of particular interest in highway traffic analysis because they involve both lateral maneuvering and longitudinal interaction between multiple traffic participants.

Cut-in behavior is selected as the primary scenario type for two main reasons. First, it is a common and representative form of vehicle interaction in multilane highway traffic, especially in dense or moderately congested conditions where lane changes frequently influence the motion of surrounding vehicles. Second, the safety relevance of a cut-in event cannot be fully understood from the lane-changing vehicle alone,

but must be interpreted together with the responses and relative states of nearby vehicles. This makes cut-in scenarios well suited for a comparative study between a THW-based classical baseline and surprise-based metrics, since both approaches aim to capture different aspects of interaction-related risk.

At this stage, the study focuses on risky cut-in interactions as the target scenario in highway traffic. The purpose of this subsection is to define the scenario domain and motivate its relevance for traffic safety analysis, while the detailed identification of risky cases is introduced in the subsequent methodology sections.

3.2.3 Coordinate System Representation and Relative-State Construction

To analyze vehicle interactions, it is more useful to describe surrounding vehicles from the perspective of the ego vehicle rather than only using the original global coordinates in the highD dataset. This is because safety-related situations, such as car-following and cut-in interactions, mainly depend on the relative position and relative motion between vehicles. Therefore, before further analysis, the raw trajectory data are transformed into an ego-centered relative-state representation.

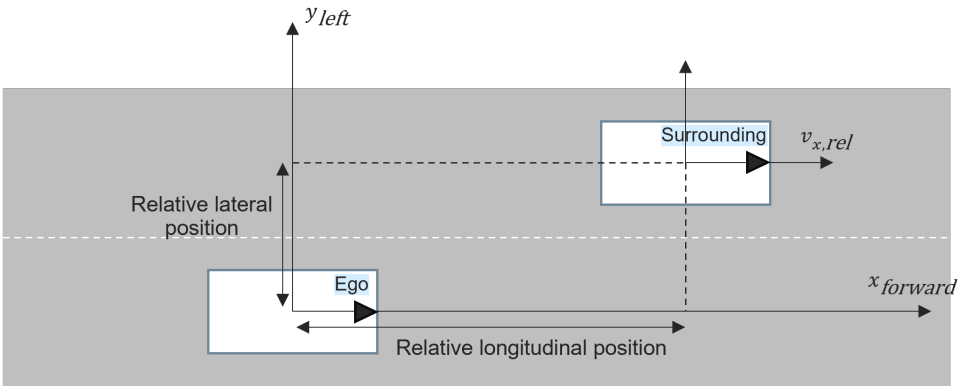

Figure 3.2: Ego-centered coordinate system for representing surrounding vehicle states.

As shown in Figure 3.2, the ego vehicle is used as the local reference vehicle. The origin of the coordinate system is placed at the geometric center of the ego vehicle. The longitudinal axis x_{forward} points along the driving direction of the ego vehicle, while the lateral axis y_{left} points to the left side of the ego vehicle.

For each surrounding vehicle, its state is expressed relative to the ego vehicle in this local coordinate system. The relative longitudinal position describes how far the surrounding vehicle is in front of or behind the ego vehicle, while the relative lateral position describes its offset to the left or right. In addition, the relative velocity components $v_{x,\text{rel}}$ and $v_{y,\text{rel}}$ describe the motion of the surrounding vehicle with respect to the ego vehicle in the longitudinal and lateral directions.

The relative state of a surrounding vehicle at frame t is represented as

$$\mathbf{x}_{\text{rel}}(t) = \begin{bmatrix} x_{\text{forward}}(t) \\ v_{x,\text{rel}}(t) \\ y_{\text{left}}(t) \\ v_{y,\text{rel}}(t) \end{bmatrix}. \tag{3.1}$$

To ensure consistency across recordings and driving directions, coordinate signs are normalised based on the ego vehicle's direction of travel. After normalisation, positive x_{forward} always represents the forward direction of the ego vehicle, and positive y_{left} always represents the left side of the ego vehicle.

Both analysis pipelines use the constructed relative state $\mathbf{x}_{\text{rel}}(t)$. In the surprise-based pipeline, relative position and relative velocity are used to generate probabilistic predictions of surrounding vehicle motion, which are then compared to the observed future relative position. In the THW-based classical pipeline, relative spacing and follower speed are used to calculate post-cut-in THW for event screening.

3.3 Trajectory Prediction and Belief Generation Model

This thesis treats trajectory prediction and belief generation as two sequential components. The trajectory prediction model first estimates the expected future motion of surrounding vehicles and generates a deterministic mean trajectory. The belief generation model then adds uncertainty around this mean trajectory and represents the prediction as a probabilistic belief [15, 10]. Therefore, trajectory prediction describes where a vehicle is expected to be, while belief generation describes the uncertainty of this expectation. This belief is later used as the reference for calculating surprise-based metrics.

A kinematics-based constant-velocity model was selected for this purpose [18]. The aim was not to develop a complex trajectory forecasting model, but to define a simple and consistent expectation against which observed vehicle motion could be compared. Under this assumption, a surrounding vehicle is expected to continue with its current relative motion. When the observed future motion differs from this expectation, the difference can be interpreted as a mismatch between expected and observed behaviour.

This modelling choice is suitable for the purpose of this thesis because it provides a transparent baseline for highway driving scenarios. In normal short-term car-following, vehicle motion is often relatively smooth, so the constant-velocity assumption gives a reasonable reference expectation. In contrast, manoeuvres such as cut-ins, hard braking, or sudden acceleration can lead to clear deviations from this reference. These deviations are directly relevant for surprise-based safety evaluation.

3.3.1 Kinematic-Based Trajectory Prediction

In this thesis, a kinematics-based trajectory prediction model is used to calculate the future mean trajectory of surrounding vehicles. The model relies on a deterministic constant-velocity assumption. Its goal is not to capture complex driver intentions or manoeuvre decisions, but rather to provide a straightforward and interpretable prediction baseline for the subsequent belief generation process.

The prediction is made using the ego-centered coordinate system. For each ego vehicle at each valid frame, nearby vehicles are first chosen from a predefined neighbourhood region. A vehicle is retained if it appears in the same frame as the ego vehicle, is not the ego vehicle itself, meets the same-direction condition when driving-direction information is available, and is within the selected interaction window. This study limits the forward range to 0 to 200 m and the lateral range to -25 to 25 m.

For each selected surrounding vehicle, the current relative state follows the same order as the CV state in Equation 2.10:

$$\mathbf{x}_t = \begin{bmatrix} x_t \\ v_{x,t} \\ y_t \\ v_{y,t} \end{bmatrix}, \tag{3.2}$$

where x_t denotes the relative longitudinal position in the ego-forward direction, y_t denotes the relative lateral position toward the left side of the ego vehicle, and $v_{x,t}$ and $v_{y,t}$ are the corresponding relative velocity components.

The future mean trajectory is then generated by applying the unified CV transition model from Equation 2.11. For prediction step k, this gives

$$\bar{\mathbf{x}}_{t+k} = \mathbf{F}_{\mathrm{2D}}^{k}\mathbf{x}_t, \qquad k = 1, 2, \ldots, H, \tag{3.3}$$

where H is the prediction horizon and the sampling period in $\mathbf{F}_{\mathrm{2D}}$ is set to $T = 0.04$ s for the 25 Hz highD data.

In the implementation, the prediction horizon is set to $H = 125$ frames, corresponding to a 5-second look-ahead duration. Therefore, for each selected surrounding vehicle, the model generates a sequence of predicted future relative positions from the current frame to 5 seconds into the future.

3.3.2 Belief Generation Based on Kinematic Trajectory Prediction

Algorithm 1 Belief generation based on a kinematics-based trajectory model

Require: highD trajectories, metadata, prediction horizon H, sampling period T
Ensure: Gaussian belief predictions for surrounding vehicles

```
Load trajectory data and metadata
Define the ego-centered coordinate system and interaction region
for each selected recording r do
    Build frame-level vehicle-state indices
    for each ego vehicle e do
        for each frame t do
            Transform surrounding vehicles into the ego-centered frame
            Select vehicles within the interaction region
            for each selected surrounding vehicle j do
                Extract relative state x_t = [x_t, v_{x,t}, y_t, v_{y,t}]^T
                Initialize covariance P_0
                for k = 1 to H do
                    Propagate mean state using Equation 2.11
                    Extract μ_{t+k} from state components 1 and 3
                    Propagate covariance using Equation 2.12
                    Extract σ_x(t+k) and σ_y(t+k) from the position variances
                    Construct Gaussian belief B_{t+k}
                    Store prediction and available future observation
                end for
            end for
        end for
    end for
end for
Return belief prediction table for surprise-metric computation
```

Algorithm 1 summarizes how the kinematics-based prediction results are converted into probabilistic beliefs. Since the ego-centered coordinate representation and the constant-velocity prediction model have already been introduced, this subsection focuses on how the predicted trajectory is represented as a Gaussian belief distribution for later surprise computation.

Using the current state $\mathbf{x}_t$ and the predicted mean state $\bar{\mathbf{x}}_{t+k}$ defined in the trajectory-prediction subsection, the Gaussian mean is extracted from the longitudinal and lateral position components:

$$\boldsymbol{\mu}_{t+k} = \begin{bmatrix} [\bar{\mathbf{x}}_{t+k}]_1 \\ [\bar{\mathbf{x}}_{t+k}]_3 \end{bmatrix}. \tag{3.4}$$

where $\boldsymbol{\mu}_{t+k}$ is the predicted mean position vector at future step $t+k$, $\bar{\mathbf{x}}_{t+k}$ is the predicted mean state obtained from the CV model, and $[\bar{\mathbf{x}}_{t+k}]_1$ and $[\bar{\mathbf{x}}_{t+k}]_3$ denote its longitudinal and lateral position components, respectively.

To convert this deterministic prediction into a probabilistic belief, the model propagates the full position–velocity covariance using the same system matrix and process-noise covariance given in Equation 2.12:

$$\mathbf{P}_{k+1} = \mathbf{F}_{\mathrm{2D}} \mathbf{P}_k \mathbf{F}_{\mathrm{2D}}^{\mathsf{T}} + \mathbf{Q}_{\mathrm{2D}}. \tag{3.5}$$

Here, $\mathbf{P}_k$ is the propagated covariance of the state $[x, v_x, y, v_y]^{\mathsf{T}}$. The process-noise parameters q_x and q_y determine how quickly the uncertainty grows over the prediction horizon.

Only the position components of the propagated covariance matrix are used in the final belief representation. Therefore, the marginal position standard deviations are obtained as

$$\sigma_x(t+k) = \sqrt{[\mathbf{P}_k]_{1,1}}, \qquad \sigma_y(t+k) = \sqrt{[\mathbf{P}_k]_{3,3}}. \tag{3.6}$$

Each future prediction step is then represented as a two-dimensional Gaussian belief:

$$\mathcal{B}_{t+k} = \mathcal{N}\left(\boldsymbol{\mu}_{t+k}, \begin{bmatrix} \sigma_x^2(t+k) & 0 \\ 0 & \sigma_y^2(t+k) \end{bmatrix}\right). \tag{3.7}$$

In this representation, the mean vector describes the most expected future relative position of the surrounding vehicle, while the covariance matrix describes the uncertainty around this prediction. The exported belief uses a diagonal covariance matrix, retaining only the marginal uncertainty in the x_{forward} and y_{left} directions.

The belief generation stage therefore produces a sequence of Gaussian distributions over future relative positions, rather than a risk label. For each prediction step, the generated belief is stored together with the corresponding observed future position when it is available. In the following surprise computation stage, the observed position is evaluated under the predicted belief distribution. If the observation lies in a low-probability region of the belief, the motion is interpreted as more unexpected. In this way, the belief generation model provides the probabilistic reference distribution required for computing surprise-based safety metrics.

3.3.3 Estimation of Process Noise and Initial Covariance

The covariance propagation described above requires the uncertainty parameters, including the process-noise parameters q_x and q_y and the initial covariance P_0. These parameters determine the spread of the Gaussian belief distribution. The process-noise parameters control how fast the prediction uncertainty increases over the prediction horizon, while the initial covariance describes the uncertainty of the current position-velocity state at the start of prediction.

In this thesis, the initial covariance is represented separately for the longitudinal and lateral axes as $P_{x,0}$ and $P_{y,0}$, respectively. Here, $P_{x,0}$ describes the initial uncertainty of the longitudinal position-velocity state $[x, v_x]^{\top}$, while $P_{y,0}$ describes the initial uncertainty of the lateral position-velocity state $[y, v_y]^{\top}$. Together with q_x and q_y, these matrices define the covariance sequence $P_{x,k}$ and $P_{y,k}$ used in the Gaussian belief generation.

The uncertainty parameters are estimated from trajectory segments that are close to constant-velocity motion. These near-constant-velocity segments are used because their remaining deviations from ideal constant-velocity motion can be interpreted as the natural uncertainty of the kinematic prediction model. A segment is selected as near-CV when the relative longitudinal acceleration, relative lateral acceleration, and lateral velocity remain small over a continuous sequence of frames. In the current implementation, the default selection conditions are

$$|a_x| \leq 0.50 \text{ m/s}^2, \qquad |a_y| \leq 0.35 \text{ m/s}^2, \qquad |v_y| \leq 0.50 \text{ m/s}. \tag{3.8}$$

with a minimum continuous run length of 15 frames.

These thresholds are used as calibration settings for extracting a near-constant-velocity subset, rather than as universal limits of normal driving. Their selection was guided by the highD sampling rate, typical vehicle motion ranges in previous studies, and the amount of data retained after filtering. Since highD is recorded at 25 Hz, one frame corresponds to 0.04 s [12]. The longitudinal acceleration threshold of 0.50 m/s^2 therefore corresponds to only 0.02 m/s velocity change per frame and at most 0.30 m/s over the required 15-frame interval. This value is lower than typical comfortable acceleration values used in car-following models, which are often around 0.8–1.5 m/s^2 [22]. It is therefore suitable for selecting segments with stable longitudinal motion.

The lateral thresholds were chosen to exclude clear lateral manoeuvres. The lateral acceleration limit was set lower than the longitudinal limit because highway lane-keeping motion is expected to have weaker lateral dynamics. Over 15 frames, 0.35 m/s^2 corresponds to a lateral velocity change of about 0.21 m/s, which still allows small within-lane corrections. The lateral velocity threshold of 0.50 m/s was used to exclude obvious lane-changing motion. Since reported lane-change durations are commonly around 5–6 s, a standard lane width of about 3.5 m implies an average lateral velocity of approximately 0.6 m/s during a lane change [20, 21]. Therefore, the selected lateral limits allow small natural fluctuations while excluding most clear lane-changing movements.

The process-noise parameters q_x and q_y are estimated from the velocity increments within the accepted near-CV segments. For two consecutive frames, the velocity increment is defined as

$$\Delta v_i = v_{i+1} - v_i. \tag{3.9}$$

Under an ideal constant-velocity assumption, this increment would be zero. In real trajectory data, small non-zero increments remain due to natural driving variations, measurement noise, and modelling errors. Therefore, the variance of these velocity increments is used as an empirical measure of motion uncertainty. The longitudinal and lateral process-noise parameters are computed separately as

$$q_x = \max\left(\frac{\mathrm{Var}(\Delta v_x)}{\Delta t}, 10^{-8}\right), \qquad q_y = \max\left(\frac{\mathrm{Var}(\Delta v_y)}{\Delta t}, 10^{-8}\right). \tag{3.10}$$

The division by Δt converts the discrete velocity-increment variance into the process-noise spectral density used in the covariance model.

The lower bound 10^{-8} is applied to avoid zero or degenerate process-noise values. This value is used only as a numerical floor and is not intended to affect the calibrated uncertainty parameters. In the calibration results, the estimated process-noise intensities were $q_x = 1.74 \times 10^{-3}$ and $q_y = 9.20 \times 10^{-4}$, both of which are several orders of magnitude larger than 10^{-8}. Therefore, the floor only serves as a safeguard against numerical degeneracy in exceptional cases.

The initial covariance P_0 is estimated from the fitting errors of short near-CV trajectory segments. For each sufficiently long near-CV segment, a constant-velocity trajectory is fitted to the observed motion. In one coordinate direction, this fitting can be written as

$$p_i \approx p_0 + v_0 t_i, \tag{3.11}$$

where p_i is the observed position at time t_i, and p_0 and v_0 are the fitted initial position and velocity. The fitted initial state is then compared with the observed initial state of the segment, producing initial-state residuals:

$$e_{p,0} = p_0^{\text{obs}} - \hat{p}_0, \qquad e_{v,0} = v_0^{\text{obs}} - \hat{v}_0. \tag{3.12}$$

By collecting these residuals over all accepted near-CV fitting segments, the initial position and velocity variances are estimated. The initial covariance matrices are then represented as diagonal covariance blocks:

$$P_{x,0} = \begin{bmatrix} \text{Var}(e_{x,0}) & 0 \\ 0 & \text{Var}(e_{v_x,0}) \end{bmatrix}, \qquad P_{y,0} = \begin{bmatrix} \text{Var}(e_{y,0}) & 0 \\ 0 & \text{Var}(e_{v_y,0}) \end{bmatrix}. \tag{3.13}$$

Here, $e_{x,0}$ and $e_{y,0}$ are the initial position residuals in the longitudinal and lateral directions, while $e_{v_x,0}$ and $e_{v_y,0}$ are the corresponding initial velocity residuals. A minimum segment length of 8 frames is used before a segment is included in the initial covariance estimation.

Since the highD dataset is recorded at 25 Hz, 8 frames correspond to approximately 0.32 s. This minimum length was used as a practical calibration setting to ensure that the initial residuals are estimated from a short but continuous motion segment, rather than from isolated frames. Very short segments may produce unstable residual estimates because they are more sensitive to frame-level noise and small local fluctuations. At the same time, the window should not be too long, because P_0 is intended to describe the initial uncertainty at the beginning of a prediction segment, not the accumulated uncertainty over a longer trajectory. Therefore, the 8-frame requirement provides a balance between numerical stability and retaining enough near-CV samples for covariance estimation.

The estimated q_x, q_y, $P_{x,0}$, and $P_{y,0}$ are then used as the inputs to the covariance propagation described above. As a result, the deterministic constant-velocity mean prediction is extended into a Gaussian belief distribution whose uncertainty is calibrated from near-constant-velocity driving data.

3.3.4 Main Parameter Settings

The main parameter settings of the kinematics-based belief generation model are summarized in Tables 3.2 and 3.3. Table 3.2 presents the deterministic trajectory

prediction settings, while Table 3.3 summarizes the Gaussian belief representation and the main uncertainty-calibration settings.

Table 3.2: Main parameter settings for the kinematics-based trajectory prediction model.

Parameter	Value	Description
Frame rate	25 Hz	Dataset sampling rate
Time step Δt	0.04 s	Frame interval
Prediction horizon H	125 frames	Number of prediction steps
Prediction duration	5.0 s	Look-ahead time
Forward range	0 to 200 m	Longitudinal selection window
Lateral range	-25 to 25 m	Lateral selection window
Motion assumption	Constant velocity	Fixed relative velocity
Mean output	$\mu_x(t+k), \mu_y(t+k)$	Predicted future relative position

It should be noted that the prediction horizon H used in the trajectory prediction model is conceptually different from the history window h used later in the surprise calculation. The horizon H defines the maximum future duration over which the deterministic trajectory and Gaussian belief are propagated. In this thesis, $H = 125$ frames, corresponding to 5.0 s at 25 Hz. By contrast, the history window h defines how far back in time the prior belief is generated for the comparison between prior and posterior beliefs, while the lookahead time z defines the future time point at which the predicted belief and the observed motion are compared. Therefore, H is a trajectory-prediction setting, whereas h and z are temporal settings used in the surprise-based evaluation.

Table 3.3: Main parameter settings for Gaussian belief generation and NCV uncertainty calibration.

Parameter	Value	Description
Belief representation	Diagonal Gaussian	Future position belief
Belief mean	CV prediction	Predicted mean position
Exported uncertainty	$\sigma_x(t+k), \sigma_y(t+k)$	Position uncertainty
Exported covariance	Diagonal	Marginal covariance only
Internal covariance state	Position–velocity	Internal propagation state
Process-noise parameters	q_x, q_y	Uncertainty growth
Initial covariance	$P_{x,0}, P_{y,0}$	Initial state uncertainty
Longitudinal acceleration threshold	$\lvert a_x \rvert \leq 0.50$	Near-CV selection
Lateral acceleration threshold	$\lvert a_y \rvert \leq 0.35$	Near-CV selection
Lateral velocity threshold	$\lvert v_y \rvert \leq 0.50$	Near-CV selection
Minimum near-CV run length	15 frames	Process-noise samples
Minimum fitting length	8 frames	Initial covariance fit
Numerical floor for q	10^{-8}	Non-zero process noise
Numerical floor for P_0	10^{-8}	Non-zero covariance
Estimation scope	Per recording	Recording-wise calibration

The prediction horizon and forward selection range were chosen to correspond with the temporal and spatial scales of highway cut-in interactions. The prediction horizon is defined as $H = 125$ frames. The highD dataset is recorded at 25 Hz, resulting in a 5.0 s look-ahead window. This duration is appropriate for cut-in analysis because lane-changing behaviour is not an instantaneous occurrence. Previous research has shown that lane-change durations are typically around 5–6 s [20]. As a result, a 5-second prediction horizon is long enough to cover the majority of a typical lane-change or cut-in manoeuvre while remaining within a short-term prediction range.

The forward selection range is set to 0 to 200 m, covering the safety-relevant region ahead of the ego vehicle within this prediction horizon. The highD dataset includes naturalistic vehicle trajectories captured on German highways [12], where vehicles can travel at high speeds. According to German motorway design guidelines [23], the recommended speed for long-distance motorways is 130 km/h. This speed is approximately 36.1 m/s, implying that a vehicle can travel around 180 m in 5 s. A forward range of 200 m is sufficient to cover vehicles that may become relevant within the 5-second prediction window, while also leaving a small spatial buffer.

The belief generation settings extend the deterministic prediction into a probabilistic representation. The predicted future position is used as the Gaussian mean, while the marginal standard deviations $\sigma_x(t+k)$ and $\sigma_y(t+k)$ describe the uncertainty in the longitudinal and lateral directions. The detailed estimation of q_x, q_y, and P_0 has been described in Section 3.3.3. Therefore, this subsection only summarizes the final parameter settings used to generate the Gaussian belief sequence.

3.3.5 Prediction Output

This stage produces a predicted future path for each selected surrounding vehicle. This path represents the expected motion based on the constant-velocity assumption.

To visualise the prediction result, compare the predicted path to the observed future trajectory from the highD dataset. If the vehicle moves steadily, the two trajectories are similar. If the vehicle accelerates, brakes, or changes lanes, the observed trajectory deviates from the predicted path.

This deviation indicates how different the actual motion is from the expected motion. In the subsequent belief generation step, the predicted path serves as the probabilistic belief's mean trajectory.

Figure 3.3 illustrates an example of the belief generation result in the ego-centered coordinate system. The constant-velocity model provides the predicted mean trajectory of the surrounding vehicle, while the Gaussian ellipses represent the associated uncertainty of the future position at successive prediction steps. The observed future trajectory is plotted together with the predicted belief sequence for comparison. In this way, the figure shows not only the expected future motion of the surrounding vehicle, but also how the uncertainty evolves along the prediction horizon. When the actual motion deviates from the constant-velocity assumption, the difference between the observed trajectory and the predicted belief becomes more apparent.

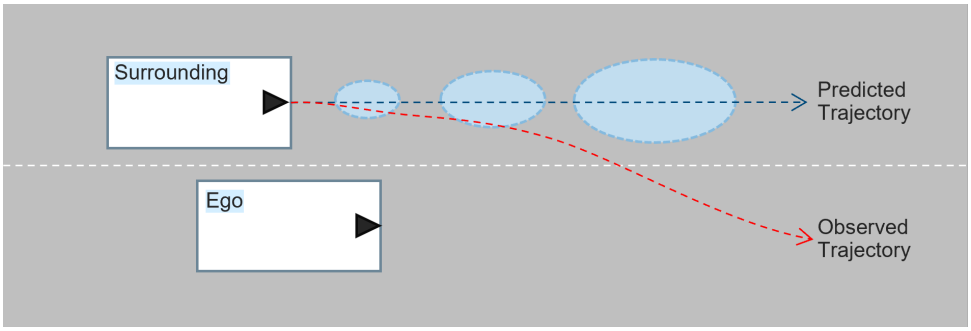

Figure 3.3: Example output of the kinematics-based trajectory prediction and belief generation model

3.4 Scenario Identification Based on the THW-Based Classical Baseline

3.4.1 THW-Based Classical Metric Used in This Study

In this study, THW is retained as the classical surrogate safety metric in the cut-in detection pipeline. Its general definition and interpretation have already been introduced in Equation 2.1, so this subsection only describes how the metric is implemented for event screening.

For each detected lane-change event, THW is computed between the cut-in vehicle and its post-change follower over the valid post-change interval, where both vehicles are simultaneously observed in the trajectory data. Following the THW definition in Equation 2.1, the distance term d is implemented as the net bumper-to-bumper longitudinal gap between the two vehicles, and the speed term is the longitudinal speed of the following vehicle.

THW is evaluated frame by frame in the post-change segment and then summarized at the event level using the minimum finite post-change value, denoted as $\mathrm{THW}_{\mathrm{min}}^{\mathrm{post}}$. This value represents the smallest temporal spacing observed after the cut-in and is used as the classical event-level severity indicator. To avoid unstable or physically unclear values, THW is only evaluated when the follower speed is above the predefined minimum speed threshold. Frames with non-positive gaps or invalid speed values are excluded from the event-level summary.

A cut-in event is accepted by the THW-based classical branch when $\mathrm{THW}_{\mathrm{min}}^{\mathrm{post}} \leq 3.0\,\mathrm{s}$. This threshold is used as an operational criterion for identifying short temporal gaps after lane changes. Therefore, in the present methodology, THW is not used as a general traffic-flow descriptor, but as a frame-level metric that is aggregated into an event-level indicator for THW-based classical cut-in detection.

3.4.2 Scenario Screening and Triggering Strategy

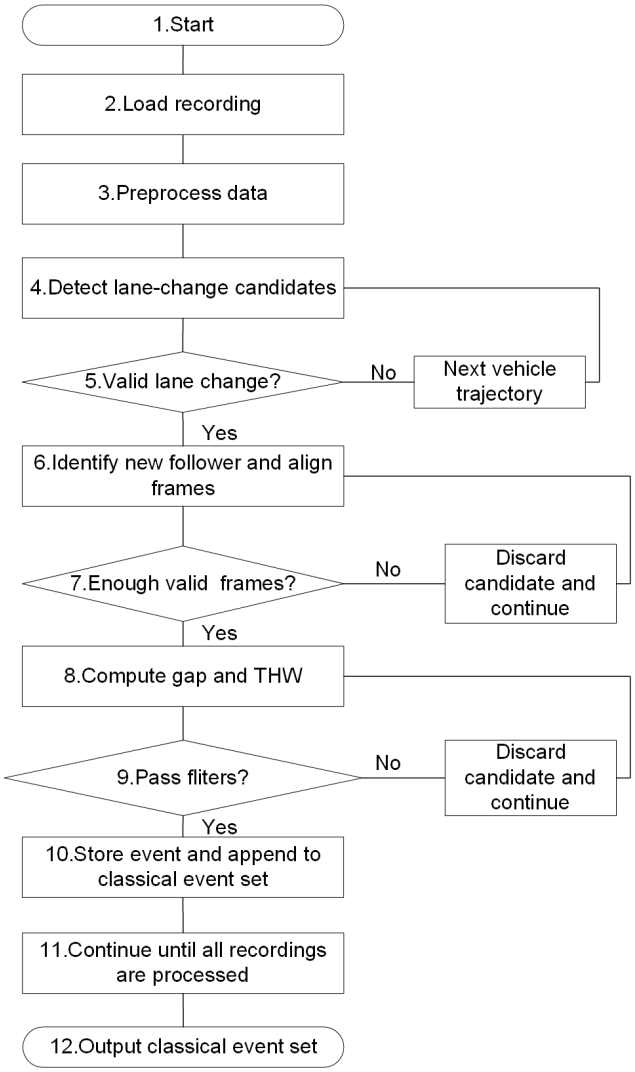

Figure 3.4: THW-based classical cut-in scenario screening workflow.

Figure 3.4 depicts the workflow of the THW-based classical cut-in scenario screening process. The procedure begins by loading each highD recording and preprocessing the trajectory data, which includes sorting vehicle trajectories and defining an analysis window centred on potential lane-change moments. Lane-change candidates are then identified using changes in lane ID for each vehicle trajectory.

For each valid lane-change candidate, the algorithm identifies the new following vehicle and uses common frame indices to align the lane-changing vehicle and the follower's trajectories. Candidates with insufficient valid post-change frames are rejected. The remaining candidates' bumper-to-bumper gap and THW are calculated during the post-change window. THW is calculated using the longitudinal gap and speed of the following vehicle.

The final triggering decision is made using the THW-related filter in conjunction

with additional geometric and kinematic constraints. A candidate is accepted only if the minimum post-change THW is less than the predefined threshold and the gap and relative-speed conditions indicate a significant close-following interaction. Accepted candidates are stored as THW-based classical cut-in events and appended to the THW-based event set E_T. This event set is then used to compare with the surprise-based event set.

3.4.3 Filter Settings

The filter settings used in the THW-based classical cut-in screening workflow are summarized in Table 3.4. In this implementation, THW is used as the classical surrogate safety metric for triggering cut-in events. This is a methodological choice made to obtain an inclusive set of close-following cut-in interactions, rather than only crash-imminent events. TTC and DRAC are relevant classical surrogate safety metrics introduced in the theory chapter, but they are not used as event-triggering criteria or in the final precision–recall comparison. The THW-based trigger provides a practical basis for identifying relevant cut-in interactions for comparison with the surprise-based metrics. Other conditions, such as follower existence, common-frame availability, gap constraints, and relative-speed constraints, are used as quality filters to ensure that each detected lane-change candidate represents a meaningful close-following interaction.

Table 3.4: Main filter settings for THW-based classical cut-in screening.

Filter	Setting
Lane-change candidate	`laneId` change
New follower	Required
Analysis window	-2 s to $+4$ s around lane change
Valid post-change frames	≥ 3 frames
Minimum follower speed	1.0 m/s
THW trigger	$\min(\mathrm{THW}_{post}) \leq 3.0$ s
Minimum post-change gap	0.5 m
Maximum change-frame gap	150.0 m
Relative-speed condition	$\Delta v < 0$

The filters in Table 3.4 serve two purposes: candidate construction and event triggering. The lane-change filter identifies possible cut-in candidates based on changes in `laneId`. This aligns with the highD dataset's structure, which includes lane-level vehicle trajectory data and a large number of recorded lane changes for highway scenario analysis [12]. After identifying a lane-change candidate, the new-follower condition is used to ensure that the lane change results in a new car-following interaction. Without this condition, ordinary lane changes that do not affect the next vehicle could be included.

The analysis window between 2 s before and 4 s after the lane change captures both the pre-change context and the immediate post-change interaction. The post-change section of the window is particularly important because the following vehicle

is only affected after the cut-in vehicle has entered its lane. The requirement for at least three valid post-change frames ensures that THW and gap values are not calculated using a single frame. Because the highD data is sampled at 25 Hz, three frames correspond to a short but non-zero time interval, which primarily serves as a data-validity check rather than a safety threshold.

The primary triggering metric in this THW-based classical screening workflow is THW. THW is commonly used as a surrogate safety indicator for car-following risk because it directly correlates the distance between two vehicles with the speed of the following vehicle. Previous traffic safety studies have discussed headway-based indicators alongside other surrogate measures such as TTC. Recommended following-time rules are often expressed in seconds, such as the commonly used two- or three-second range [6, 24]. Hence, the threshold $\min(\mathrm{THW}_{post}) \leq 3.0$ s is used as a relatively inclusive criterion to identify potentially relevant close-following cut-in events, rather than a strict crash-imminent threshold. It should be noted that some countries, such as Sweden, recommend at least 3s THW, while in other countries the recommendations are shorter (e.g., Germany recommending at least X s THW). These recommendations clearly shows the inclusiveness of the THW threshold used.

The final filters are quality and consistency constraints. The minimum follower speed prevents unstable THW values when the denominator approaches zero. The minimum post-change gap eliminates physically invalid or overlapping cases, whereas the maximum change-frame gap eliminates interactions that are too far apart to be meaningful. Finally, the relative-speed condition $\Delta v < 0$ requires the cut-in vehicle to be slower than the following vehicle during the lane-change. This ensures that the follower approaches from behind, which corresponds to the intended close-following cut-in scenario.

3.5 Scenario Identification Based on Surprise-Based Metrics

3.5.1 Surprise Metrics Used in This Study

This study employs four surprise metrics to describe unexpected behaviour in the predicted belief of nearby vehicle motion: Shannon Surprisal, Residual Information, Bayesian Surprise, and Antithesis. Their theoretical definitions have already been introduced in Equations 2.4, 2.7, 2.8, and 2.9. Therefore, this subsection only describes how these metrics are used in the surprise-based screening pipeline.

In the pipeline, one value of each surprise metric is calculated for every sampled frame of a vehicle interaction and linked to the corresponding recording, ego vehicle, and frame index.

Since the highD data are sampled at 25 Hz, this means that each metric produces one value every 0.04 s. For example, for one ego–surrounding-vehicle interaction, the sequence of Residual Information, Bayesian Surprise, or Antithesis values over consecutive frames shows how the surprise response changes before, during, and after

a cut-in. These frame-level sequences form the foundation for subsequent event-level summaries in the surprise-based scenario identification process. The detailed screening strategy and triggering conditions based on these metrics are described in the subsections below.

3.5.2 Scenario Screening and Triggering Strategy

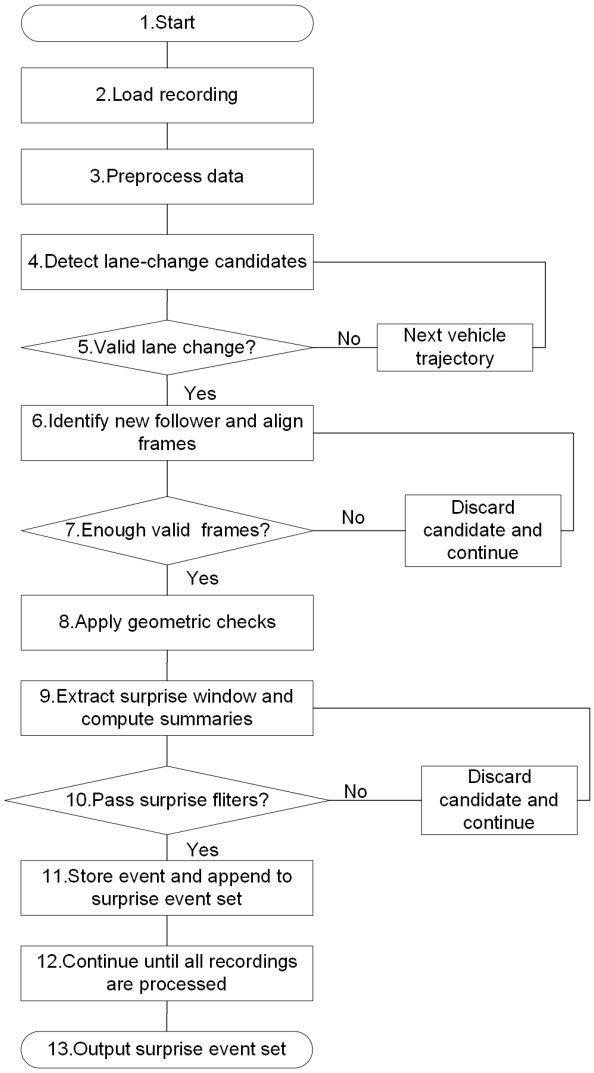

Figure 3.5: Surprise-based cut-in scenario screening and triggering workflow.

Figure 3.5 depicts the workflow of the surprise-based cut-in scenario screening and triggering process. The procedure begins by loading each highD recording together with the corresponding frame-level surprise data and baseline percentile thresholds. After preprocessing the trajectory data, lane-change candidates are identified from changes in lane ID for each vehicle trajectory.

For each valid lane-change candidate, the algorithm identifies the new following vehicle and aligns the trajectories of the lane-changing vehicle and the follower using

common frame indices. Candidates with insufficient valid post-change frames are rejected. Basic geometric checks are then applied to ensure that the candidate represents a meaningful interaction before the surprise-based evaluation is performed.

For the remaining candidates, the corresponding surprise window is extracted from the frame-level surprise time series. The four surprise metrics, including Shannon Surprisal, Residual Information, Bayesian Surprise, and Antithesis, are summarized within the analysis window using event-level values such as peak and mean responses. These summaries describe the intensity of the expectation violation during the candidate cut-in event.

The final triggering decision is based on the surprise metrics and their baseline percentile thresholds. In the current implementation, Residual Information is used as the main surprise trigger, while Bayesian Surprise and Antithesis are used as auxiliary conditions. Shannon Surprisal is retained as a summarized metric but is not used as the primary trigger. Candidates satisfying the surprise-based triggering rule are stored as surprise-based cut-in events and appended to the surprise event set E_S. This event set is later compared with the THW-based classical event set E_T.

The temporal settings used in the main surprise-based detection pipeline were $h = 1.0$ s and $z = 0.2$ s. Here, the history window h defines how far back the prior belief is generated, while the lookahead time z defines the future time point at which the predicted belief and observed motion are compared. These settings were used for the main surprise-based event detection and for the comparison with the THW-based classical method. To examine the sensitivity of the results, additional analyses were conducted with $h = 0.5$ s, 1.0 s, and 2.0 s under a fixed $z = 0.2$ s, and with $z = 0.2$ s, 1.0 s, and 2.0 s under a fixed $h = 1.0$ s.

3.5.3 Filter Settings

The main filter settings used in the surprise-based cut-in screening pipeline are summarized in Table 3.5. The pipeline uses a combination of geometric candidate filters and data-derived surprise thresholds. The geometric filters are used to ensure that the selected candidates represent meaningful cut-in interactions, while the surprise thresholds determine whether the candidate shows an abnormal expectation violation compared with normal driving. In the trigger notation, A denotes the Antithesis value and $p95$ denotes the 95th percentile of the corresponding normal-driving baseline.

Table 3.5: Main filter settings for surprise-based cut-in screening.

Filter	Setting
Lane-change candidate	`laneId` change
New follower	Required
Analysis window	-2 s to $+4$ s around lane change
Valid post-change frames	≥ 3 common frames
Minimum gap	0.5 m
Maximum insertion gap	100.0 m
Minimum follower speed	1.0 m/s for THW calculation
Surprise data alignment	`recording_id`, `ego_id`, and frame
Baseline thresholds	Percentiles from normal-driving baseline
Active trigger	$RI > RI_{p95}$ and ($BS > BS_{p95}$ or $A > A_{p95}$)
Post-trigger check	Gap < 150.0 m, $\Delta v < 0$

The filter settings in Table 3.5 are chosen to balance valid cut-in candidate construction with data-driven abnormality detection. The lane-change and new-follower filters are used because the highD dataset provides lane-level trajectory information and naturalistic highway lane-change scenarios, making lane-ID transitions and new following relationships suitable indicators for cut-in candidate extraction [12]. The gap, follower-speed, and relative-speed constraints serve as quality filters, removing invalid or weakly interacting cases. These settings align with the car-following interpretation of headway-based safety indicators, which emphasises the longitudinal gap and speed of the following vehicle in assessing interaction risk [6].

The surprise thresholds are derived from the normal-driving baseline, rather than being manually selected constants. Stable car-following baseline frames are matched with frame-level surprise outputs using `recording_id`, `ego_id`, and frame. The resulting normal-driving surprise distribution is used to compute percentile thresholds. This follows the idea of defining abnormality in relation to a reference distribution [25]. This study uses $p95$ as an inclusive threshold for unusually high surprise values. A candidate is accepted into E_S if $RI > RI_{p95}$ and ($BS > BS_{p95}$ or $A > A_{p95}$).

3.5.4 Normal-Driving Baseline and Data-Driven Threshold Calibration

After calculating the frame-level surprise metrics, a normal-driving baseline is constructed to determine the thresholds for RI, BS, and Antithesis. The same baseline construction procedure is used for all three metrics, so that their thresholds are derived from a common normal-driving reference set.

Stable car-following frames are first selected from normal driving records. A frame is retained only when the ego vehicle remains in the same lane, follows the same preceding vehicle within the selected time window, and has a valid preceding vehicle located in front of it. In addition, the gap and THW values are required to stay within normal car-following ranges. Frames with sudden changes in gap or THW are removed, and repeated frame samples caused by overlapping windows are excluded

to avoid double counting. The numerical filtering settings used in this step are summarized in Table 3.6.

Table 3.6: Numerical filtering settings for constructing the normal-driving baseline.

Filtering aspect	Value used in this study
Stability window	100 frames, corresponding to 4.0 s at 25 Hz.
Sliding step size	50 frames, corresponding to 2.0 s at 25 Hz.
Lane and leader stability	Constant lane; constant non-zero preceding vehicle.
Valid car-following relation	Preceding vehicle in front of the ego vehicle; positive net gap.
Ego speed	$\geq 3.0 m/s$.
Ego longitudinal acceleration	$\lvert a_x \rvert \leq 1.5 m/s^2$.
Ego lateral velocity	$\lvert v_y \rvert \leq 0.3 m/s$.
Leader speed	$\geq 3.0 m/s$.
THW range	$2.5s \leq \text{THW} \leq 6.0s$.
Frame-to-frame gap continuity	Gap jump $\leq 2.0m$.
Frame-to-frame THW continuity	THW jump $\leq 0.3s$.
Duplicate removal	One sample is kept for each ego vehicle and frame.

The selected baseline frames are then matched with the corresponding frame-level surprise results. This step links each normal-driving frame with its residual information, Bayesian surprise, and antithesis values. If a metric value is missing or invalid for a certain frame, that frame is not used for the threshold calculation of that metric.

The threshold is calculated separately for each surprise metric. For residual information, all valid residual information values from the normal-driving baseline are collected, and the 95th percentile of this distribution is used as the residual information threshold. The same procedure is applied to Bayesian surprise and antithesis. Therefore, although the three metrics share the same normal-driving baseline frames, each metric has its own threshold because their numerical ranges and distributions are different.

In this study, a metric value above its corresponding 95th percentile threshold means that the response is unusually high compared with normal driving. This exceedance is used only as a screening condition. It does not directly mean that the frame or event is risky. The final risky cut-in detection is still determined by the event-level screening rules introduced in the following section.

4

Results

This chapter presents the results of the proposed evaluation framework. It first reports the surprise-metric thresholds obtained from the normal-driving baseline. It then compares the risky cut-in events detected by the THW-based baseline and the surprise-based method, including their overlap and event-level differences. The chapter further evaluates detection performance using precision–recall analysis and examines the sensitivity of the surprise-based metrics to the history window h and lookahead time z. These results provide the basis for assessing whether surprise-based metrics offer complementary information to the THW-based baseline in highway cut-in scenarios.

4.1 Threshold Values of Surprise-Based Metrics from the Normal-Driving Baseline

The threshold values for the surprise-based metrics were calculated using the normal-driving baseline. Rather than using manually selected fixed thresholds, this study defines abnormal surprise responses relative to the distribution observed during normal driving. This is important because the numerical values of surprise-based metrics depend on the prediction model, uncertainty settings, and selected temporal parameters.

Figures 4.1–4.3 show the distributions of Residual Information, Bayesian Surprise, and Antithesis calculated from the normal-driving baseline frames. In this context, a frame-level baseline sample refers to one valid normal-driving frame after matching the normal-driving baseline with the corresponding surprise-metric output. The dashed vertical line marks the corresponding $p95$ threshold for each metric.

The $p95$ threshold is adopted as the operational screening threshold for each surprise-based metric. Compared with lower percentiles such as $p75$ and $p90$, $p95$ focuses on clearly higher surprise responses. Compared with a more extreme threshold such as $p99$, it is less restrictive and does not limit the screening only to rare outliers. Thus, the $p95$ threshold provides a practical balance between sensitivity and strictness. In the subsequent risky cut-in detection stage, these thresholds are used to identify candidate events whose Residual Information, Bayesian Surprise, or Antithesis responses are unusually high compared with the normal-driving baseline.

Samples to the right of this line are located in the upper tail of the normal-driving

baseline distribution. They are therefore treated as unusually high surprise responses compared with typical highway interactions.

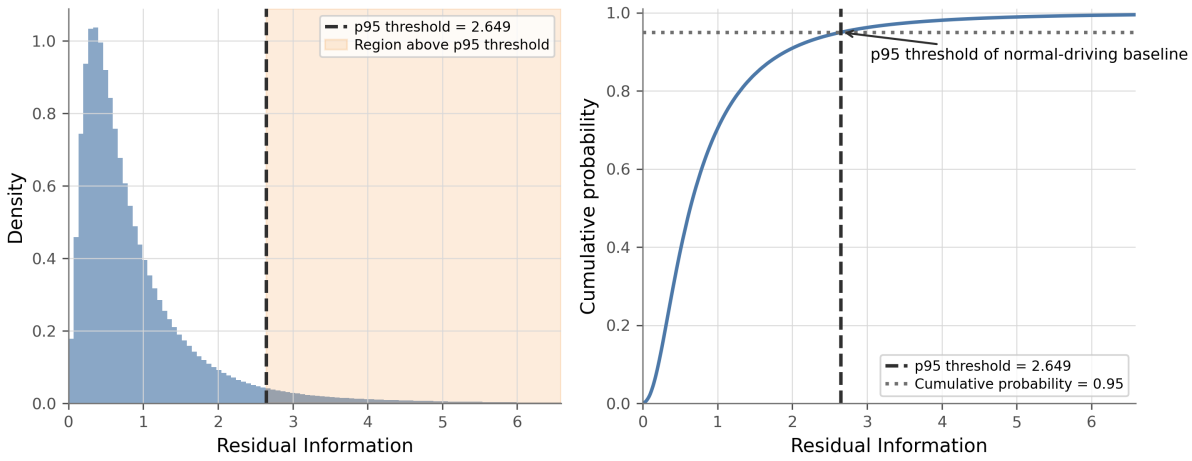

Figure 4.1: Normal-driving Residual Information distribution with the $p95$ threshold.

Figure 4.1 shows the distribution of Residual Information values in the normal-driving baseline. The baseline contains 6,671,521 valid Residual Information samples from 60 recordings. Most Residual Information values are small: 50% of the samples are below 0.634, 75% are below 1.126, and 90% are below 1.905. This indicates that, in most normal-driving frames, the observed vehicle position is close to the position predicted by the belief model. However, the distribution also has a long upper tail. The maximum Residual Information value reaches 185.828, although such large values are rare. The $p95$ threshold is 2.64, meaning that 95% of normal-driving samples have Residual Information values below this value. In total, 333,576 samples (per definition 5.00% of the total number of samples) of the Residual Information baseline, lie above the $p95$ threshold.

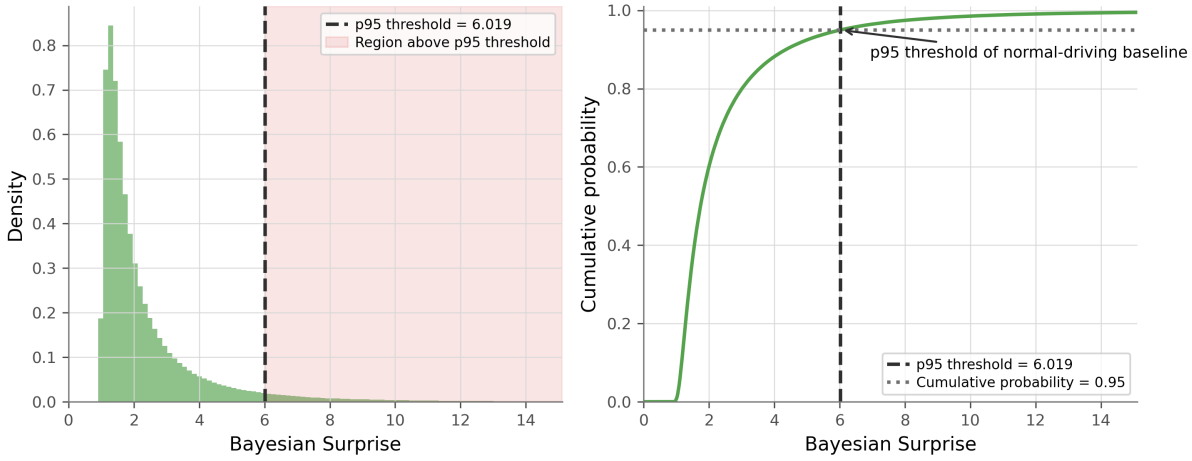

Figure 4.2: Normal-driving Bayesian Surprise distribution with the $p95$ threshold.

Figure 4.2 presents the normal-driving baseline distribution of Bayesian Surprise across 60 recordings, including 5,478,035 valid samples. This number differs from the Residual Information baseline because Bayesian Surprise requires a valid matched pair of prior and posterior Gaussian beliefs. Samples with missing aligned beliefs, invalid covariance values, or non-finite results are excluded.

The Bayesian Surprise values are mainly concentrated in the lower range, with a mean of 2.45 and a median of 1.738, while the maximum value reaches 187.304. The distribution is right-skewed: 75% of samples are below 2.646, 90% are below 4.352, and 95% are below the $p95$ threshold of 6.019. The dashed line marks this threshold, above which 273,902 samples are located (by definition, 5.00% of the valid Bayesian Surprise baseline samples).

Although the theoretical lower bound of Bayesian Surprise is zero, the empirical normal-driving distribution in this study does not start from zero. This is because Bayesian Surprise is computed as the KL divergence between a posterior belief generated at the current time and a prior belief generated at an earlier time but evaluated at the same future target time. Therefore, the prior corresponds to a longer prediction horizon than the posterior. Even under normal driving, the two Gaussian beliefs generally differ in both mean and covariance, especially because the prediction uncertainty increases with horizon length. As a result, the KL divergence is positive in most valid samples, and values close to zero are rarely observed in the empirical baseline distribution.

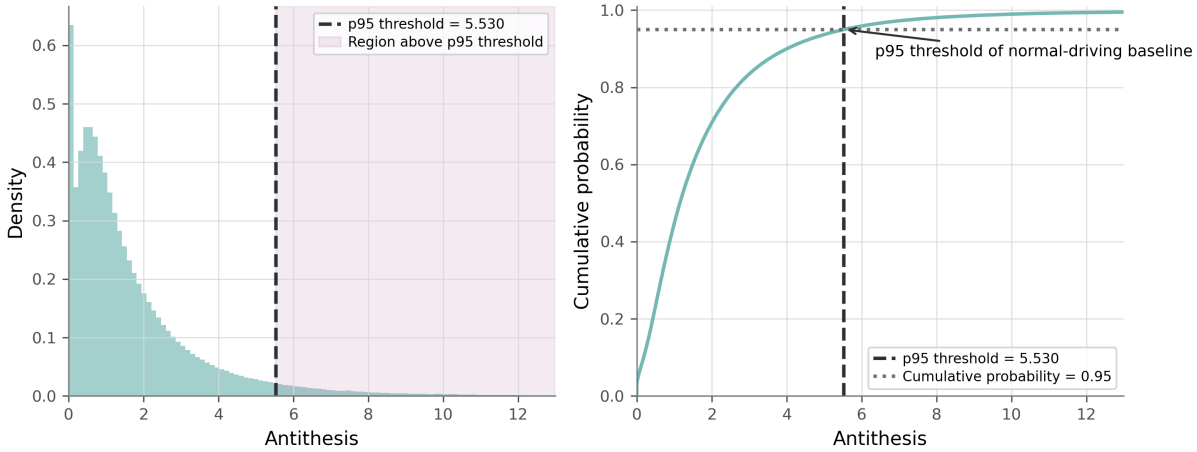

Figure 4.3: Normal-driving Antithesis distribution with the $p95$ threshold.

Figure 4.3 presents the normal-driving baseline distribution of Antithesis across 60 recordings, including 5,478,035 valid samples. The Antithesis values are mainly concentrated in the lower range, with a mean of 1.787 and a median of 1.150, while the maximum value reaches 29.966. The distribution is also right-skewed: 75% of samples are below 2.262, 90% are below 4.005, and 95% are below the $p95$ threshold of 5.530. The dashed line marks this threshold, above which 273,902 samples, or 5.00% of the Antithesis baseline, are located.

Overall, the three baseline distributions show a similar pattern: most normal-driving samples remain in a low-value region, while a small proportion forms an upper tail. This pattern is expected because normal highway interactions, especially stable car-following situations, are usually smooth and relatively predictable under the constant-velocity-based belief model. At the same time, high surprise values can still occur within the normal-driving baseline. These exceedances may be caused by mild acceleration, small lane-position adjustments, sensor noise, local interaction changes, or short-term deviations from the constant-velocity assumption. Therefore, values above the $p95$ threshold should be interpreted as unusually high relative to normal driving, rather than as inherently safety-critical events.

4.2 Overlap and Complementarity Between Classical and Surprise-Based Risky Cut-in Detection

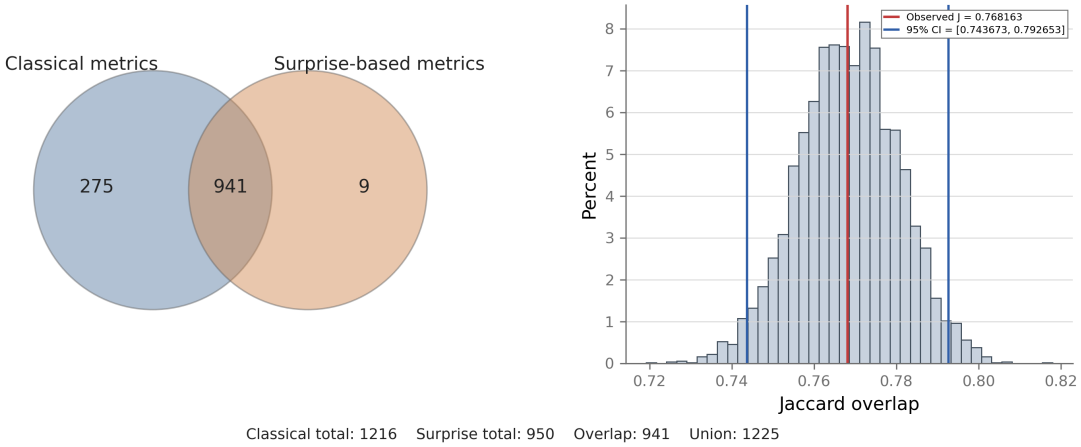

Figure 4.4: Overlap of risky cut-in events detected by the THW-based baseline and surprise-based method, together with the bootstrap distribution of the Jaccard overlap.

Figure 4.4 shows the overlap between the risky cut-in events detected by the THW-based baseline and the surprise-based method, together with the bootstrap distribution of the Jaccard overlap. In total, 1,225 unique detected events are obtained from the union of the two methods. Among them, 941 events are detected by both methods, accounting for 76.82% of the total. In addition, 275 events are detected only by the THW-based screening method, corresponding to 22.45%, while 9 events are detected only by the surprise-based method, corresponding to 0.73%.

The large intersection indicates that most detected risky cut-in events are identified by both methods. This suggests that many cut-in events contain both classical surrogate safety characteristics, such as short THW or critical relative motion, and abnormal surprise responses under the belief-model-based evaluation. However, the two

non-overlapping regions show that the methods are not identical. The classical-only events represent cases that satisfy the selected surrogate safety thresholds, whereas the surprise-only events represent cases where the observed vehicle behaviour is unexpected relative to the prediction model but does not satisfy the selected classical screening rule.

To quantify the degree of agreement between the two detected event sets, the Jaccard index is used, defined as the size of the intersection divided by the size of the union. Based on the Venn result, the observed Jaccard overlap is 0.768, meaning that approximately 76.8% of all unique detected events are shared by the two methods. The bootstrap distribution in Figure 4.4 shows the Jaccard overlap. The red vertical line marks the observed value, while the blue vertical lines indicate the 95% bootstrap confidence interval. Using 5,000 bootstrap samples, the bootstrap mean is 0.768, the median is 0.768, and the standard deviation is 0.012. The resulting 95% confidence interval is $[0.744, 0.793]$, indicating that the overlap estimate is stable under resampling.

Overall, the Jaccard overlap of about 0.77 indicates a high degree of agreement between the THW-based baseline and the surprise-based method. At the same time, the overlap is not complete. This is important because it shows that the two methods do not simply duplicate each other. Instead, they emphasize different aspects of risky cut-in behaviour. The THW-based baseline mainly reflects post-cut-in temporal proximity, while the surprise-based method captures deviations from the motion pattern expected by the belief model.

4.3 Precision-Recall Comparison Between Classical and Surprise-Based Detection

Before presenting the precision–recall results, a manually labelled validation set is used as the reference for evaluation. This set is an event-level subset sampled from highD trajectory candidates, including events detected by the classical method, events detected by the surprise-based method, and negative samples. Each selected event was inspected using trajectory visualizations and labeled as either safety-critical or non-critical. The labels were not used to generate either detection result, but only to evaluate how well each risk score separates risky cut-ins from non-critical cases. Therefore, this subset should be interpreted as a validation benchmark rather than as an estimate of the natural prevalence of safety-critical events.

Precision–recall analysis is then used to evaluate the metric-level detection performance against these manual labels [28]. Precision describes the proportion of detected events that are truly risky, while recall describes the proportion of manually labelled risky events that are successfully detected. By varying the decision threshold, the precision–recall curve shows how performance changes from strict selection to more inclusive detection. The average precision (AP) summarizes this curve into a single value, where a higher AP indicates stronger overall performance across different recall levels.

In this section, the THW-based risk score is used as the baseline and is compared separately with Residual Information, Bayesian surprise, and Antithesis. The purpose is not only to compare their AP values, but also to examine what type of information each metric captures. The THW-based score mainly reflects whether a cut-in creates a short longitudinal time gap in front of the ego vehicle. In contrast, the surprise-based metrics describe whether the observed motion of the surrounding vehicle deviates from the prediction-based belief.

4.3.1 Comparison with Residual Information

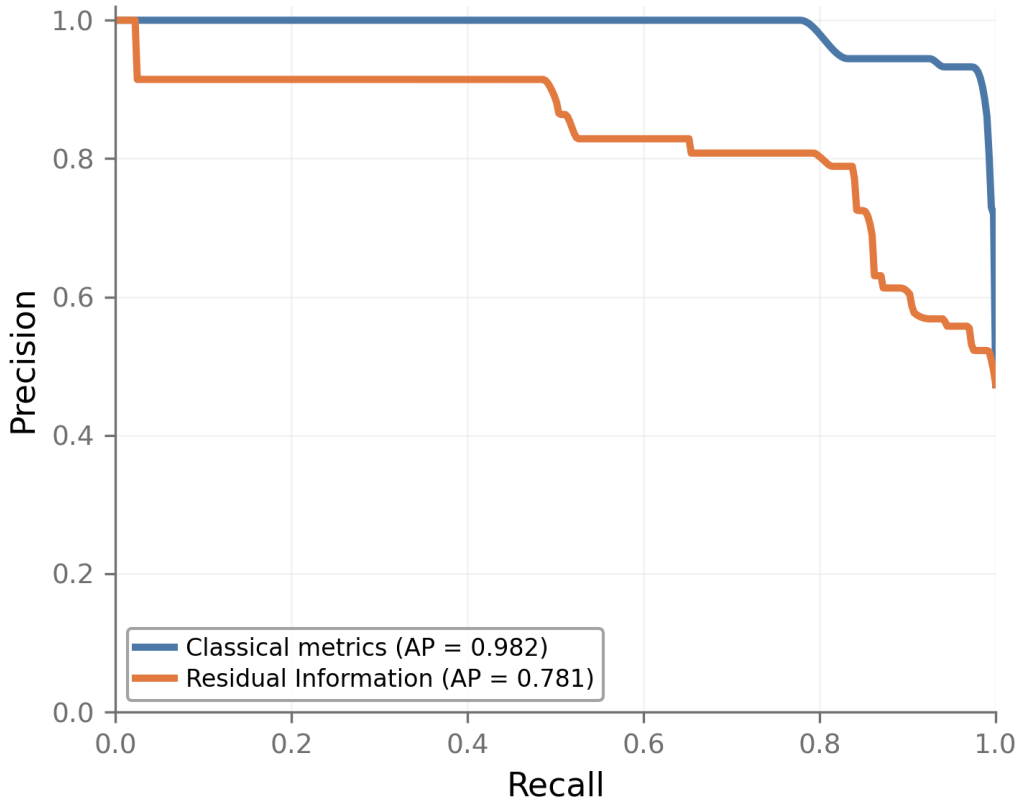

Figure 4.5: Precision–recall comparison between the THW-based risk score and Residual Information for risky cut-in detection.

Figure 4.5 compares the precision–recall performance of the THW-based risk score and Residual Information. The THW-based score achieves an AP of 0.982, while Residual Information achieves an AP of 0.781. The large difference between the two curves indicates that the manually labelled risky cut-in events are more strongly aligned with THW-based risk characteristics, especially reduced longitudinal spacing and short time headway after the cut-in.

The THW-based curve remains close to perfect precision over most of the recall range. This result is reasonable because many risky cut-ins become safety-critical when the surrounding vehicle enters the ego lane and reduces the available temporal

gap to the following vehicle. In such cases, THW directly captures the resulting temporal urgency by measuring how short the post-change headway becomes.

Residual Information shows weaker performance in comparison. Although it starts with perfect precision at very low recall, the curve quickly drops to around 0.91 precision and remains below the THW-based curve across almost the entire recall range. This means that some events with high Residual Information are not labelled as risky cut-ins. In practical terms, a large prediction–observation mismatch can indicate unexpected vehicle behaviour, but this unexpectedness does not necessarily imply that the interaction is safety-critical, not the least since the situations may not be urgent enough to warrent a critical event classification.

The decline of the Residual Information curve becomes more pronounced in the high-recall region. When the threshold is relaxed to recover more labelled risky events, more false positives are included, and precision drops substantially. This suggests that Residual Information is sensitive to deviations from the prediction model, but some of these deviations may correspond to smooth lane changes, non-threatening manoeuvres, or interactions with sufficient longitudinal distance.

4.3.2 Comparison with Bayesian Surprise

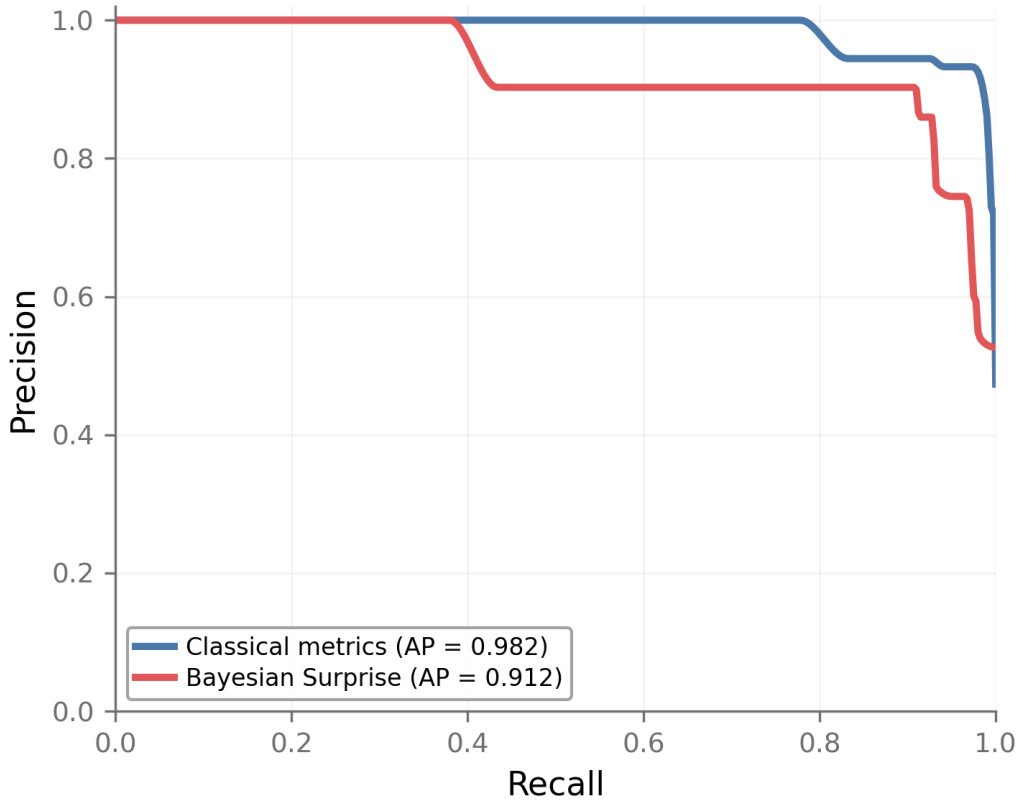

Figure 4.6: Precision–recall comparison between the THW-based risk score and Bayesian surprise for risky cut-in detection.

Figure 4.6 compares the precision–recall performance of the THW-based risk score and Bayesian surprise. The THW-based score achieves an AP of 0.982, while Bayesian surprise achieves an AP of 0.912. Bayesian surprise performs better than Residual Information, but it still remains below the THW-based score.

The Bayesian surprise curve shows good performance at low and moderate recall. It remains at perfect precision until about 0.39 recall, meaning that the events with the highest Bayesian surprise values are all consistent with the manual risky-event labels in this range. After this point, precision decreases to around 0.90 and remains relatively stable over a wide recall interval. This suggests that strong belief changes are often related to risky cut-in behaviour, but they are not as consistently aligned with the manual labels as the THW-based score.

The curve drops more clearly when recall becomes high. When the threshold is relaxed to recover more labelled risky events, additional events with large belief changes are included, but some of them are not labelled as risky. This indicates that a large update from prior belief to posterior belief does not always imply safety-criticality. A surrounding vehicle may cause a noticeable belief change because its motion differs from the expected prediction, while the actual interaction may still occur with sufficient distance or limited direct influence on the ego vehicle - that is, the urgency is low.

Therefore, Bayesian surprise does not outperform the THW-based score in this comparison. Its value lies in describing changes in the belief about surrounding vehicle behaviour, rather than directly measuring physical closeness or time gap. This makes Bayesian surprise a useful complementary indicator, but the result also shows that belief change alone is not sufficient for reliable risky cut-in detection, especially with respect to situational urgency.

4.3.3 Comparison with Antithesis

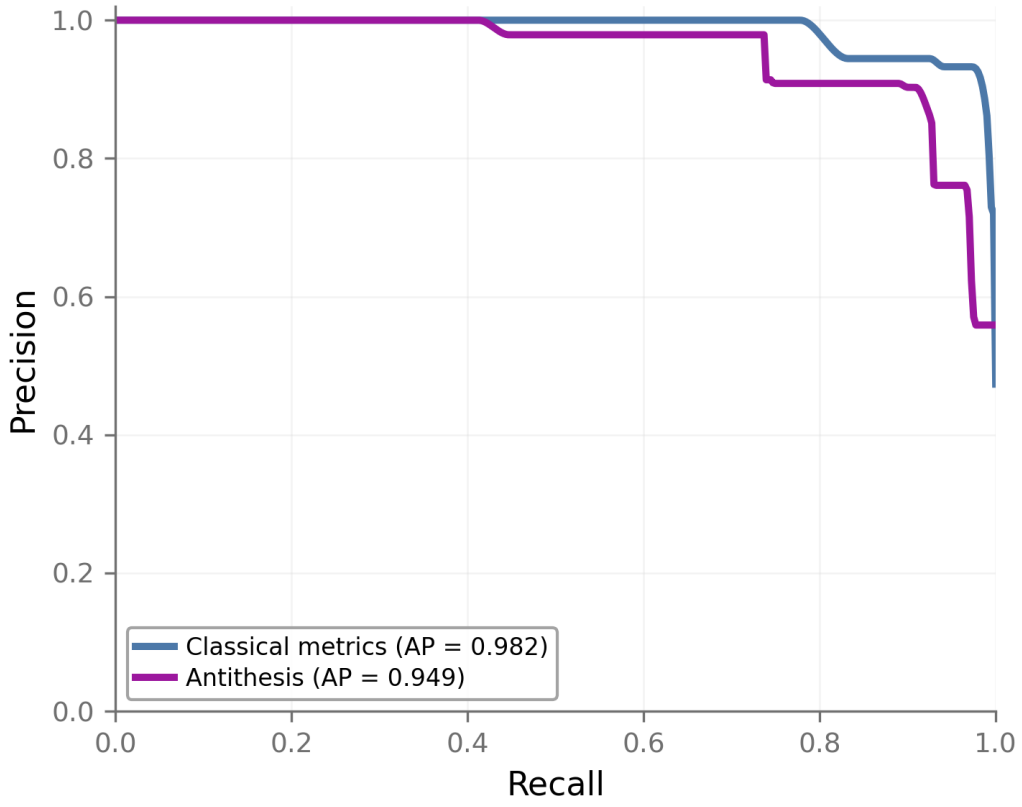

Figure 4.7: Precision–recall comparison between the THW-based risk score and Antithesis for risky cut-in detection.

Figure 4.7 compares the precision–recall performance of the THW-based risk score and Antithesis. The THW-based score achieves an AP of 0.982, while Antithesis achieves an AP of 0.949. Both values are high, indicating that both short-headway information and surprise-based information are useful for identifying risky cut-in events in the labelled subset. However, the higher AP of the THW-based score shows that the manually labelled risky cut-ins are still more strongly associated with conventional risk characteristics.

Antithesis shows strong performance at low and moderate recall. The curve remains at perfect precision until about 0.42 recall and stays close to 0.98 precision until around 0.74 recall. This means that the events with the highest Antithesis values are mostly consistent with the manual risky-event labels. In practical terms, a high Antithesis value often corresponds to surrounding-vehicle behaviour that strongly conflicts with the expected belief, which can be associated with unexpected cut-in manoeuvres.

However, the Antithesis curve drops more clearly in the high-recall region. When the threshold is relaxed to recover more labelled risky events, additional events with high Antithesis values are included, but not all of them are labelled as risky. This again

indicates that unexpected behaviour does not necessarily imply safety-criticality. A vehicle may behave differently from the prediction model, but the manoeuvre may still occur with sufficient distance, smooth lateral motion, or limited direct influence on the ego vehicle.

Therefore, Antithesis does not outperform the THW-based score in this comparison, but it provides the strongest performance among the three surprise-based metrics. Compared with Residual Information and Bayesian surprise, Antithesis appears to select unexpected changes that are more closely related to risky cut-in interactions. This suggests that Antithesis can help describe unexpected motion patterns beyond proximity-based risk alone.

Across the three precision–recall comparisons, the AP values are 0.982 for the THW-based score, 0.949 for Antithesis, 0.912 for Bayesian surprise, and 0.781 for Residual Information. Thus, Antithesis achieves the closest performance to the THW-based baseline among the tested surprise-based metrics.

4.4 Sensitivity of Surprise-Based Detection to h and z

The surprise-based detection results depend on the temporal settings used for belief comparison. The history window h controls how far back the prior belief is generated, while the lookahead time z defines the future point used for comparing prediction and observation. This section evaluates whether changing these parameters affects the precision–recall performance and average precision of the surprise-based metrics.

4.4.1 Effect of History Window h on Detection Performance

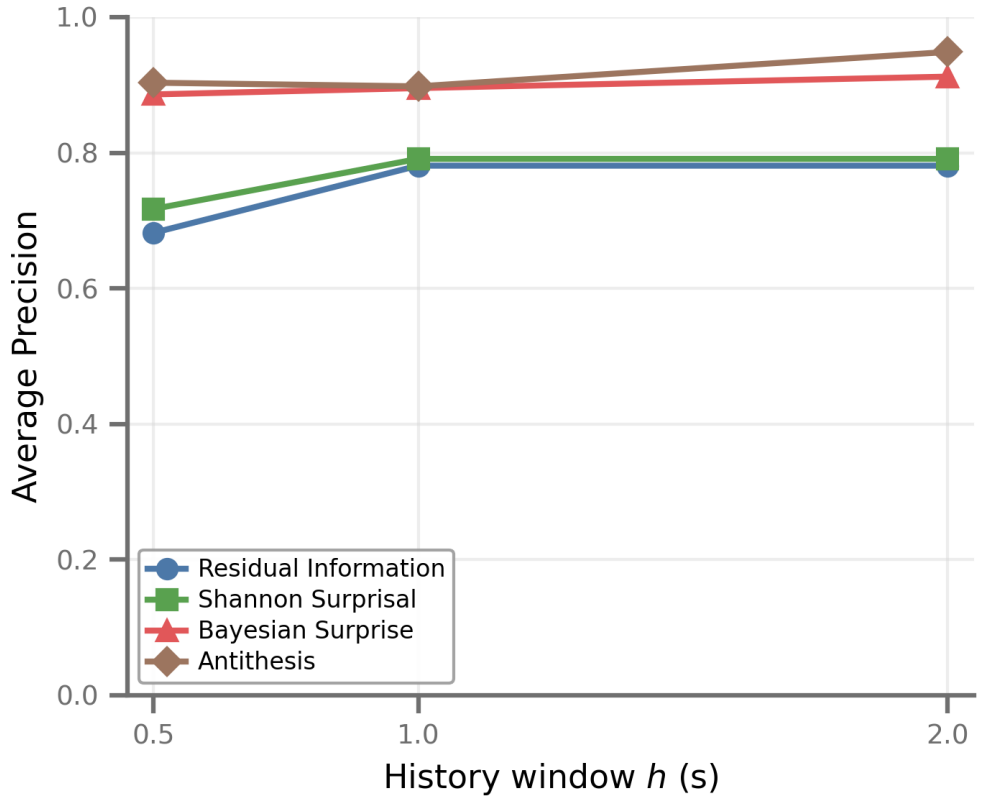

Figure 4.8: Effect of the history window h on the average precision of the surprise-based metrics with $z = 0.2$ s.

Figure 4.8 shows the sensitivity of the surprise-based detection performance to the history window h. The average precision values are compared under three history-window settings, $h = 0.5$ s, $h = 1.0$ s, and $h = 2.0$ s, while the lookahead time is fixed at $z = 0.2$ s. These values were chosen to cover short, medium, and longer prior-belief intervals within the surprise-based comparison framework. The use of a short lookahead time follows the idea in previous surprise-based traffic-safety studies, where surprise is evaluated over a near-future response interval rather than over a long-term trajectory forecast [10, 11].

Residual Information and Shannon Surprisal show the most visible sensitivity to h. Their average precision values are lower when $h = 0.5$ s, but increase when the history window is extended to 1.0 s. After $h = 1.0$ s, the improvement becomes very small, and the performance remains almost unchanged at $h = 2.0$ s. This indicates that a very short history window may provide insufficient temporal context for these two metrics, while a history window of 1.0 s is already sufficient to capture most of the useful information.

Bayesian Surprise is less sensitive to the history window. Its average precision remains high across all tested values of h, with only a slight increase as h becomes

longer. Antithesis also shows limited sensitivity, although it reaches its highest average precision at $h = 2.0$ s. This suggests that Antithesis may benefit slightly from a longer temporal context, but the improvement is still moderate.

These results show that increasing h from 0.5 s to 1.0 s can improve the performance of some surprise metrics, especially Residual Information and Shannon Surprisal. However, further increasing h to 2.0 s does not lead to a large additional improvement for most metrics. Therefore, the choice of h affects the results, but its influence is moderate.

4.4.2 Effect of Lookahead Time z on Detection Performance

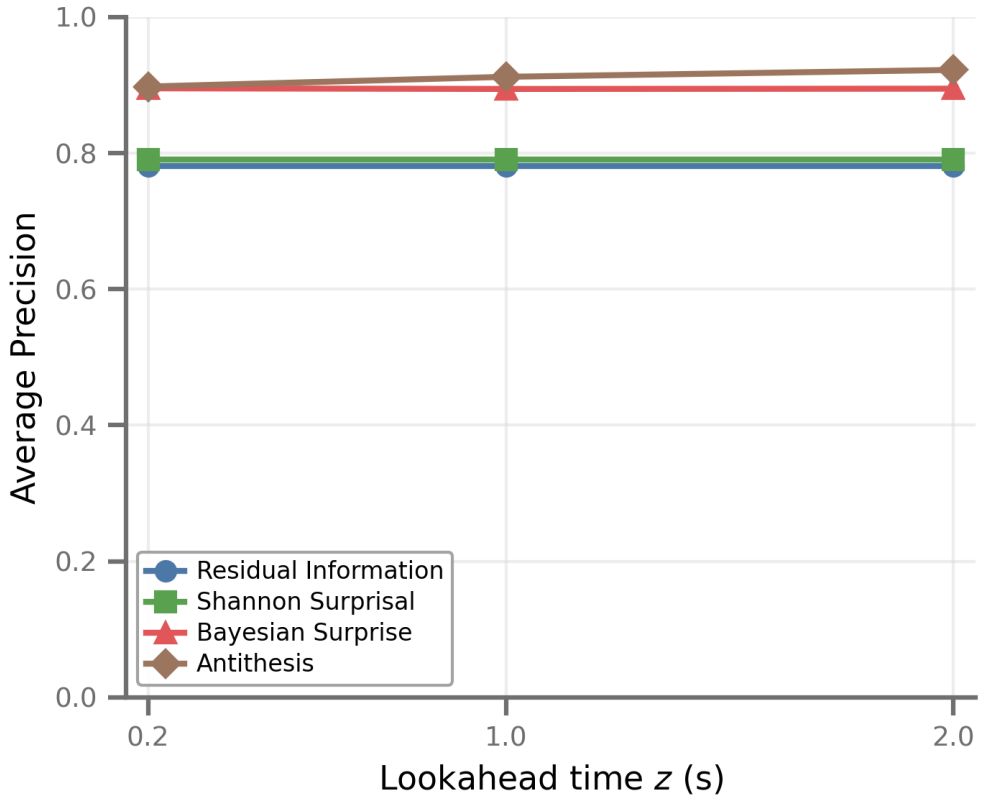

Figure 4.9: Effect of the lookahead time z on the average precision of the surprise-based metrics with $h = 1.0$ s.

Figure 4.9 shows the sensitivity of the surprise-based detection performance to the lookahead time z. The average precision values are compared under three lookahead settings, $z = 0.2$ s, $z = 1.0$ s, and $z = 2.0$ s, while the history window is fixed at $h = 1.0$ s. The tested values were chosen to cover short, medium, and longer prediction-observation comparison intervals. The shortest setting, $z = 0.2$ s, follows the near-term comparison used in previous surprise-based traffic-safety studies, whereas $z = 1.0$ s and $z = 2.0$ s are included to test whether the detection performance changes when the lookahead time is extended [10].

Residual Information and Shannon Surprisal remain almost unchanged across the three tested values of z. This behaviour is related to the definitions of these two metrics. Both metrics mainly evaluate how unlikely the observed vehicle state is under the predicted belief. Shannon Surprisal is based on the negative log-likelihood of the observation, while Residual Information measures the difference between the likelihood at the most likely state and the likelihood at the observed state. Therefore, they mainly describe the local prediction–observation mismatch. Under the current implementation, changing the lookahead time does not substantially change this mismatch-based ranking of events, so their average precision values remain nearly constant.

5

Discussion

This chapter interprets the results in relation to the two research questions. The main finding is that the THW-based classical score and the surprise-based metrics are related but not interchangeable. THW mainly reflects physical proximity and temporal urgency, while surprise-based metrics describe behavioural unexpectedness relative to the generated belief.

The chapter first discusses the main result patterns, including overlap, precision–recall performance, and temporal-parameter sensitivity. It then answers the research questions, summarizes the main limitations, and outlines future directions for improving belief generation through reachability constraints, comfort- and dread-zone concepts, and learning-based trajectory prediction.

5.1 Discussion of Results

5.1.1 Overlap and Complementarity Between Detection Methods

The overlap between the THW-based classical method and the surprise-based method suggests that the two approaches are related, but not equivalent, which is consistent with the broader view that surrogate safety and criticality indicators capture different dimensions of traffic risk [1, 6]. This is an important finding because it indicates that risky cut-in events cannot be fully described by a single type of safety metric. The THW-based score and the surprise-based metrics describe different aspects of the same traffic interaction. The former focuses on the longitudinal time gap, while the latter focus on prediction inconsistency.

The large overlapping region indicates that many detected cut-in events involve both aspects. In these cases, the interaction has a short longitudinal time gap according to the THW-based classical rule, and it also produces an abnormal response relative to the expected belief. This suggests that short-gap cut-ins and unexpected motion are often connected in risky cut-in scenarios. A vehicle that cuts in with a short THW may also deviate strongly from the motion predicted from its earlier state. Therefore, the agreement between the two methods is reasonable and supports the validity of both approaches for identifying many relevant cut-in interactions.

However, the non-overlapping cases are particularly important for understanding

complementarity. These cases show that the two methods do not simply detect the same phenomenon using different mathematical formulations. THW-only events represent situations where the minimum THW falls below the selected criticality threshold, but the vehicle motion may still be relatively consistent with the earlier belief [6, 10]. In such cases, the event can have a short longitudinal time gap without being highly unexpected from the prediction model's perspective. Conversely, surprise-only events represent situations where the observed motion differs strongly from the expected belief, even though the minimum THW remains above the selected THW threshold.

This distinction is central to the interpretation of the surprise-only events. These events should not be interpreted simply as critical events missed by the THW-based classical method. Their larger THW values indicate that they are not immediately urgent conflicts under the selected THW threshold. Instead, their importance lies in showing that a cut-in can be behaviourally unusual even when it is not physically critical in the conventional sense. In other words, the surprise-based method identifies prediction-abnormal behaviour, not necessarily immediate collision risk.

The nine events detected only by the surprise-based method are summarized in Table 5.1. All of them satisfied the dual-path surprise rule, but none passed the THW-based classical screening rule because their post-change minimum THW values remained above 3.0 s.

Table 5.1: Summary of the nine cut-in events detected only by the surprise-based method.

Event	Dominant surprise response	Peak value of selected metric	Peak-to-$p95$ ratio	Minimum THW (s)
1	Bayesian Surprise	23.195	3.85	3.331
2	Antithesis	7.691	1.39	3.773
3	Bayesian Surprise	58.567	9.73	3.227
4	Bayesian Surprise	25.909	4.30	3.175
5	Bayesian Surprise	39.662	6.59	3.157
6	Bayesian Surprise	19.937	3.31	3.563
7	Bayesian Surprise	27.707	4.60	3.007
8	Antithesis	10.338	1.87	3.177
9	Bayesian Surprise	41.988	6.98	3.698

Table 5.1 shows that seven events are dominated by Bayesian Surprise and two by Antithesis. The peak-to-$p95$ ratios indicate that several events clearly exceed their normal-driving baseline, with the largest case reaching 9.73 times the threshold. At the same time, all minimum THW values remain above 3.0 s. Event 7 is only slightly above this boundary with a minimum THW of 3.007 s, while the other cases have larger THW values. This supports the interpretation that these events are prediction-abnormal, but not classified as critical by the selected THW threshold.

This also clarifies the role of Bayesian Surprise and Antithesis in the analysis. High values of these metrics indicate that the prior belief and the updated or observed

behaviour differ substantially. This means that the vehicle motion caused a strong belief mismatch. However, such a mismatch does not automatically imply that the situation is dangerous. At this stage, the overlap analysis alone cannot determine whether the events selected only by the surprise-based method are truly the risky cut-in events targeted in this study. It only shows that these events are difficult to anticipate under the adopted belief model. Their detection accuracy therefore needs to be evaluated in the following precision–recall analysis, where the metric outputs are compared with the manually labelled risky cut-ins. Therefore, surprise-based metrics should be interpreted here as indicators of behavioural unexpectedness rather than direct measures of physical criticality.

The complementarity between the two methods should therefore be understood as a difference in perspective. The THW-based classical method is well suited for identifying interactions with short longitudinal time gaps. Surprise-based metrics add another layer of information by identifying situations in which surrounding vehicles behave differently from what the ego-centered prediction model expected. Combining the two types of metrics can therefore provide a more complete description of cut-in interactions than using either approach alone.

This interpretation is in line with previous research on classical surrogate safety measures. Different surrogate indicators do not necessarily classify the same traffic situations as critical, because they measure different aspects of risk. Headway-based indicators mainly describe the available following gap, while TTC-based indicators focus on collision imminence under assumptions about future motion [6, 7]. In addition, deceleration-based indicators such as DRAC describe the braking demand required to avoid a crash, which represents another dimension of conflict severity [8]. More generally, reviews of criticality metrics for automated driving also show that different surrogate safety measures capture different aspects of traffic criticality and are therefore better understood as complementary rather than fully interchangeable indicators [1]. Although these studies focus on classical surrogate measures rather than surprise-based metrics, they support the idea that partial overlap between detection methods is expected. In this thesis, the THW-based baseline captures post-cut-in close-following risk, whereas the surprise-based metrics capture deviations from the belief model. The non-overlapping events therefore indicate that the two approaches emphasize different dimensions of risky cut-in behaviour.

At the same time, the overlap analysis has a clear limitation. It only shows whether the THW-based classical method and the surprise-based method select the same events. It does not show whether these selected events are truly risky cut-ins according to the manual labels. Therefore, the non-overlapping events should not be directly interpreted as false positives or false negatives. They only show that the two methods focus on different aspects of the interaction. To evaluate whether these detections are accurate, the detected events need to be compared with the manually labelled risky cut-ins. This is why the overlap analysis should be interpreted together with the following precision–recall analysis.

Overall, the results suggest that the THW-based classical method and the surprise-based method are complementary rather than interchangeable. THW describes the

longitudinal time gap, while surprise-based metrics describe prediction inconsistency and behavioural unexpectedness. The main value of the surprise-based approach is therefore not that it replaces the THW-based classical score, but that it provides additional information about unexpected vehicle behaviour that is not directly captured by THW alone.

5.1.2 Precision–Recall Performance

The precision–recall analysis is used to evaluate the discriminative ability of the different scores [28]. It is therefore different from the previous overlap analysis, which mainly counted whether the two detection methods selected the same events. Because the highD dataset contains a large number of trajectory events, it was not feasible to manually label every candidate case. A smaller validation set was therefore constructed through trajectory visualization. This validation set is not intended to represent the full distribution of safety-critical cut-ins in the dataset. Instead, it provides a practical basis for examining whether each score can separate visually critical cut-ins from visually non-critical ones.

When interpreting the PR results, it is important to note that the THW-based score should not be treated as the only correct definition of a risky cut-in, since headway-based indicators mainly describe temporal spacing and do not cover all aspects of traffic criticality [6, 1]. The THW-based score mainly describes physical conflict, especially the temporal urgency created after a vehicle enters the ego lane. Its strong performance shows that many visually critical cut-ins are associated with a clear reduction in available time gap. However, this does not mean that all safety-relevant events can be explained by short headway alone. Rather, the result shows that physical urgency is one important dimension of cut-in criticality.

The surprise-based metrics describe another dimension of the same interaction: behavioural unexpectedness relative to the prediction-based belief [10, 11]. The PR results also show that the three surprise-based metrics do not behave in the same way. Residual Information has the weakest discriminative performance among them, with an AP of 0.781. This suggests that Residual Information is sensitive to general prediction–observation mismatch, but is less selective for risky cut-in detection. In practical terms, it can identify abnormal motion patterns, but not every abnormal motion pattern is safety-critical. For example, deviations may also occur during smooth lane changes, distant manoeuvres, or interactions with sufficient spacing. Such cases may be surprising to the belief model but may not be visually judged as critical.

Bayesian surprise performs better than Residual Information, with an AP of 0.912. This indicates that the change from the prior belief to the posterior belief is more closely related to visually critical cut-in interactions than general prediction mismatch alone. Bayesian surprise is therefore more suitable for identifying cases in which a surrounding vehicle produces a sustained change in expected future motion. Even so, a large belief update does not necessarily mean that the situation is immediately risky. A manoeuvre can strongly change the predicted belief while still occurring at a sufficient longitudinal distance or with limited influence on the ego

vehicle. For this reason, Bayesian surprise should be interpreted as an indicator of behavioural change rather than a direct measure of physical conflict.

Antithesis achieves the strongest performance among the surprise-based metrics, with an AP of 0.949. This value is closer to the THW-based score than the values of the other surprise-based metrics. The result suggests that Antithesis is more closely associated with unexpected behaviour that is also safety-relevant. Compared with Residual Information and Bayesian surprise, Antithesis appears to be more selective for manoeuvres that strongly conflict with the expected belief. These conflicts are more likely to correspond to cut-in interactions that are visually perceived as critical. Nevertheless, Antithesis should still be regarded as a complementary indicator, because unexpectedness alone is not equivalent to safety-criticality.

These results also help explain why surprise-based metrics can highlight events that are not emphasized by the THW-based classical baseline. A cut-in may be behaviourally unexpected without producing a very short headway. Conversely, another cut-in may be physically urgent without producing an unusually large belief mismatch. Therefore, the differences in PR performance should be understood as differences in metric sensitivity. They should not be interpreted as evidence that one metric family is universally correct and the other is incorrect.

Together with the overlap analysis, the PR results provide a broader evaluation of the proposed surprise-based method. The overlap analysis shows whether the method detects a meaningful and complementary set of events at the dataset level. The PR analysis further shows whether the corresponding scores can distinguish visually critical and non-critical interactions in the manually labelled validation set. From these two perspectives, the surprise-based method is feasible not as a replacement for the THW-based classical baseline, but as a complementary approach. Its main contribution is that it adds information about behavioural unexpectedness to the evaluation of safety-critical traffic interactions.

Therefore, the contribution of the surprise-based method lies not in replacing the THW-based classical baseline, but in providing an additional behavioural-unexpectedness perspective. Residual Information is more suitable for broad abnormal-behaviour screening, Bayesian surprise is useful for identifying structured belief changes, and Antithesis appears most relevant for detecting unexpected behaviours with potential safety significance.

5.1.3 Sensitivity to Temporal Parameters

The sensitivity analysis of h and z provides an independent examination of how the temporal settings affect the behaviour of the surprise-based metrics, which is relevant because surprise values depend on the temporal relationship between prior belief, prediction horizon, and observation [10, 11]. Unlike the previous comparison between THW-based and surprise-based detection, the purpose here is not to determine whether one method is generally better than another. Instead, this analysis examines whether the surprise values are stable when the temporal definition of the belief comparison is changed.

The effect of the history window h can be understood from how the prior belief is constructed. In the CV-based belief model, the vehicle state is propagated according to the state transition equation in Equation (2.11), with uncertainty introduced through the process-noise covariance in Equation (2.12). Therefore, changing h changes the time point from which the prior expectation is formed. A very short history window, such as $h = 0.5$ s, leaves only limited temporal separation between the prior state and the observed cut-in behaviour. As a result, the belief may remain close to the current motion state, and the behavioural change may not yet be sufficiently visible. When h is increased to 1.0 s, the prior belief contains more historical context, making the contrast between expected and observed motion clearer.

This explains why Residual Information and Shannon Surprisal improve when h increases from 0.5 s to 1.0 s. Shannon Surprisal is defined in Equation (2.4) as the negative log probability of the observed outcome under the prior belief. Residual Information, defined in Equation (2.7), further compares the likelihood of the observed outcome with the likelihood of the most expected outcome. Both metrics therefore depend directly on the probability assigned by the prior belief to the realised vehicle state. If the history window is too short, the prior belief may not differ enough from the observed state, which weakens the prediction–observation contrast. A moderate history window makes this contrast more informative and improves the ranking of risky cut-in events.

However, increasing h beyond 1.0 s does not necessarily add useful information. When the history window becomes too long, the prior belief may be generated from a traffic state that is less directly related to the current cut-in interaction. The limited improvement from $h = 1.0$ s to $h = 2.0$ s therefore suggests that most of the useful temporal context is already captured within approximately one second. From this perspective, $h = 1.0$ s provides a reasonable balance between short-term responsiveness and sufficient historical contrast.

The weaker sensitivity of Bayesian Surprise and Antithesis to h is also consistent with their definitions. Bayesian Surprise, shown in Equation (2.8), measures the KL divergence between the posterior belief and the prior belief. It therefore responds to the overall belief update rather than only to the likelihood of one observed state. Antithesis, defined in Equation (2.9), is even more selective because it only accumulates belief increase in regions that were outside the prior expectation. For these two metrics, the exact length of h is less important than whether the observation produces a clear change in the belief distribution. This helps explain why their performance is more stable across the tested history windows.

The lookahead time z affects the future point at which the prediction and observation are compared. In the CV model, a longer prediction horizon means that the state transition in Equation (2.11) is applied over a longer time interval. At the same time, uncertainty accumulates through the process-noise term in Equation (2.11) and the covariance in Equation (2.12). In theory, this wider belief distribution can reduce the contrast between the expected state and the observed state, especially for metrics such as Shannon Surprisal and Residual Information that depend directly on prior likelihood values. However, the results show that the tested metrics remain

relatively stable when z changes from 0.2 s to 1.0 s and 2.0 s. This indicates that the event ranking is not strongly dependent on one specific future comparison point.

Overall, the results suggest that h has a more visible influence than z on the behaviour of the surprise-based metrics. The history window changes the temporal context used to form the prior belief, while the lookahead time mainly changes the prediction horizon and the amount of propagated uncertainty. Residual Information and Shannon Surprisal benefit from a sufficient history window because they are directly linked to the prior likelihood terms in Equations (2.4) and (2.7). Bayesian Surprise and Antithesis are more robust because they focus on belief updates, as shown in Equations (2.8) and (2.9). Therefore, the sensitivity analysis supports the robustness of the surprise-based framework and suggests that $h = 1.0$ s and $z = 0.2$ s are reasonable temporal settings for capturing near-term behavioural unexpectedness.

5.2 Answers to Research Questions

Based on the results, the two research questions can be answered as follows.

5.2.1 How do surprise-based traffic safety metrics compare with the THW-based classical surrogate safety metric in identifying risky driving scenarios under the same highway conditions?

The results show that the surprise-based metrics and the THW-based classical surrogate safety metric are related, but not interchangeable. The overlap analysis shows that 941 events are detected by both methods. This corresponds to 76.82% of all 1,225 unique detected events, with a Jaccard overlap of 0.768. This high overlap indicates that many risky cut-ins contain both physical conflict and expectation violation. In other words, cut-ins that reduce the post-change THW often also deviate from the motion predicted by the belief model.

The non-overlapping events clarify the difference between the two perspectives. The THW-based screening method detects 275 events that are not detected by the surprise-based method. These cases satisfy the selected THW condition and therefore mainly represent short-gap interactions. By contrast, the surprise-based method detects 9 events that are not detected by the THW-based screening method. Their surprise values are high, but their minimum THW values remain above the THW threshold of 3.0 s. These events are therefore unexpected according to the belief model, but they are not classified as physically critical by the selected THW rule.

The precision–recall results further support this conclusion. The THW-based score achieves the highest AP value of 0.982, showing that short headway is the strongest single indicator under the selected highway cut-in conditions and manual labels. Among the surprise-based metrics, Antithesis performs best with an AP of 0.949,

followed by Bayesian Surprise with an AP of 0.912 and Residual Information with an AP of 0.781. Therefore, the surprise-based metrics should not be interpreted as direct substitutes for THW. Their main value is that they add a behavioural-unexpectedness perspective to the evaluation of risky cut-ins.

5.2.2 How do the history window h and lookahead time z affect the detection performance of surprise-based metrics for risky cut-in identification?

The sensitivity analysis shows that the history window h has a clearer influence on detection performance than the lookahead time z. The history window changes the temporal context used to form the prior belief. In the CV-based belief model, this prior is generated through the state propagation in Equation (2.11), with uncertainty controlled by the process-noise covariance in Equation (2.12). When h increases from 0.5 s to 1.0 s, the AP values of Residual Information and Shannon Surprisal improve, suggesting that a very short history window does not provide enough temporal contrast between expected and observed motion.

This behaviour is consistent with the metric definitions. Shannon Surprisal in Equation (2.4) and Residual Information in Equation (2.7) both depend directly on the prior likelihood assigned to the observed state. They are therefore sensitive to whether h provides enough context for a meaningful prediction–observation mismatch to appear. However, increasing h further to 2.0 s gives only limited additional improvement, indicating that most useful temporal context is already captured around $h = 1.0$ s.

Bayesian Surprise and Antithesis are more robust to changes in h, because they focus on belief updates rather than only on the likelihood of one observed state. Bayesian Surprise measures the KL divergence between posterior and prior beliefs, as shown in Equation 2.8. Antithesis further restricts the response to regions that were outside the prior expectation and became more likely after the update, as defined in Equation 2.9. This explains why their performance changes less strongly across the tested history windows.

The lookahead time z has a smaller effect within the tested range. Although a longer z increases the prediction horizon and allows more uncertainty to accumulate through the process noise, the AP values remain relatively stable when z changes from 0.2 s to 1.0 s and 2.0 s. Overall, the results suggest that a moderate history window, especially $h = 1.0$ s, is useful for constructing a sufficiently informative prior belief, while the tested lookahead-time range is less critical for the final risky cut-in detection performance. The setting $z = 0.2$ s is therefore reasonable for capturing near-term behavioural unexpectedness.

5.3 Limitations

5.3.1 Limitations Related to the Risky Cut-in Definition

The results depend on how a risky cut-in event is defined. In the THW-based classical screening workflow, the threshold $\min(\mathrm{THW}_{post}) \leq 3.0$ s was used to include a broad set of close-following cut-in events. This choice is useful for exploratory comparison, but it should not be interpreted as a crash-imminent boundary.

THW was selected as the classical baseline because it directly captures the following gap after a cut-in and therefore matches the focus of this thesis on post-cut-in close-following situations. It is also simple, interpretable, and stable as a baseline for comparison with surprise-based metrics. TTC and DRAC could be included in future work to provide a stricter or more dynamic conflict definition, especially when the aim is to focus on closing-speed effects or required braking demand rather than temporal spacing alone.

A stricter definition, such as a shorter THW threshold or additional TTC- or DRAC-based conditions, would likely select fewer but more severe events. This could change both the overlap between the THW-based and surprise-based methods and the precision–recall results. The conclusions about complementarity should therefore be understood as conditional on the selected event definition: THW mainly captures physical urgency, whereas surprise-based metrics capture deviations from the expected belief. Future work could repeat the analysis under stricter criticality definitions to test whether the same relationship holds for more severe safety-critical scenarios.

5.3.2 Limitations of Manual Ground-Truth Classification

The ground truth used for risky cut-in evaluation was established through manual review of event visualizations. It was therefore not based on crash records or an externally validated event database. This introduces a degree of subjectivity, since the classification depends on the reviewer's interpretation of whether a cut-in should be considered critical or risky.

The adopted definition of criticality was also relatively inclusive. It was intended to capture potentially risky close-following cut-in situations, rather than only crash-imminent or severe conflict events. This choice is suitable for exploratory comparison, but it can influence the precision–recall results. A stricter or more conservative ground-truth definition could change which events are treated as true positives and false positives for both the THW-based baseline and the surprise-based metrics.

Compared with more systematic naturalistic driving studies, such as SHRP2 [35, 36], the manual classification in this thesis is limited in scale and procedure [35, 36]. Such studies typically use more structured review protocols, clearer event taxonomies, and often multiple reviewers or quality-control steps. This thesis does not reach the same level of large-scale validation, inter-rater agreement, or multi-reviewer consistency checking. Future work could improve the reliability of the ground truth by using

multiple independent reviewers, clearer annotation criteria, inter-rater agreement analysis, and possibly a more fine-grained severity scale for cut-in events.

5.3.3 Limitations of the Constant-Velocity Belief Model

Another limitation concerns the belief generation model [15, 18]. The surprise-based metrics were computed from a constant-velocity Gaussian prediction, which provides a transparent baseline for comparing metrics but simplifies the motion behaviour of surrounding vehicles.

This simplification is especially relevant for cut-in scenarios. The future motion of a vehicle may depend on lane-changing intention, lane geometry, available gaps, neighbouring vehicles, and the driver's decision to continue or abort a manoeuvre. These factors are not explicitly represented by the constant-velocity model, which mainly propagates the current state with calibrated uncertainty.

Because Residual Information, Bayesian Surprise, and Antithesis are all computed from the mismatch between predicted belief and observed behaviour, their values depend on the quality of that belief. Some high surprise responses may therefore reflect model limitations rather than genuinely abnormal behaviour, while some relevant manoeuvres may remain unsurprising if they are consistent with the constant-velocity assumption. The results should therefore be interpreted as conclusions under this specific belief model, not as general conclusions for all possible belief-generation approaches.

5.3.4 Limitations Related to Dataset Scope

The analysis was conducted on the highD dataset, which provides naturalistic vehicle trajectories from German highways [12]. This dataset is well suited for studying highway lane changes and cut-in manoeuvres, but it also limits the scope of the conclusions to structured highway traffic.

Other traffic environments, such as urban intersections, roundabouts, merging areas, pedestrian crossings, or mixed traffic scenarios, may involve more diverse interaction patterns and uncertainty sources. In addition, highD provides trajectory-level information but not direct evidence of driver intention, perception, attention, or subjective comfort. The study can therefore infer behavioural unexpectedness from motion and belief mismatch, but it cannot explain why a driver performed a manoeuvre or how surrounding drivers perceived it.

The observed complementarity between THW-based and surprise-based metrics may therefore partly depend on the characteristics of highway traffic in highD. Applying the same framework to other datasets, road types, and interaction scenarios would help test whether the conclusions remain stable beyond highway cut-in events.

5.4 Future Research Directions

5.4.1 Improving the Belief Generation Model with Reachability-Based Constraints

A direct extension of this work is to improve belief generation with reachability-based constraints. The Field of Safe Motion extends the earlier Field of Safe Travel by using reachability analysis to describe whether a driver still has a collision-free escape route, or an "out", in the near future [29, 30]. This idea is relevant here because the surprise metrics depend directly on the predicted belief distribution.

Reachability analysis can define the set of future states that a vehicle can physically reach under kinematic constraints [31, 30]. Such constraints could reduce probability mass in infeasible regions and align the belief distribution more closely with road geometry and lane structure. For example, states outside physically reachable areas or inconsistent with lane boundaries could be down-weighted before computing surprise.

This extension would make the belief model less dependent on unconstrained constant-velocity extrapolation. As a result, surprise responses would more clearly reflect unexpected driving behaviour rather than artefacts of an overly broad or physically unrealistic prediction.

5.4.2 Incorporating Comfort- and Dread-Zone Boundaries into Belief Generation

Future work could also incorporate comfort-zone and dread-zone boundaries into belief generation. These concepts are rooted in earlier work on safety margins and the Field of Safe Travel [29, 32, 33]. Bärgman et al. quantified such boundaries for left-turn-across-path/opposite-direction scenarios, where the comfort zone describes situations drivers accept as safe and the dread-zone boundary marks situations they are unwilling to enter voluntarily [34].

These concepts are useful because physically possible manoeuvres are not necessarily behaviourally likely. In a highway cut-in setting, a lane change with sufficient gap, moderate lateral motion, and reasonable relative speed could be assigned higher prior probability. By contrast, a cut-in with a very small gap, high lateral acceleration, or limited safety margin could be represented as less likely or more aggressive.

Using such behavioural boundaries would make the belief distribution reflect not only what a vehicle can do, but also what a human driver is likely to consider acceptable. This could improve the interpretability of surprise values, because unexpectedness would be measured against a belief model that includes both kinematic feasibility and behavioural plausibility.

However, existing numerical boundaries should not be transferred directly to this thesis. The study by Bärgman et al. focused on left-turn interactions at intersections, whereas this thesis focuses on highway cut-ins [34]. Future work would

therefore need to estimate comfort-zone-like and dread-zone-like boundaries specifically for highway lane changes.

5.4.3 Extending Belief Generation with Learning-Based Trajectory Prediction

A third direction is to replace or extend the constant-velocity model with learning-based trajectory prediction. Models such as VectorNet encode agent trajectories and map elements as vectors and use a hierarchical graph neural network to capture both local polyline features and global interactions among traffic participants and road components [19]. This is relevant for cut-in analysis because lane structure, neighbouring vehicles, and multi-agent interactions all influence whether a lane change is likely.

A learning-based model could generate a more context-aware belief distribution [19]. Instead of extrapolating only from the current velocity, it could learn whether a vehicle is likely to continue lane keeping, initiate a lane change, complete a cut-in, or adjust to nearby traffic. Map and lane information could also help the predicted belief follow the road geometry more closely.

For surprise-based analysis, the key requirement is that the model outputs a probabilistic belief rather than only a single future trajectory. One possible approach is to combine a VectorNet-style encoder with a probabilistic or multi-modal trajectory decoder, so that multiple plausible futures and their uncertainties are represented. This would provide a stronger basis for applying Residual Information, Bayesian Surprise, and Antithesis to complex highway interactions.

6

Conclusion

This thesis investigated whether surprise-based traffic safety metrics can support the identification of risky cut-in events in highway driving scenarios. Using naturalistic trajectory data from the highD dataset, the study compared surprise-based metrics with a THW-based classical surrogate safety score and examined how temporal parameter settings affect detection performance.

A constant-velocity belief generation model was developed to represent the expected future motion of surrounding vehicles. The model propagated vehicle motion under a CV assumption and represented uncertainty with a Gaussian belief distribution. Based on this belief, Residual Information, Bayesian Surprise, and Antithesis were calculated. A normal-driving baseline was then used to define data-driven surprise thresholds from the 95th percentile of each metric.

The results show that the THW-based classical score and the surprise-based metrics are related, but not interchangeable. The overlap analysis found that 941 events were detected by both methods out of 1,225 unique detected events, corresponding to a Jaccard overlap of 0.768. This indicates that many risky cut-ins combine physical closeness with unexpected motion. However, the non-overlapping cases show that the two approaches emphasize different aspects of the interaction. THW mainly captures longitudinal time-gap reduction and physical urgency, whereas surprise-based metrics capture deviations from the prediction-based belief.

The precision–recall results support the same interpretation. The THW-based score achieved the highest AP value of 0.982, showing that short post-change headway is a strong indicator for the manually labelled risky cut-ins in this study. Among the surprise-based metrics, Antithesis performed best with an AP of 0.949, followed by Bayesian Surprise with 0.912 and Residual Information with 0.781. These results suggest that surprise-based metrics can identify meaningful behavioural unexpectedness, but they should be interpreted as complementary indicators rather than direct replacements for the THW-based classical baseline.

The sensitivity analysis showed that the history window h had a clearer effect than the lookahead time z. Increasing h from 0.5 s to 1.0 s improved the performance of some surprise-based metrics, while increasing it further to 2.0 s gave only limited additional benefit. In contrast, changes in z within the tested range had a smaller influence on detection performance. This suggests that a moderate history window is useful for forming an informative prior belief, while the tested lookahead settings

are less critical.

Overall, the thesis demonstrates that surprise-based metrics add a useful behavioural perspective to risky cut-in analysis. They help identify situations in which observed vehicle motion is unexpected relative to a probabilistic belief model. At the same time, unexpectedness is not equivalent to physical risk. A surprising manoeuvre may still occur with sufficient spacing, while a physically risky cut-in may be predictable under the belief model. For this reason, the strongest interpretation is that surprise-based metrics and the THW-based classical baseline should be used together to describe both behavioural unexpectedness and post-cut-in temporal criticality.

The conclusions should be understood in light of the study's limitations. The risky cut-in definition depends on the selected THW threshold and manual labels, the belief model is based on a simplified constant-velocity assumption, and the analysis is limited to structured highway scenarios in the highD dataset. Future work should therefore test stricter criticality definitions, apply the framework to more diverse traffic environments, and improve belief generation. Promising directions include reachability-based constraints, comfort- and dread-zone boundaries, and learning-based trajectory prediction models that can represent multiple plausible future behaviours probabilistically.

Bibliography

[1] Westhofen, L., Neurohr, C., Koopmann, T., Butz, M., Schütt, B., Utesch, F., Neurohr, B., Gutenkunst, C., & Böde, E. (2023). Criticality metrics for automated driving: A review and suitability analysis of the state of the art. *Archives of Computational Methods in Engineering, 30*, 1–35. doi:10.1007/s11831-022-09788-7

[2] Zhang, X., Tao, J., Tan, K., Törngren, M., Gaspar Sánchez, J. M., Ramli, M. R., Tao, X., Gyllenhammar, M., Wotawa, F., Mohan, N., Nica, M., & Felbinger, H. (2023). Finding critical scenarios for automated driving systems: A systematic mapping study. *IEEE Transactions on Software Engineering, 49*(3), 991–1026. doi:10.1109/TSE.2022.3170122

[3] Kalra, N., & Paddock, S. M. (2016). Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice, 94*, 182–193. doi:10.1016/j.tra.2016.09.010

[4] Roche, F. (2021). Assessing subjective criticality of take-over situations: Validation of two rating scales. *Accident Analysis & Prevention, 159*, 106216. doi:10.1016/j.aap.2021.106216

[5] Tscharn, R., Naujoks, F., & Neukum, A. (2018). The perceived criticality of different time headways is depending on velocity. *Transportation Research Part F: Traffic Psychology and Behaviour, 58*, 1043–1052. doi:10.1016/j.trf.2018.08.001

[6] Vogel, K. (2003). A comparison of headway and time to collision as safety indicators. *Accident Analysis & Prevention, 35*(3), 427–433. doi:10.1016/S0001-4575(02)00022-2

[7] Minderhoud, M. M., & Bovy, P. H. L. (2001). Extended time-to-collision measures for road traffic safety assessment. *Accident Analysis & Prevention, 33*(1), 89–97. doi:10.1016/S0001-4575(00)00019-1

[8] Fu, C., & Sayed, T. (2021). Comparison of threshold determination methods for the deceleration rate to avoid a crash (DRAC)-based crash estimation. *Accident Analysis & Prevention, 153*, 106051. doi:10.1016/j.aap.2021.106051

[9] Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research, 49*(10), 1295–1306. doi:10.1016/j.visres.2008.09.007

[10] Dinparastdjadid, A., Supeene, I., & Engström, J. (2023). Measuring surprise in the wild. *arXiv preprint arXiv:2305.07733*. doi:10.48550/arXiv.2305.07733

[11] Engström, J., Liu, S.-Y., Dinparastdjadid, A., & Simoiu, C. (2024). Modeling road user response timing in naturalistic traffic conflicts: A surprise-based framework. *Accident Analysis & Prevention, 198*, 107460. doi:10.1016/j.aap.2024.107460

[12] Krajewski, R., Bock, J., Kloeker, L., & Eckstein, L. (2018). The highD dataset: A drone dataset of naturalistic vehicle trajectories on German highways for validation of highly automated driving systems. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)* (pp. 2118–2125). IEEE. doi:10.1109/ITSC.2018.8569552

[13] SAE International. (2013). *Operational definitions of driving performance measures and statistics* (SAE J2944 Proposed Draft). SAE International.

[14] Hayward, J. C. (1972). Near-miss determination through use of a scale of danger. *Highway Research Record, 384*, 24–34.

[15] Lefèvre, S., Vasquez, D., & Laugier, C. (2014). A survey on motion prediction and risk assessment for intelligent vehicles. *Robomech Journal, 1*, 1. doi:10.1186/s40648-014-0001-z

[16] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(3), 379–423; *27*(4), 623–656. doi:10.1002/j.1538-7305.1948.tb01338.x; doi:10.1002/j.1538-7305.1948.tb00917.x

[17] Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79–86. doi:10.1214/aoms/1177729694

[18] Bar-Shalom, Y., Li, X. R., & Kirubarajan, T. (2001). Estimation for kinematic models. In *Estimation with Applications to Tracking and Navigation: Theory, Algorithms and Software* (pp. 267–299). John Wiley & Sons. doi:10.1002/0471221279.ch6

[19] Gao, J., Sun, C., Zhao, H., Shen, Y., Anguelov, D., Li, C., & Schmid, C. (2020). VectorNet: Encoding HD maps and agent dynamics from vectorized representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11525–11533).

[20] Toledo, T., & Zohar, D. (2007). Modeling duration of lane changes. *Transportation Research Record, 1999*(1), 71–78. doi:10.3141/1999-08

[21] Olsen, E. C. B., Lee, S. E., Wierwille, W. W., & Goodman, M. J. (2002). Analysis of distribution, frequency, and duration of naturalistic lane changes. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 46*(22), 1923–1927. doi:10.1177/154193120204602203

[22] Treiber, M., & Kesting, A. (2013). *Traffic Flow Dynamics: Data, Models and Simulation*. Springer. doi:10.1007/978-3-642-32460-4

[23] FGSV. (2008/2011). *Guidelines for the Design of Motorways (RAA): English*

version of Richtlinien für die Anlage von Autobahnen, edition 2008, translation 2011. Forschungsgesellschaft für Straßen- und Verkehrswesen.

[24] SWOV. (2012). *Headway times and road safety.* SWOV Fact Sheet. The Hague: SWOV Institute for Road Safety Research.

[25] Davis, N., Raina, G., & Jagannathan, K. (2020). A framework for end-to-end deep learning-based anomaly detection in transportation networks. *Transportation Research Interdisciplinary Perspectives, 5*, 100112. doi:10.1016/j.trip.2020.100112

[26] Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles, 37*, 547–579.

[27] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1–26. doi:10.1214/aos/1176344552

[28] Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233–240). doi:10.1145/1143844.1143874

[29] Gibson, J. J., & Crooks, L. E. (1938). A theoretical field-analysis of automobile-driving. *The American Journal of Psychology, 51*(3), 453–471.

[30] Johnson, L., Victor, T., & Engström, J. (2026). The Field of Safe Motion: Operationalizing affordances in the Field of Safe Travel using reachability analysis. *arXiv preprint arXiv:2604.27168.*

[31] Althoff, M., Frehse, G., & Girard, A. (2021). Set propagation techniques for reachability analysis. *Annual Review of Control, Robotics, and Autonomous Systems, 4*, 369–395.

[32] Näätänen, R., & Summala, H. (1974). A model for the role of motivational factors in drivers' decision-making. *Accident Analysis & Prevention, 6*, 243–261.

[33] Summala, H. (2007). Towards understanding motivational and emotional factors in driver behaviour: Comfort through satisficing. In C. Cacciabue (Ed.), *Modelling Driver Behaviour in Automotive Environments: Critical Issues in Driver Interactions with Intelligent Transport Systems* (pp. 189–207). Springer. doi:10.1007/978-1-84628-618-6_11

[34] Bärgman, J., Smith, K., & Werneke, J. (2015). Quantifying drivers' comfort-zone and dread-zone boundaries in left-turn-across-path/opposite-direction (LTAP/OD) scenarios. *Transportation Research Part F: Traffic Psychology and Behaviour, 35*, 170–184. doi:10.1016/j.trf.2015.10.003

[35] Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences, 113*(10), 2636–2641. doi:10.1073/pnas.1513271113

[36] Virginia Tech Transportation Institute. (2015). *SHRP 2 Naturalistic Driving Study: Researcher dictionary for video reduction data*. Virginia Tech Transportation Institute.

A

Appendix

A.1 Sensitivity of Final Cut-in Event Selection to THW Threshold

Figure A.1 illustrates the sensitivity of the final cut-in event set to the selected post-change THW threshold. The number of retained events increases from 258 at 0.8 s to 1216 at 3.0 s. Compared with the 3.0 s baseline, the 2.0 s threshold retains about 78.3% of the events, while the 1.5 s and 1.2 s thresholds retain about 61.1% and 44.8%, respectively.

This confirms that the THW threshold has a clear effect on the size and strictness of the selected event set. Lower thresholds identify more severe close-following cut-ins, whereas the 3.0 s threshold provides a broader and more inclusive screening criterion for potentially relevant cut-in events.

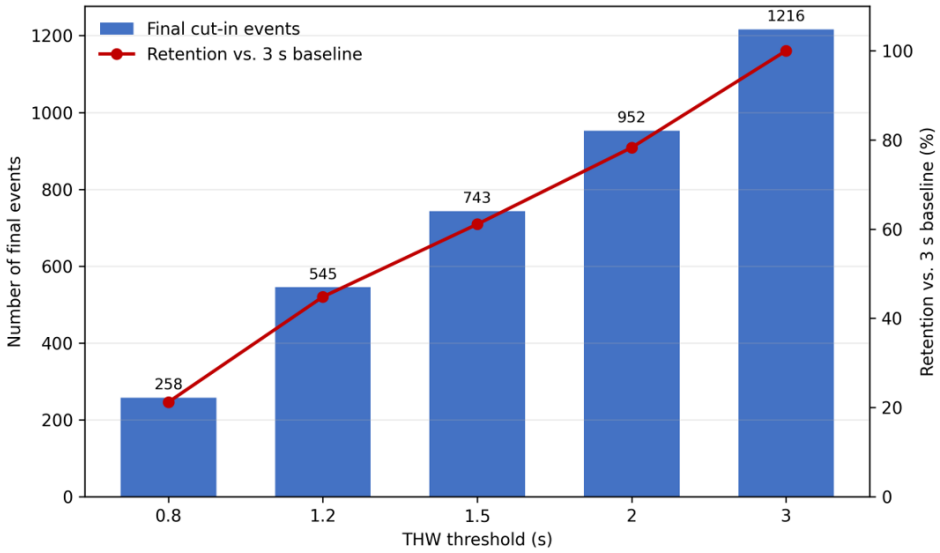

Figure A.1: Sensitivity of final cut-in event selection to different post-change THW thresholds. The bars show the number of retained final cut-in events, and the red line shows the retention rate relative to the 3.0 s baseline.

A.2 Bootstrap Confidence Intervals for AP Scores

To provide an additional indication of the uncertainty in the AP estimates, bootstrap confidence intervals were computed for the surprise-based metrics under two selected temporal settings. The first setting, $h = 0.5$ s and $z = 0.2$ s, was used to examine the uncertainty under the shortest history window considered in the sensitivity analysis. The second setting, $h = 1.0$ s and $z = 0.2$ s, corresponds to the main temporal setting used in the surprise-based detection pipeline.

The bootstrap analysis was performed at the event level using the labelled validation data. In each bootstrap iteration, labelled events were resampled with replacement, and the AP score was recomputed for each metric. The 95% confidence interval was obtained from the 2.5th and 97.5th percentiles of the resulting bootstrap AP distribution.

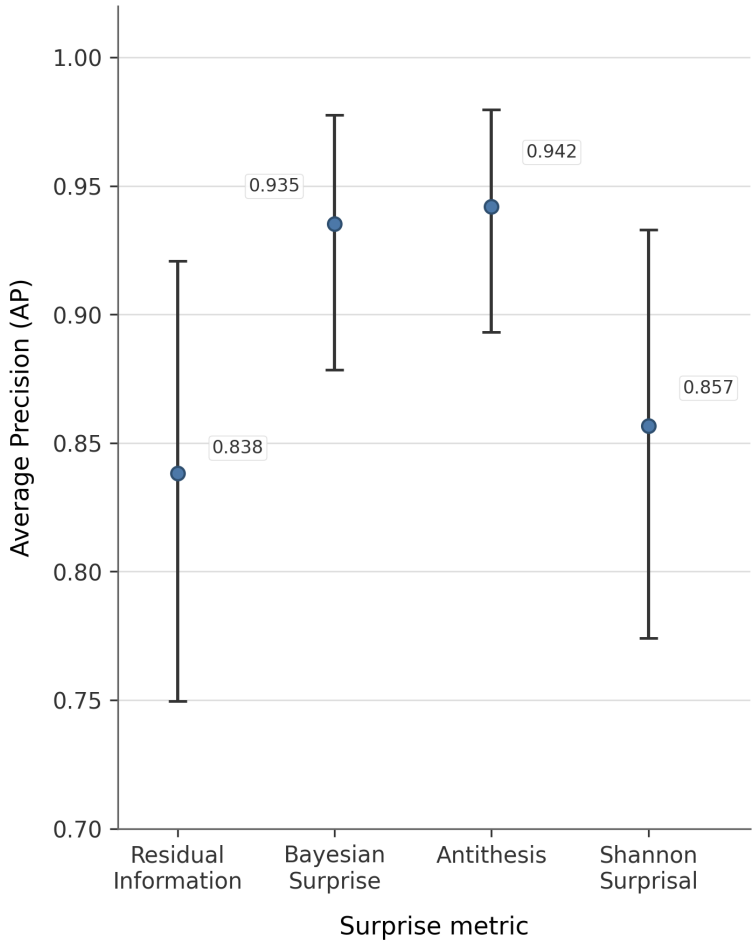

Figure A.2: Bootstrap confidence intervals of AP scores for the surprise-based metrics under $h = 0.5$ s and $z = 0.2$ s. The markers show the original AP values, while the vertical error bars indicate the 95% bootstrap confidence intervals obtained from event-level resampling of the labelled validation data.

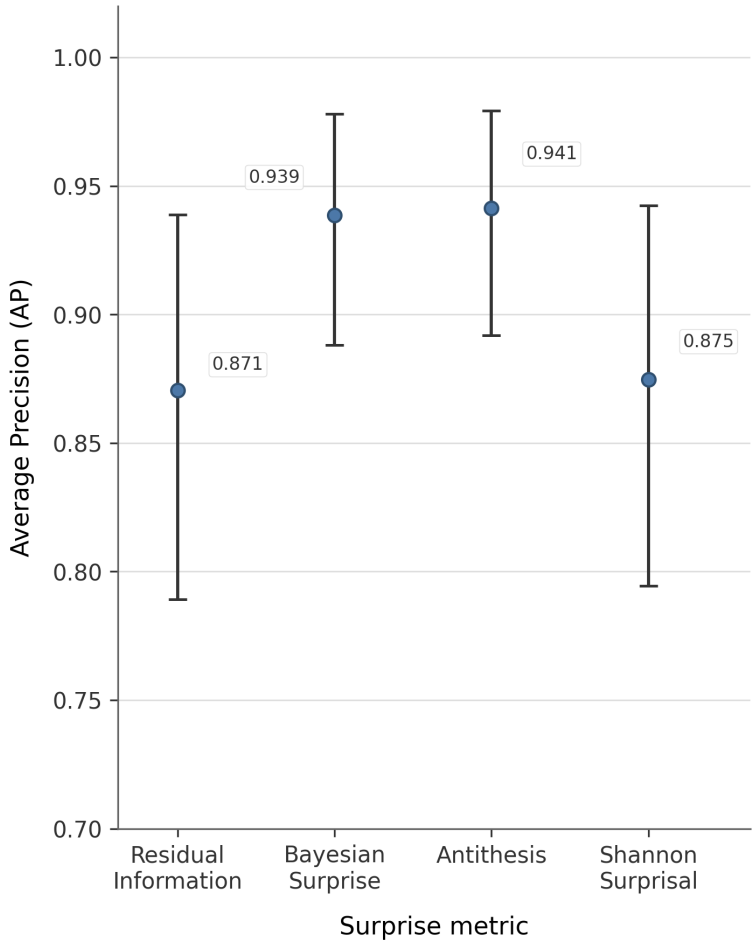

Figure A.3: Bootstrap confidence intervals of AP scores for the surprise-based metrics under $h = 1.0$ s and $z = 0.2$ s. The markers show the original AP values, while the vertical error bars indicate the 95% bootstrap confidence intervals obtained from event-level resampling of the labelled validation data.

A.3 Event Counts in the Classical Cut-in Screening Workflow

To provide additional transparency on the event selection process, Table A.1 summarizes the cumulative number of candidate events retained during the classical THW-based cut-in screening workflow. Several technical consistency checks are grouped together to keep the table concise. The table is included as supplementary information, since the main methodology section focuses on the rationale for the filter design.

Table A.1: Cumulative event counts in the classical THW-based cut-in screening workflow.

Step	Screening stage	Main condition	Remaining	Dropped	Retention (%)
1	Initial lane-change candidates	`laneId` changes between consecutive ego frames	13,949	0	100.00
2	Valid post-change follower	Post-change `followingId` is non-zero	10,827	3,122	77.62
3	New follower condition	Post-change follower differs from the pre-change follower	10,822	5	77.58
4	Trajectory and common-frame availability	Ego and follower trajectories available, with at least 3 common post-change frames	10,723	99	76.87
5	Valid post-change gap	Minimum post-change gap ≥ 0.5 m	5,011	5,712	35.92
6	THW-based trigger	Valid follower speed and minimum post-change THW ≤ 3.0 s	4,264	747	30.57
7	Change-frame gap consistency	Change-frame rows available, finite gaps, and change-frame gap < 150 m	4,247	17	30.45
8	Relative-speed condition	Positive closing speed, $\Delta v > 0$, meaning the follower is faster	1,216	3,031	8.72

A.4 Event Counts in the Surprise-Based Cut-in Screening Workflow

To provide additional transparency on the event selection process, Table A.2 summarizes the cumulative number of candidate events retained during the surprise-based cut-in screening workflow. The counts are cumulative, meaning that each row shows the number of events remaining after the corresponding screening stage has been applied. Several technical checks are grouped together to keep the table concise. The table is included as supplementary information to show how the initial lane-change candidates were reduced to the final surprise-based detections.

Table A.2: Cumulative event counts in the surprise-based cut-in screening workflow.

Step	Screening stage	Main condition	Remaining	Dropped	Retention (%)
1	Initial lane-change candidates	`laneId` changes between consecutive ego frames	13,949	0	100.00
2	Valid post-change follower	Post-change `followingId` is non-zero	10,827	3,122	77.62
3	New follower condition	Post-change follower differs from pre-change follower	10,822	5	77.58
4	Trajectory and common-frame availability	Ego and follower trajectories available, with at least 3 common post-change frames	10,723	99	76.87
5	Post-change geometric sanity	Change-frame gap $\leq$ 100 m, post-change gap $\geq$ 0.5 m, and valid follower speed	4,285	6,438	30.72
6	Surprise metrics available and finite	Finite event-level peak RI, BS, and A values are available in the event window	4,026	259	28.86
7	Surprise-threshold trigger	Peak RI exceeds $p99$ with BS or A above $p95$; or RI, BS, and A all exceed $p95$	3,396	630	24.35
8	Post-trigger physical consistency	Change-frame rows available, gaps $<$ 150 m, and positive closing speed, $\Delta v > 0$	950	2,446	6.81

A.5 Representative Example of Manual Labelling

To make the manual labelling process more transparent, one representative borderline cut-in case is shown in Figure A.4. This example is not intended to define a strict quantitative classification rule. Instead, it illustrates how manual judgment was applied in practice. In this thesis, the manual labeling used a relatively inclusive definition of criticality: an event could be labeled as critical when it created a close-following or interaction-demanding situation, even if it was not crash-imminent.

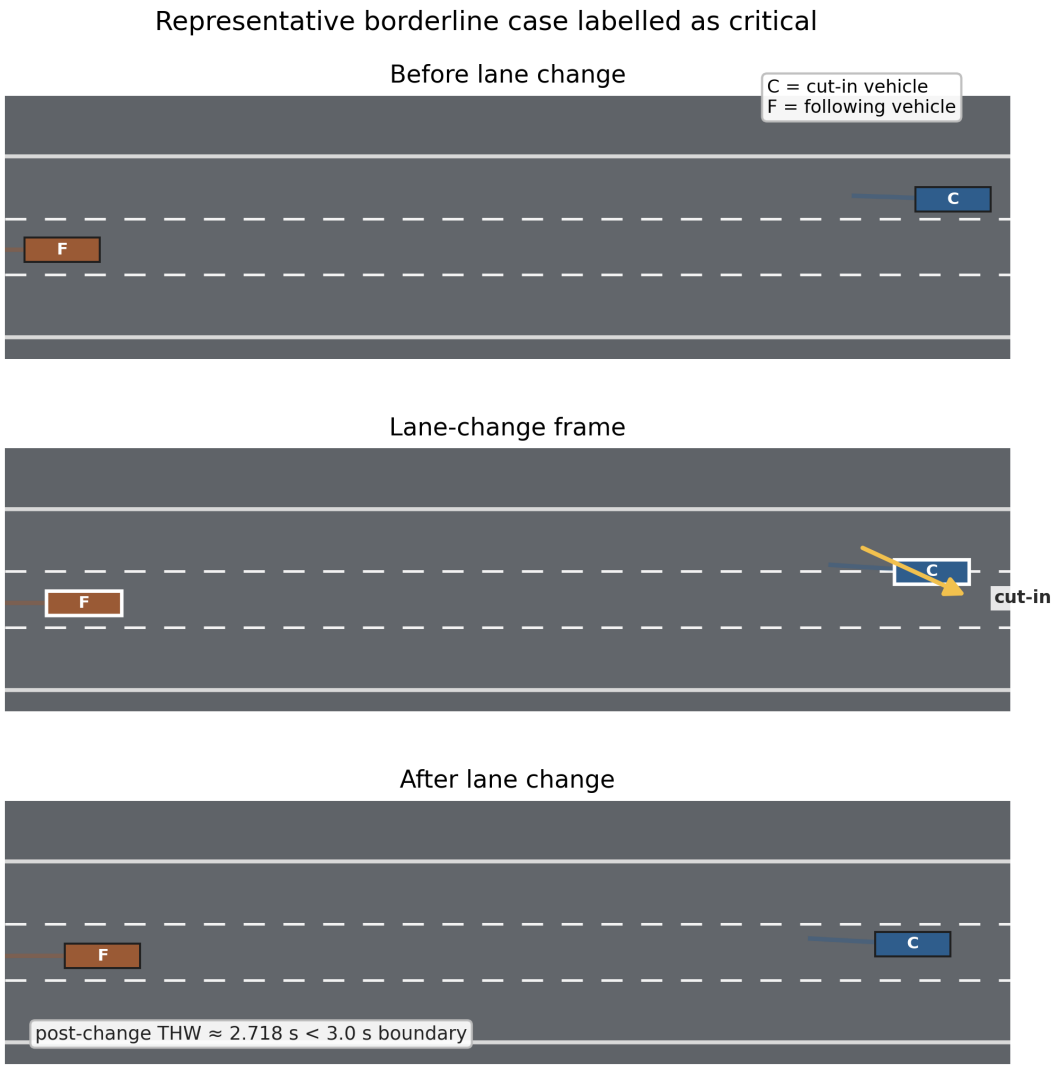

Figure A.4: Representative borderline cut-in event labelled as critical. The event was not crash-imminent, but the post-change time headway was approximately 2.718 s, which is below and close to the 3.0 s criticality boundary used in this thesis. This example illustrates the relatively inclusive manual labelling criterion adopted for the validation set.

DEPARTMENT OF MECHANICS AND MARITIME SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY