Testing and Evaluation of Autonomous Driving Systems

From Simulated to Real-world Test Environments

Andrea Stocco

SIESTA summer school

02.09.2026

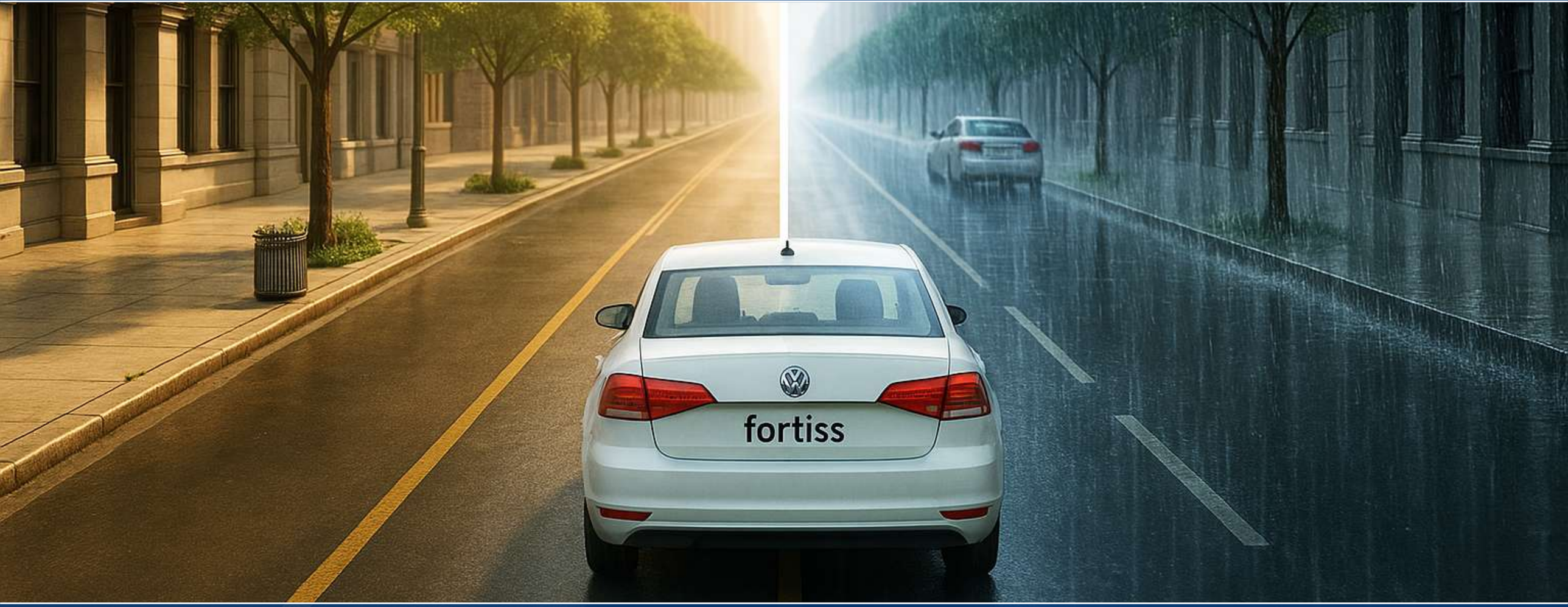


Image generated with DALL·E 3

whoami

Cyber Security at *Tallinn University of Technology* and *University of Tartu*, Estonia

PhD in Computer Science at *University of Genova*, Italy

PostDoctoral Researcher of the *Software Analysis and Testing* group at University of British Columbia, Canada

PostDoctoral Researcher of the *Software Institute at* Università della Svizzera italiana, Switzerland

whoami

Lead of the Automated Software Testing (AST) Field of Competence

Email

stocco@fortiss.org
andrea.stocco@tum.de

Chair of Software Engineering for Data-intensive Applications

Website/X/LinkedIn

tsigalko18.github.io/
@tsigalko18

www.linkedin.com/in/andreastocco/

Acknowledgments

Stefano Carlo Lambertenghi

Xingcheng Chen

Lev Sorokin

Oliver Weißl

Davide Yi Xian Hu

Manfredi Napolitano

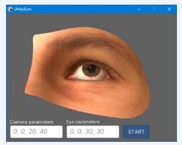
https://github.com/ast-fortiss-tum/

And many other collegues and co-authors…

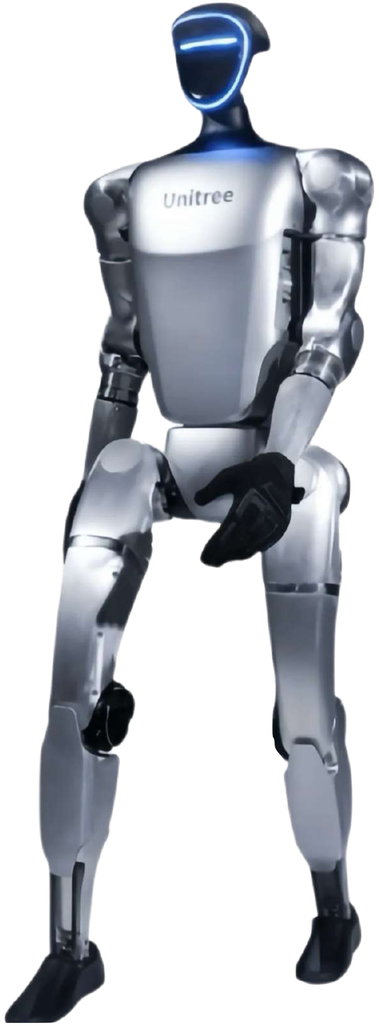
Automated Software Testing

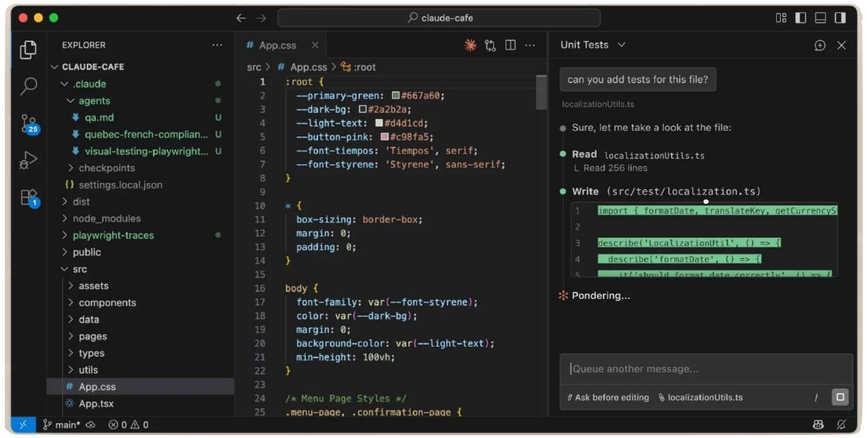
Main domain areas

- **Traditional Software Systems**
  - Search-based Test Generation
  - Regression Testing (e.g., test prioritization, test minimization)
- **Web Applications**
  - End-to-End GUI Testing
  - Automated Crawling
- **Deep Learning Systems**
  - Test Generation
  - Failure Prediction (black-box, white-box)
- **Cyber Physical Systems**
  - Reality Gap Assessment and Mitigation
  - Automotive, UAVs, Elevators
- **LLM Systems**
  - Test Generation
  - Conversational Systems, UAVs, Elevators



AI-based Systems





1982: What if we could build an intelligent car?

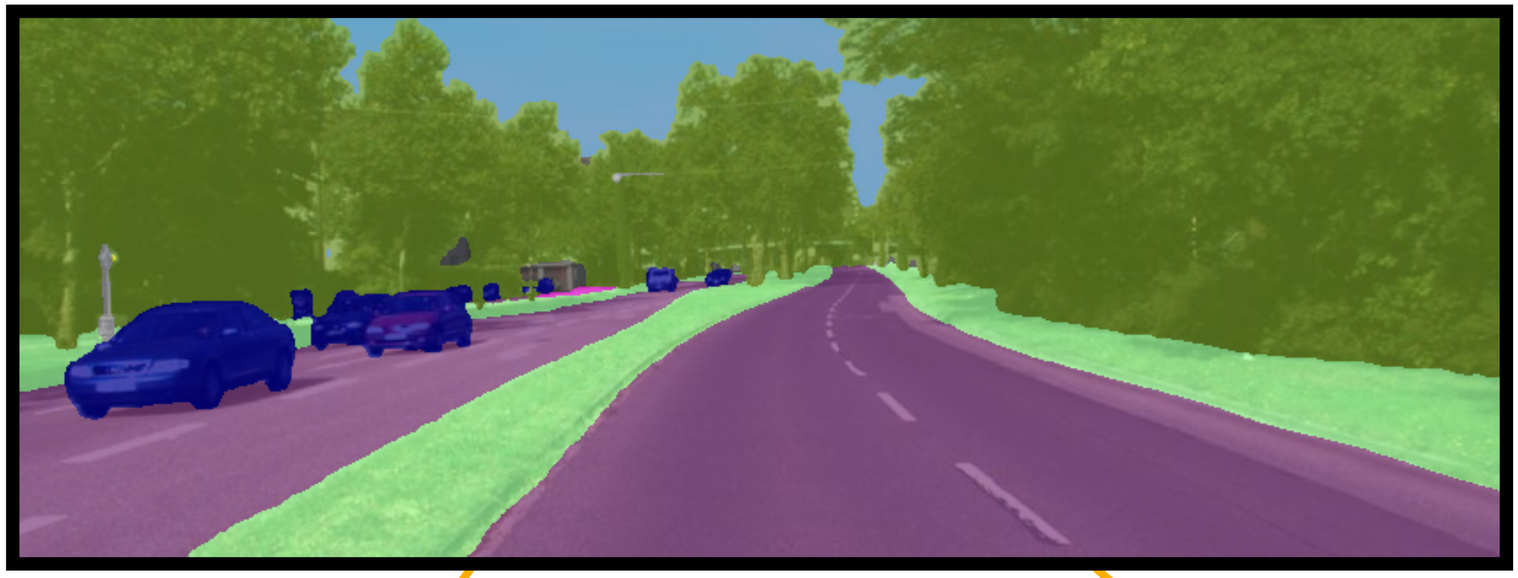
2026: We can. Now, how can we trust it?

Camera/LiDAR/Radar

AI

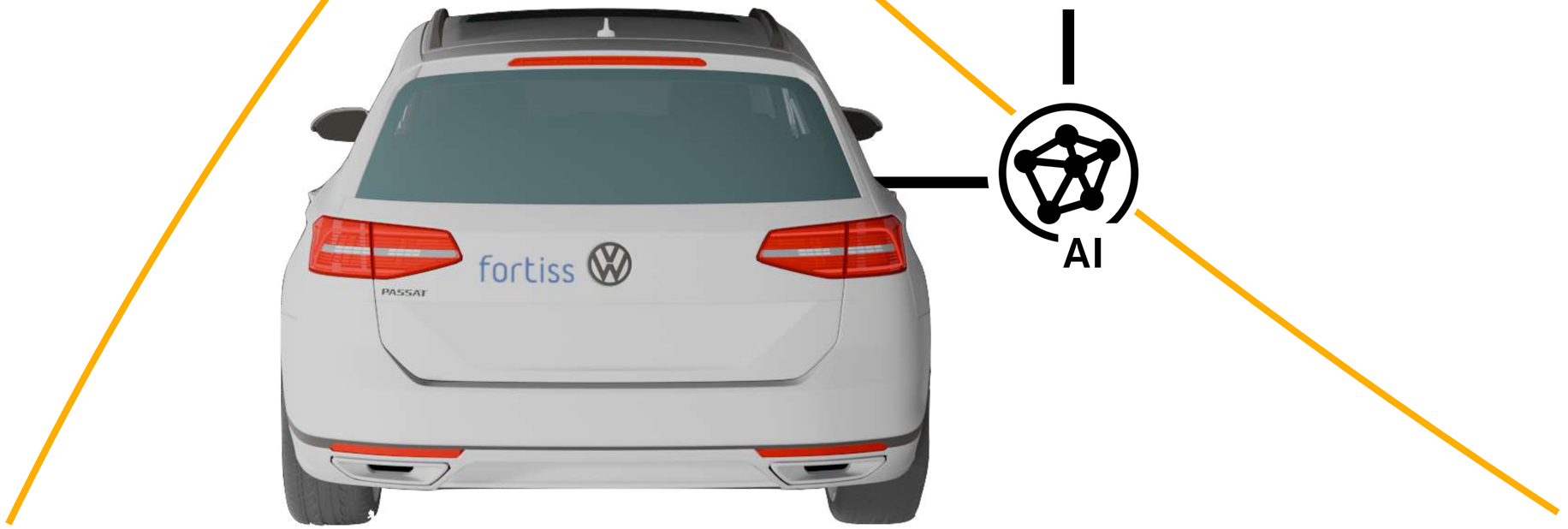
Deep Neural Networks

Visual Language Models



MODULAR

Segmentation



MODULAR

Segmentation

END-2-END

Lane-keeping +
Adaptive Cruise Control

AI

Steer
0.1

Throttle
0.3

LEVEL OF DRIVING AUTOMATION

0

NO AUTOMATION

Manual control. The human performs all driving task (steering, acceleration, braking etc.).

1

DRIVER ASSISTANCE

The vehicle features a single automated system (e.g. it monitors speed through cruise).

2

PARTIAL AUTOMATION

ADAS. The vehicle can perform steering and acceleration. The human still monitors all tasks and can take control at any time.

3

CONDITIONAL AUTOMATION

Environmental detection capabilities. The vehicle can perform most driving tasks, but human override is still required.

4

HIGH AUTOMATION

The vehicle performs all driving tasks under specific circumstances. Geofencing is required. Human override is still an option.

5

FULL AUTOMATION

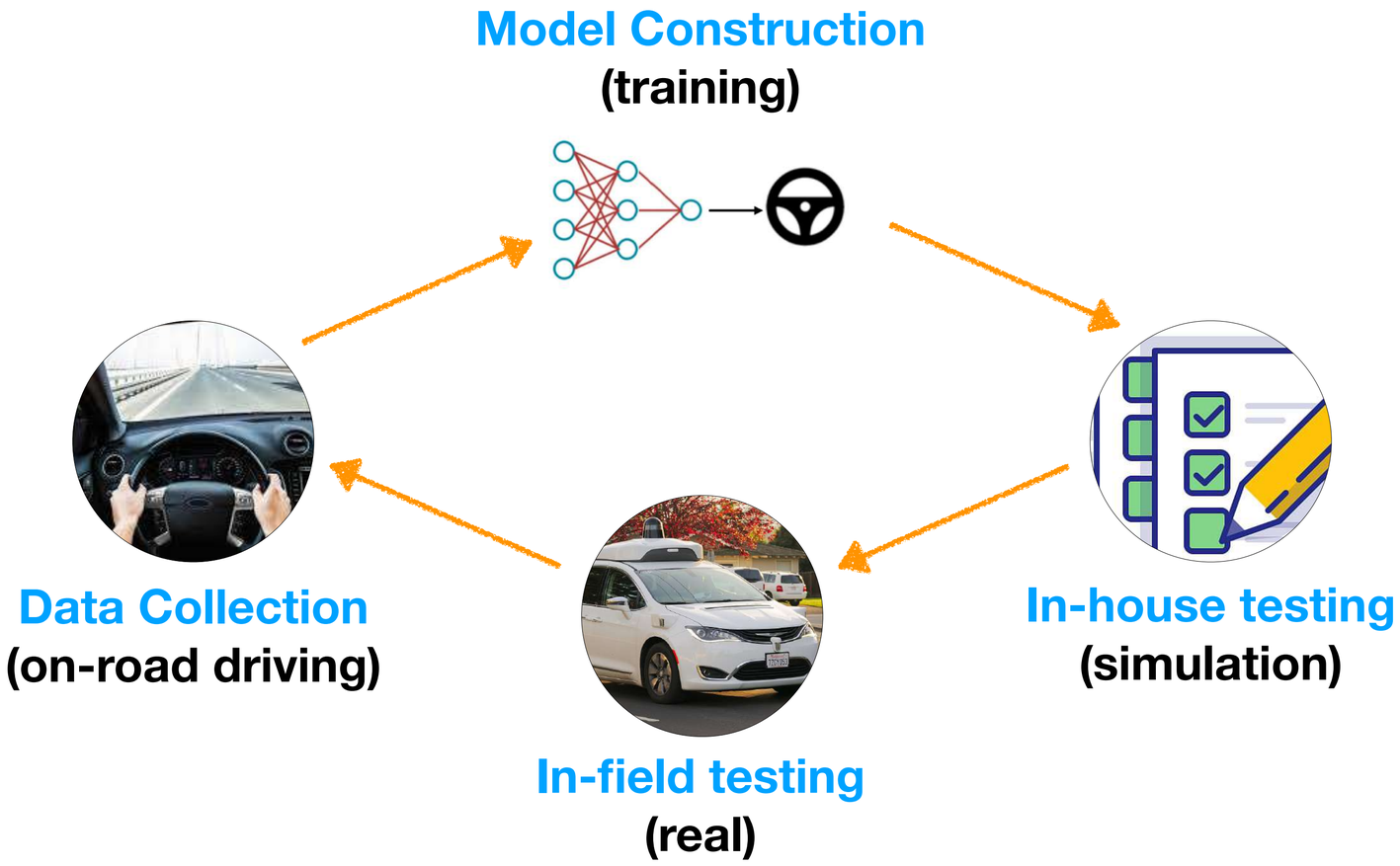
The vehicle performs all driving tasks under all conditions. Zero human attention or interaction is required.

The human monitors the driving environment

The automated system monitors the driving environment

Automated Driving System (ADS) testing

How to ensure that an ADS system is ready for deployment?



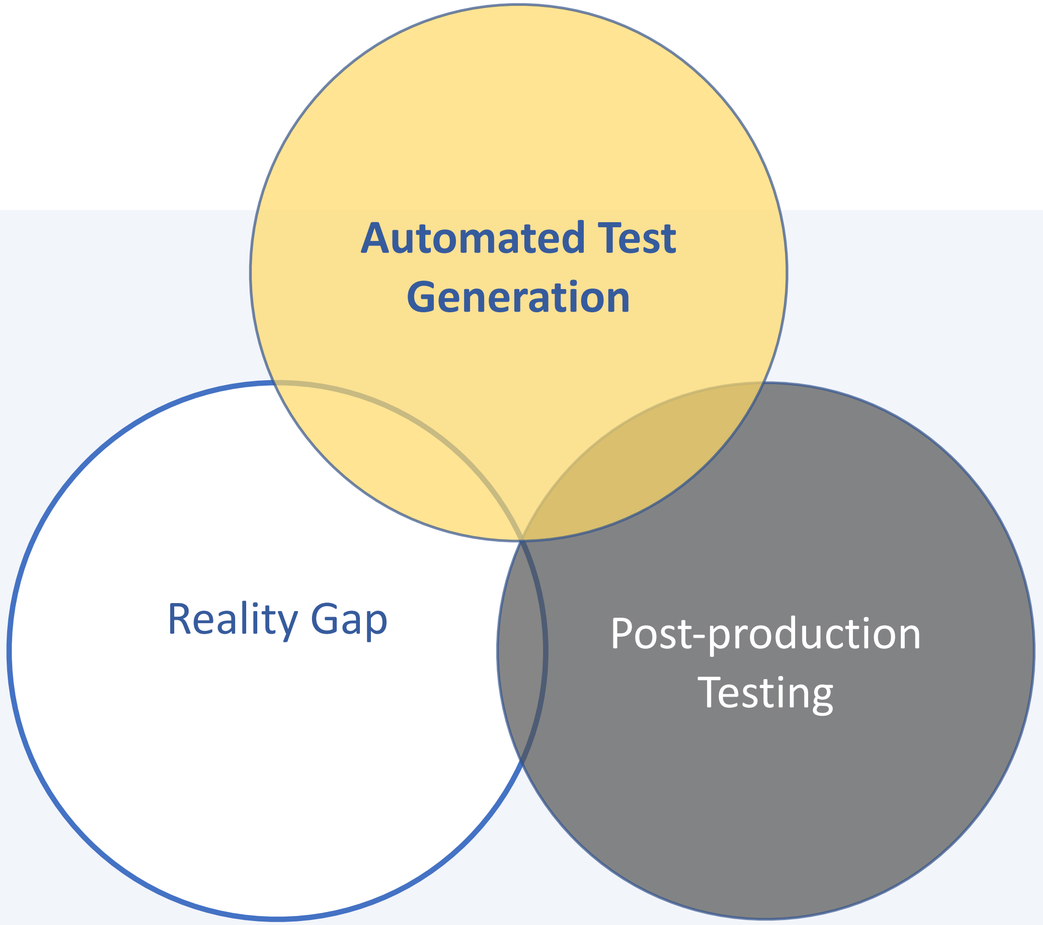
ADS Testing

Main research topics

- **Automated Test Generation**
  How can we automatically generated complex scenario-based tests efficiently and effectively?

- **Reality Gap**
  How can we measure and mitigate the reality gap between simulation and in-field testing?

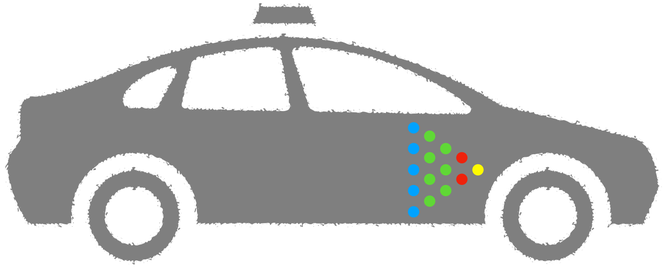
- **Post-production Testing**
  How to ensure a high dependability of deep neural network driven-cyber-physical systems (CPS) in production?



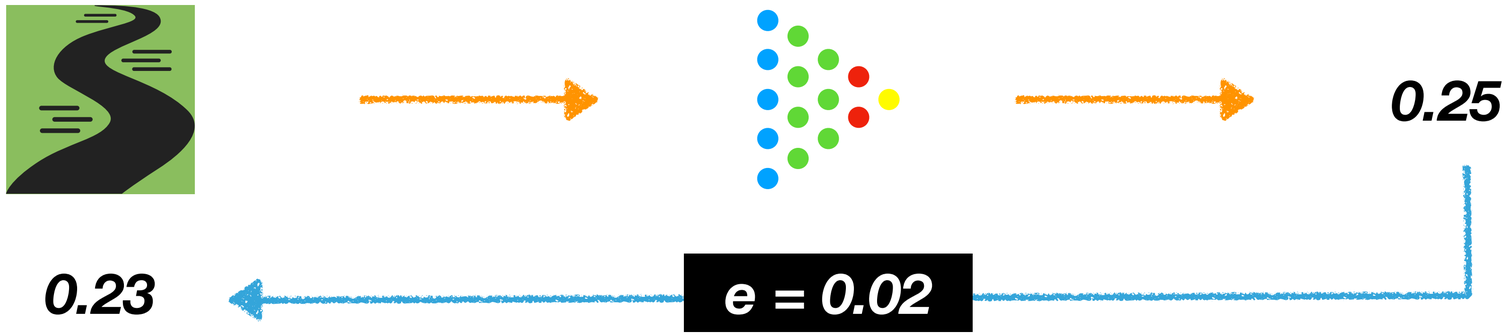
Model vs System level Testing

Model

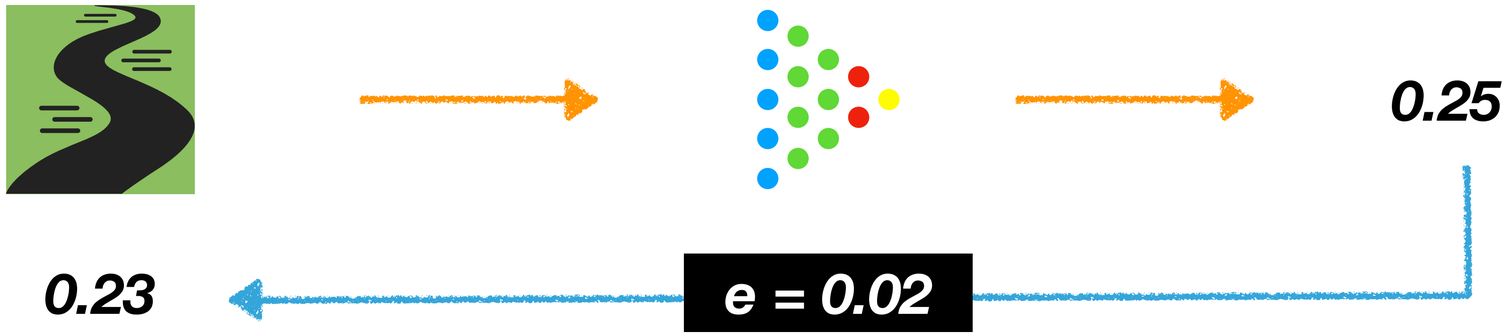
System



Model level Testing



Model level Testing



- Quick Feedback
- Cheap

- No driving
- Simple Oracle

Model level Testing Literature

DeepXplore
Pei et al., 2017


(a) Input 1

(b) Input 2 (darker version of 1)

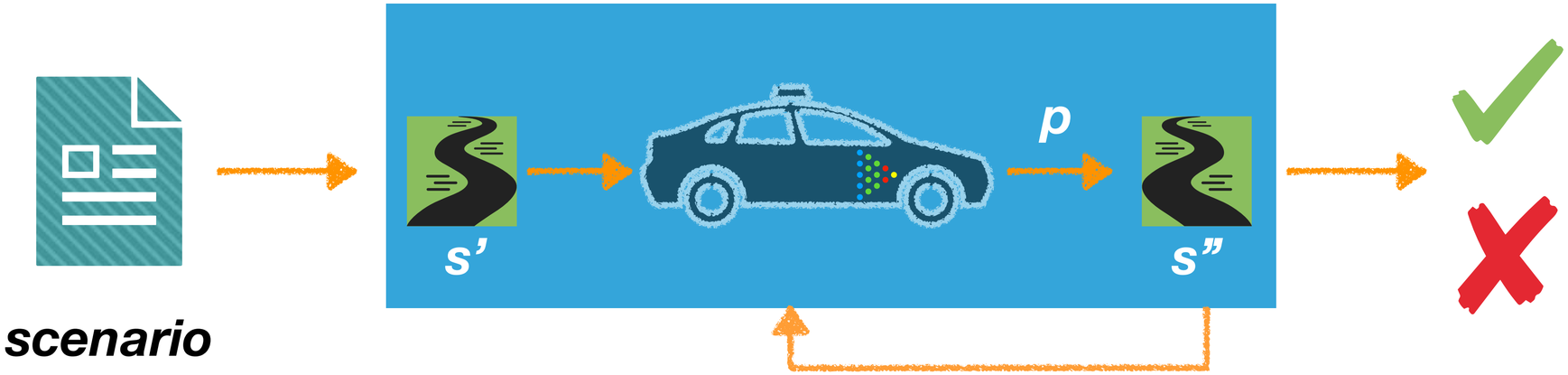
DeepTest
Tian et al., 2017

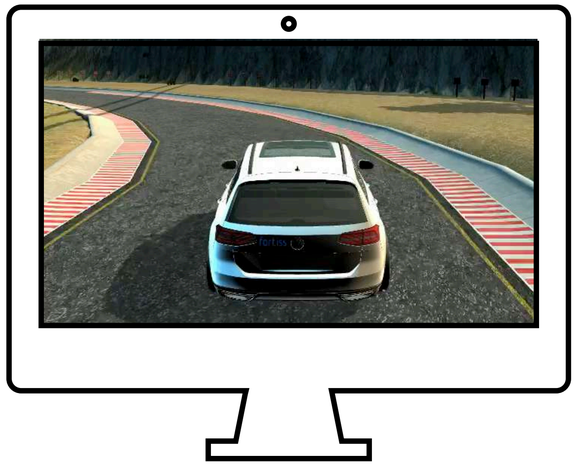
original

brightness(50)

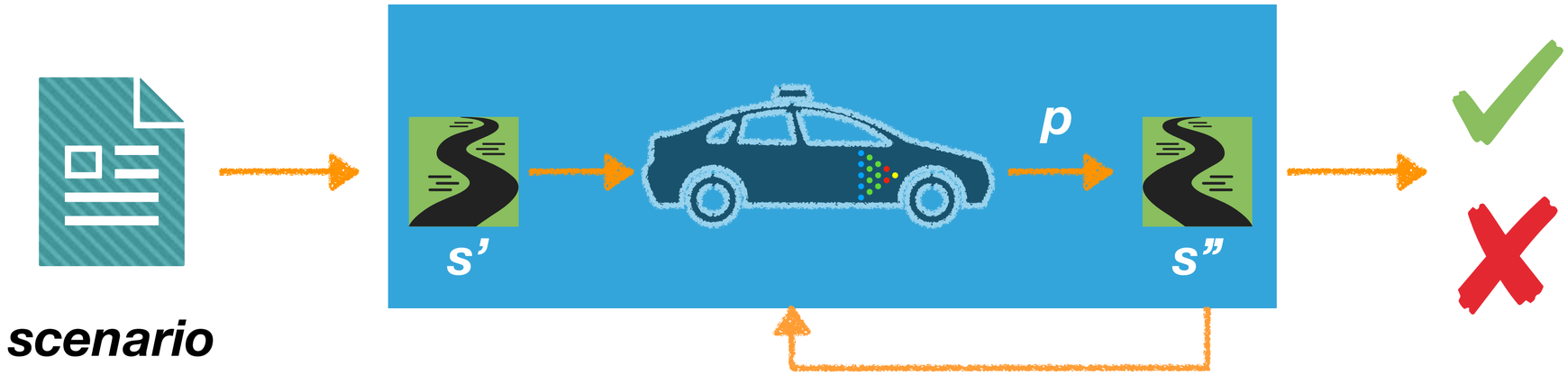
DeepRoad
Zhang et al., 2018

System level Testing





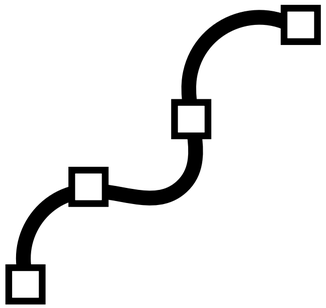
System level Testing



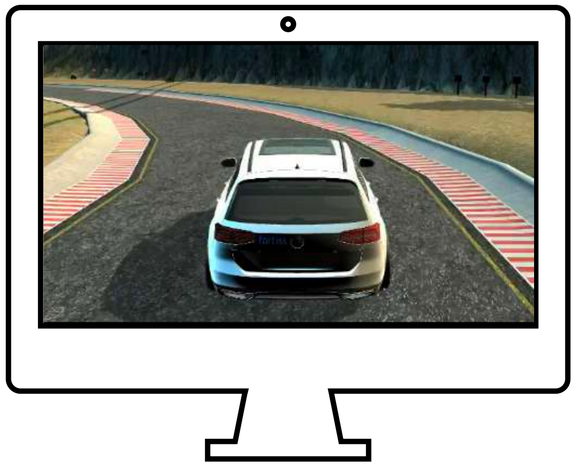
- Driving
- More realistic

- Time consuming
- Complex Oracle

Test Scenarios for ADS



Lane-keeping Driving Assistance System



Modular or End-to-End Driving System

Testing for Robustness

PerturbationDrive

33 perturbations
5 intensities

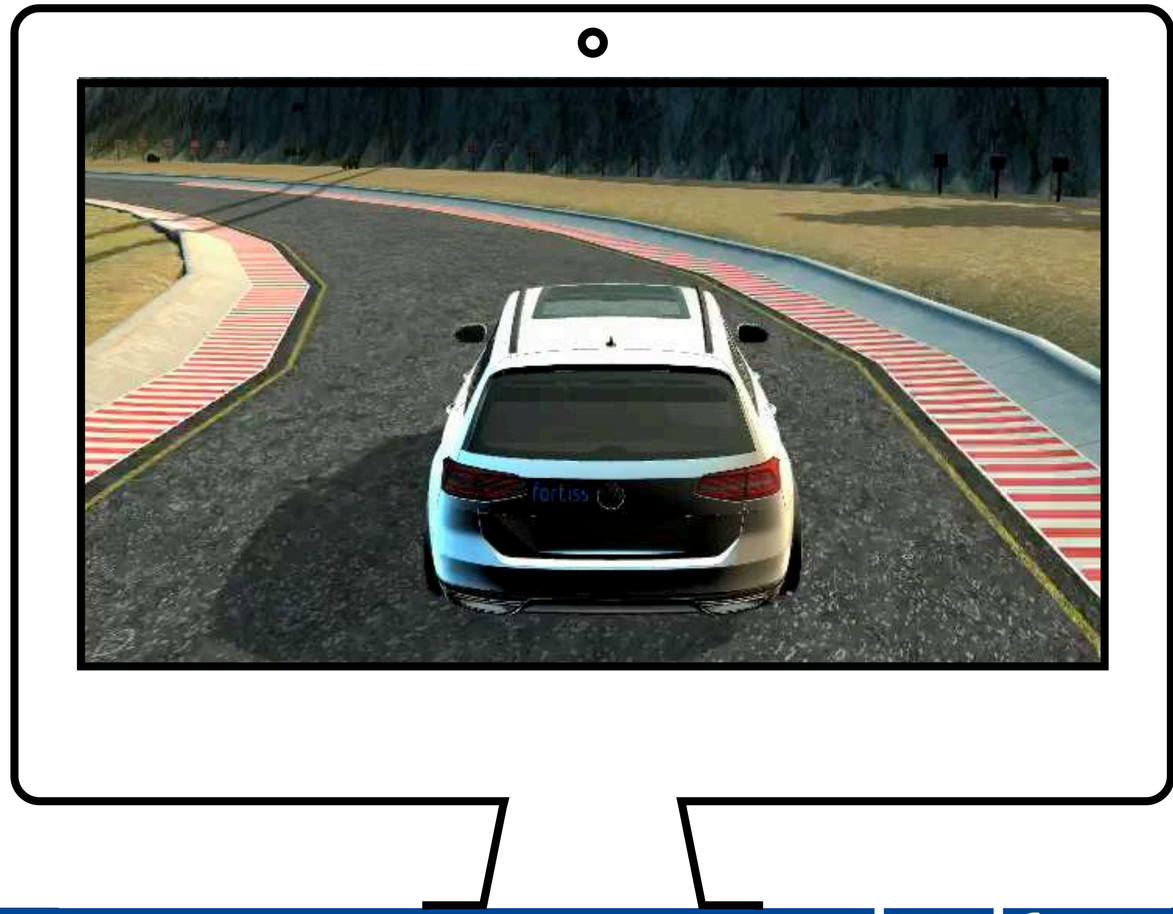
Model & system level

Multiple simulators

```
> apply("Phase scrambling",intensity=4)
```

PerturbationDrive

AI



RQ1

Most effective perturbations

SIM 1
Small DT

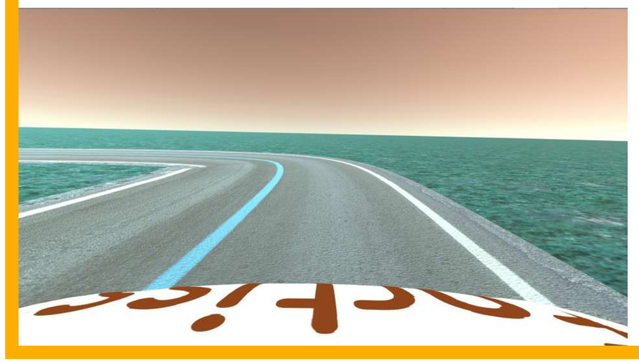
LK
ACC

Phase scrambling

↓54%
Success rate

SIM2
Udacity

False color

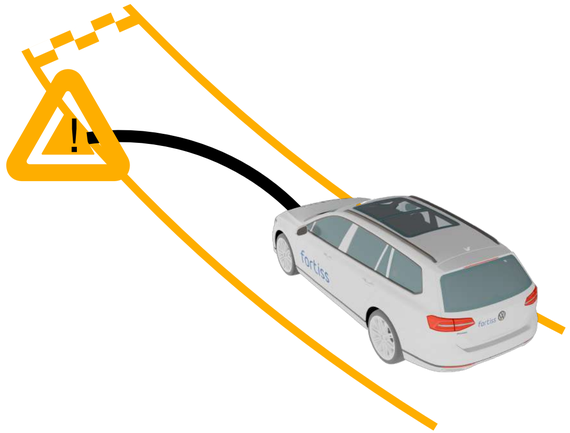


↓76%
Success rate

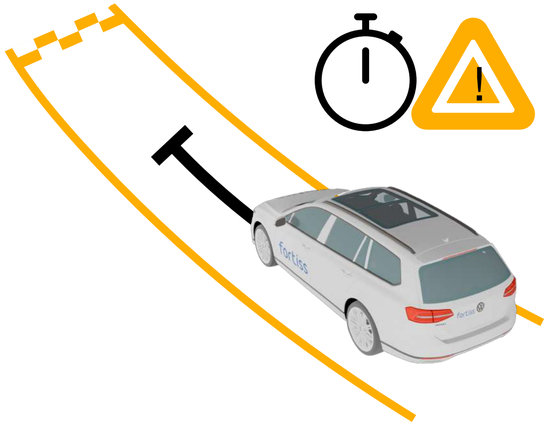
RQ1

LK ACC

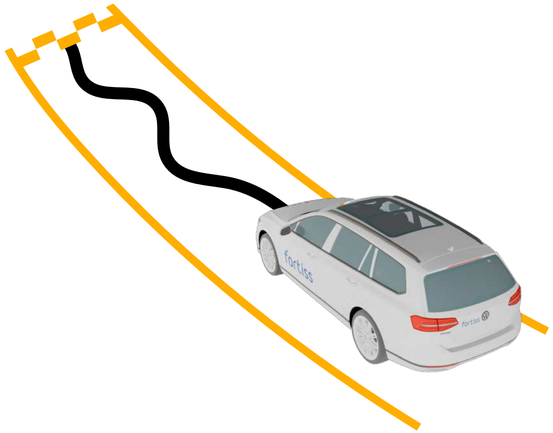
SIM2
Udacity



Out of road



Out of time

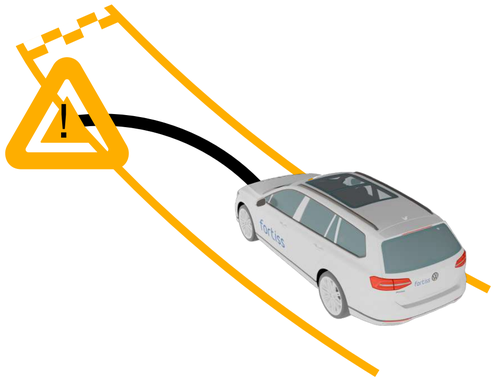


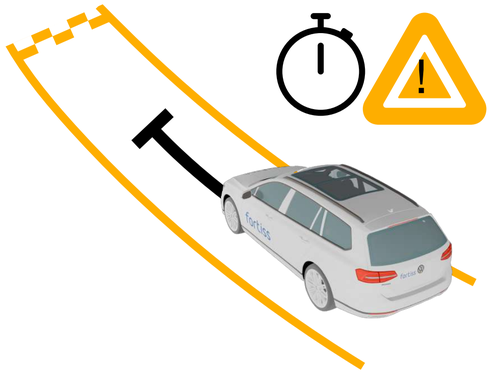
Jitter

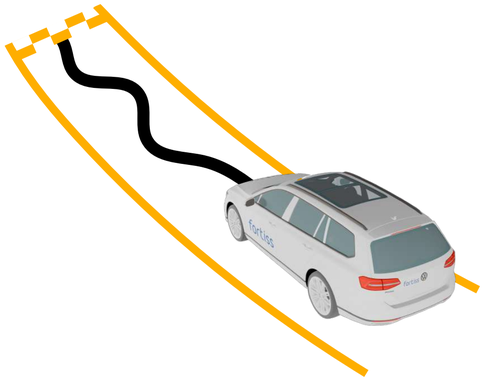
RQ1

LK ACC

SIM2 Udacity

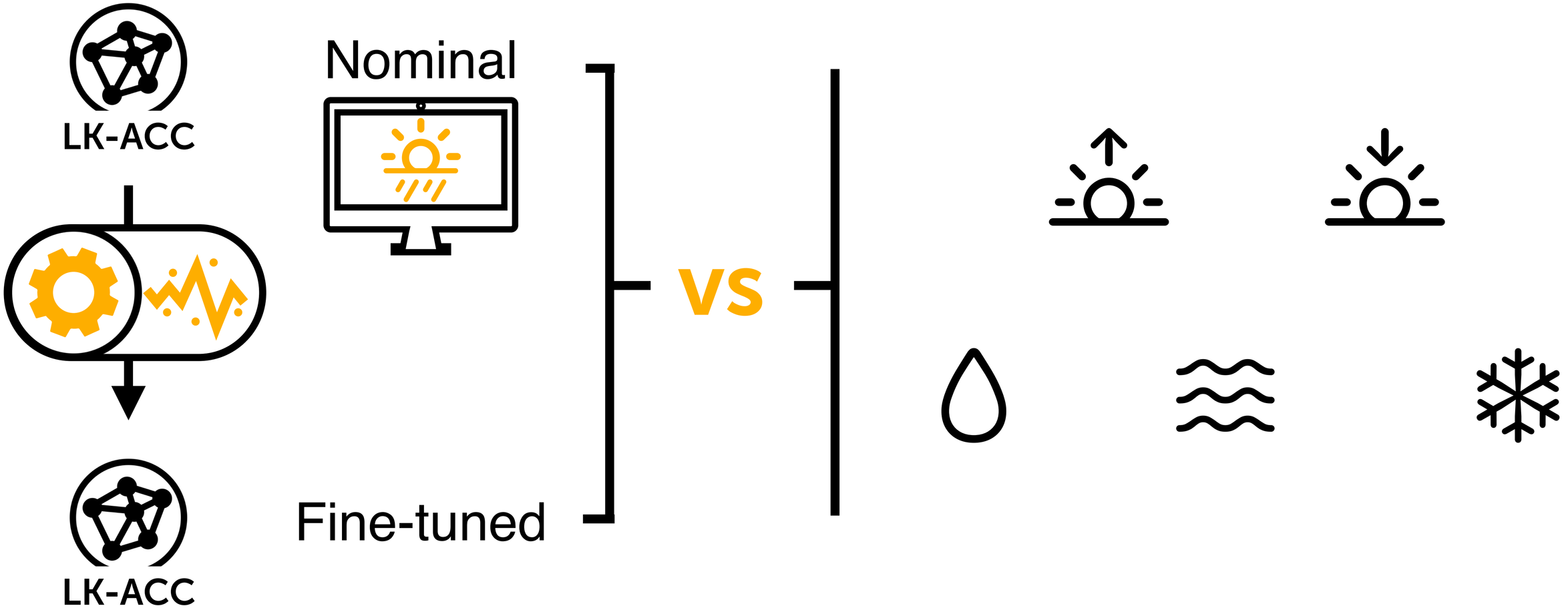




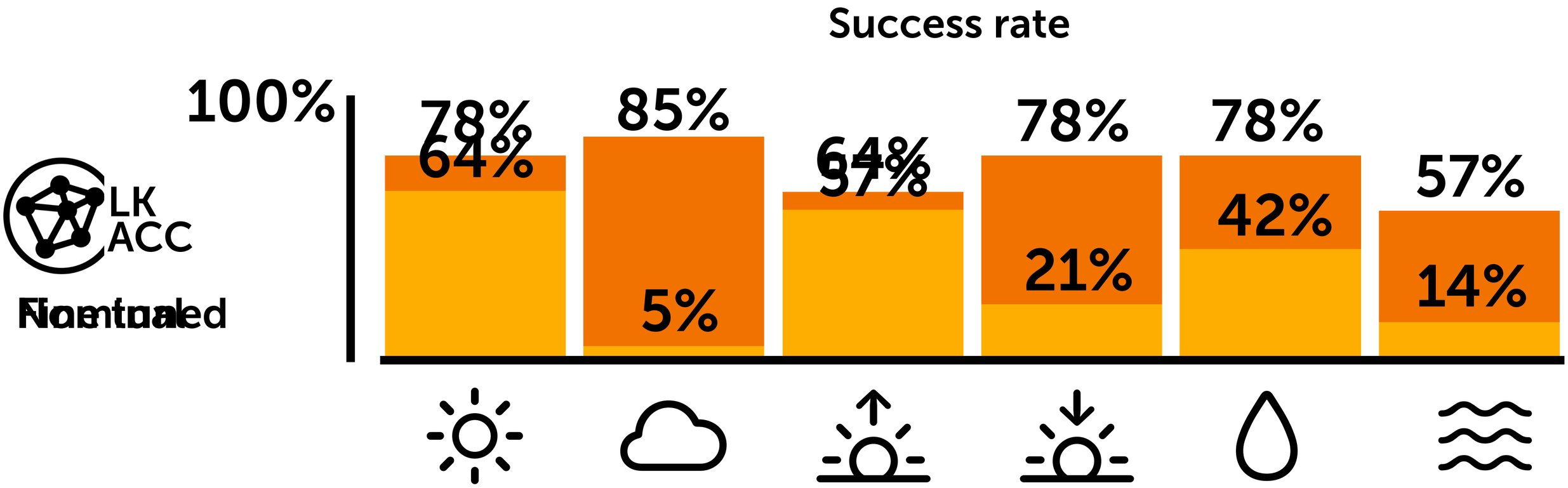


		Out of road	Out of time	Jitter
False color	↓76% Success rate	+38	+0	↑530%
Jpeg artefacts	↓34% Success rate	+1	+16	↑130%
Saturation decrease	↓48% Success rate	+15	+9	↑621%

RQ2 Generalisation

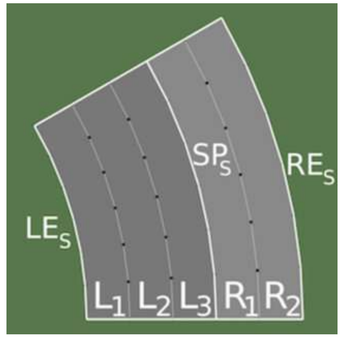


RQ2

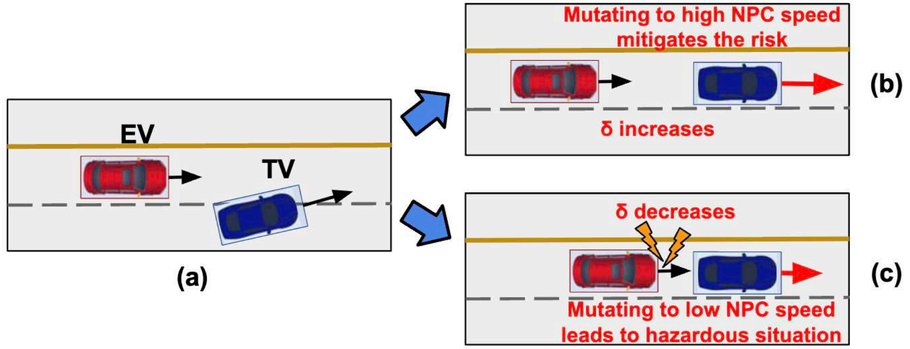


Testing for Generalization

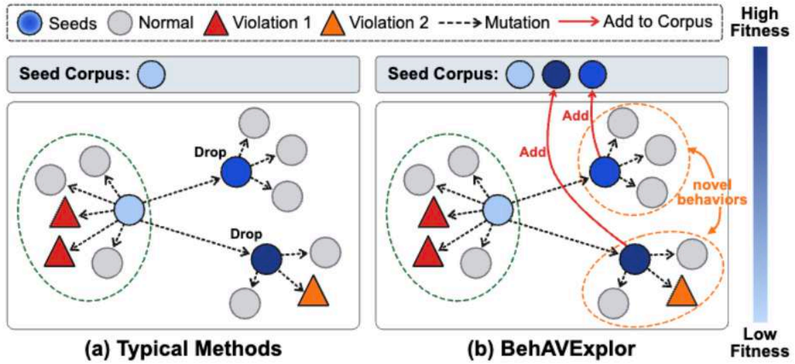
AsFault
Gambi et al., 2019



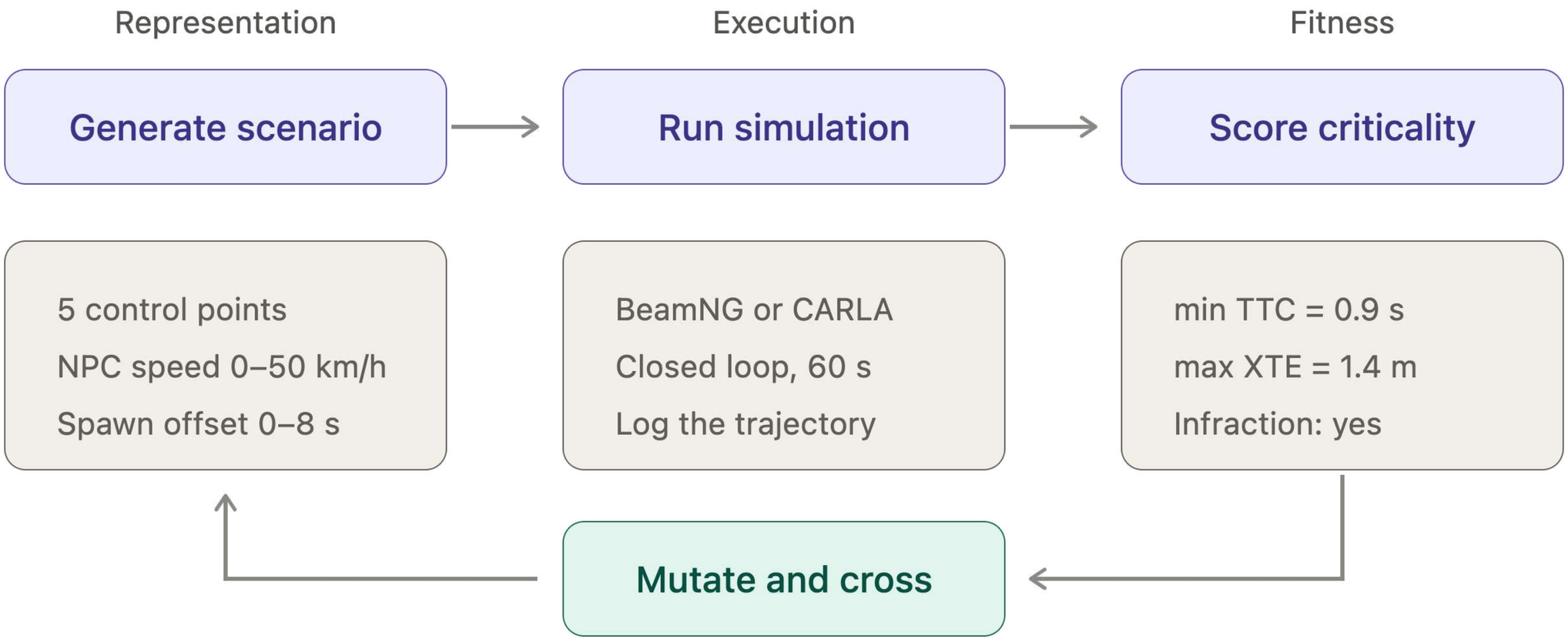
AVFuzzer
Li et al., 2020



BehAVExplor
Cheng et al., 2023



Search-based Test Generation



Repeat until the budget is spent, then return the most critical scenarios found

Chair of Software Engineering for data-intensive Applications
School of Computation, Information and Technology
Technical University of Munich

OpenSBT

Open Search-based Testing

ADS & ADAS

- **End-to end DNNs (level 2)**
- **Full AD stacks**

Critical Test Scenarios Identification

High fidelity sim-based Validation

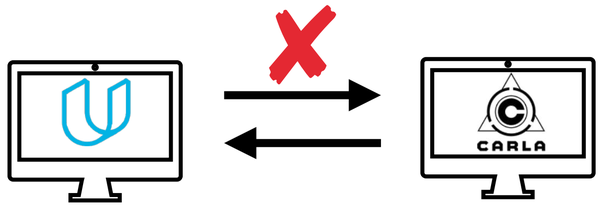
CARLA UDACITY BeamNG

https://git.fortiss.org/opensbt

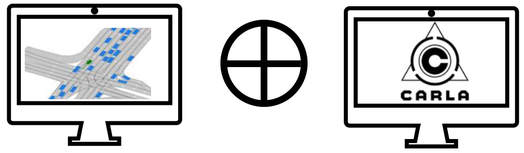
Sorokin et al. OpenSBT: A Modular Framework for Search-based Testing of Automated Driving Systems.

Automated Test Generation

Digital Siblings (EMSE 2025)



SimFusion (under review)

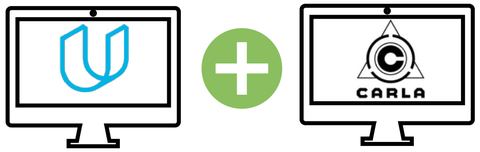


Misbehaviour Forecasting for Focused Testing (ICSE 2026)

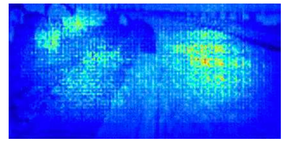
Validity

Efficiency

Simulator Ensembles (EMSE 2026)

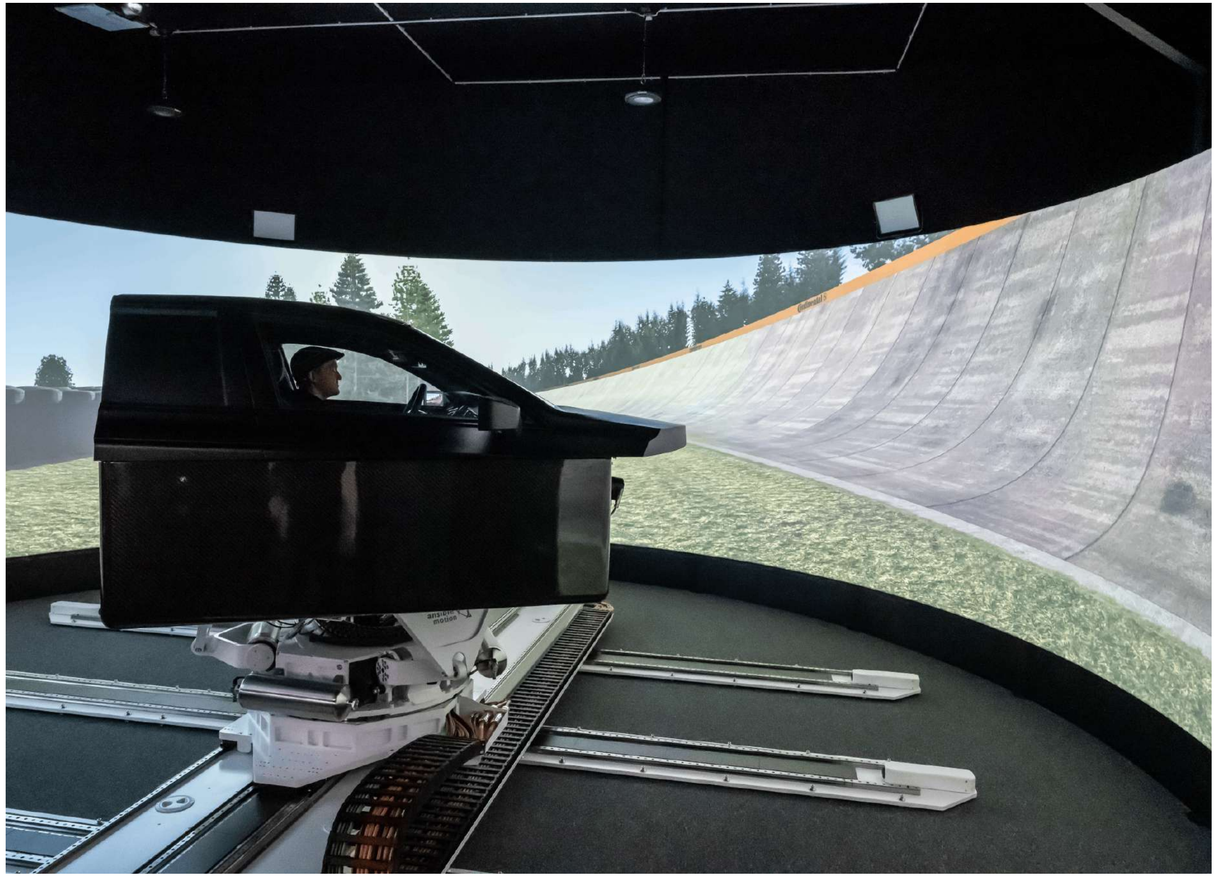


XAI-based Test generation (EMSE 2026)

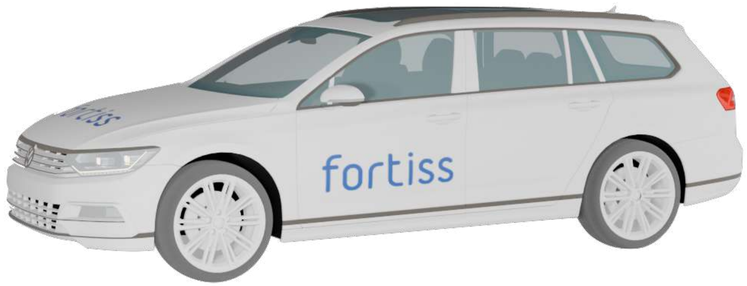


How to increase validity?

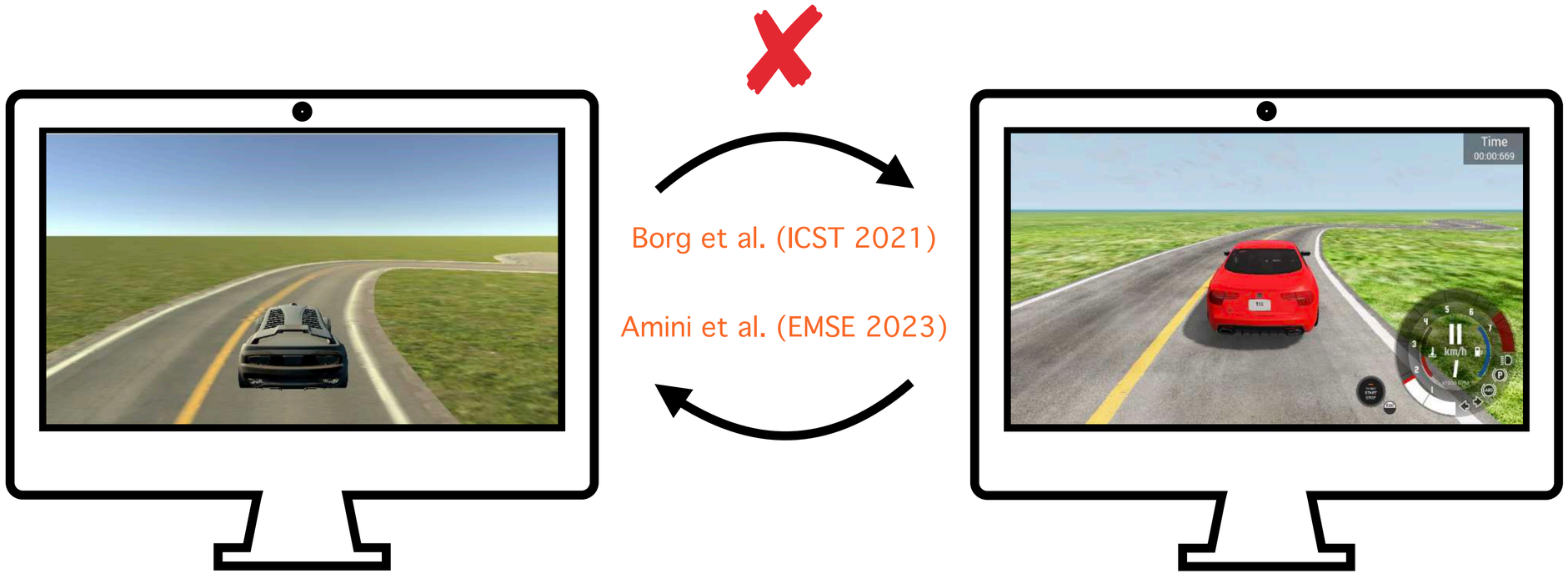
Real vehicles?



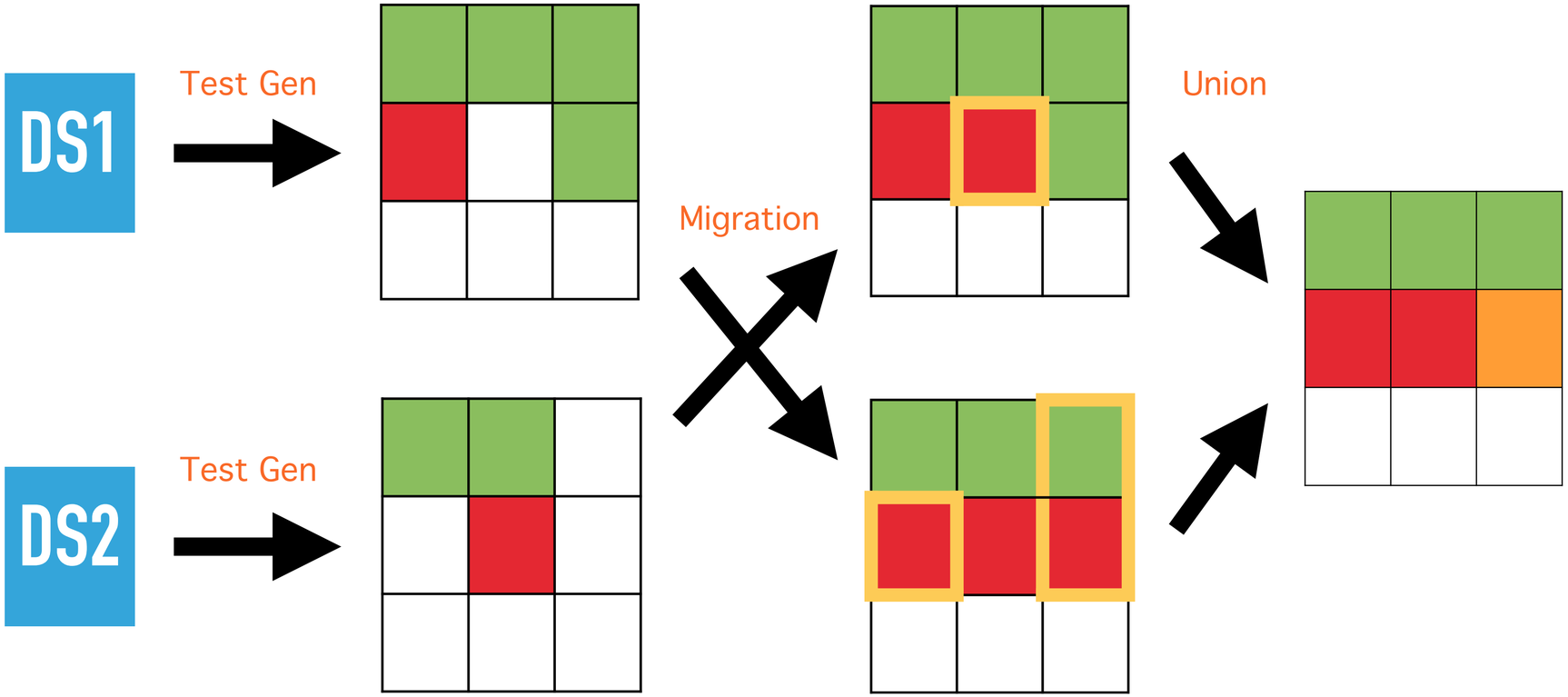
Digital Twins



General-purpose simulators

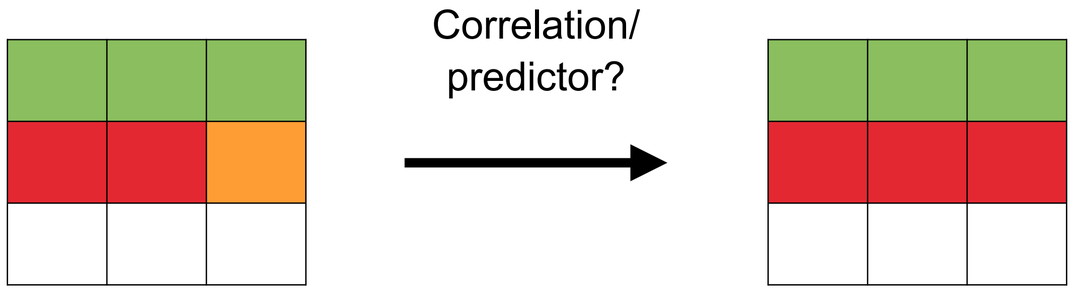


Digital Siblings

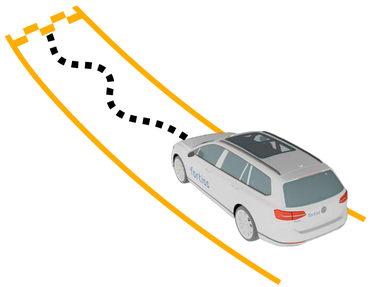


M. Biagiola, A. Stocco, V. Riccio, P. Tonella. Two is Better Than One: Digital Siblings to Improve Autonomous Driving Testing. EMSE

Digital siblings vs digital twin

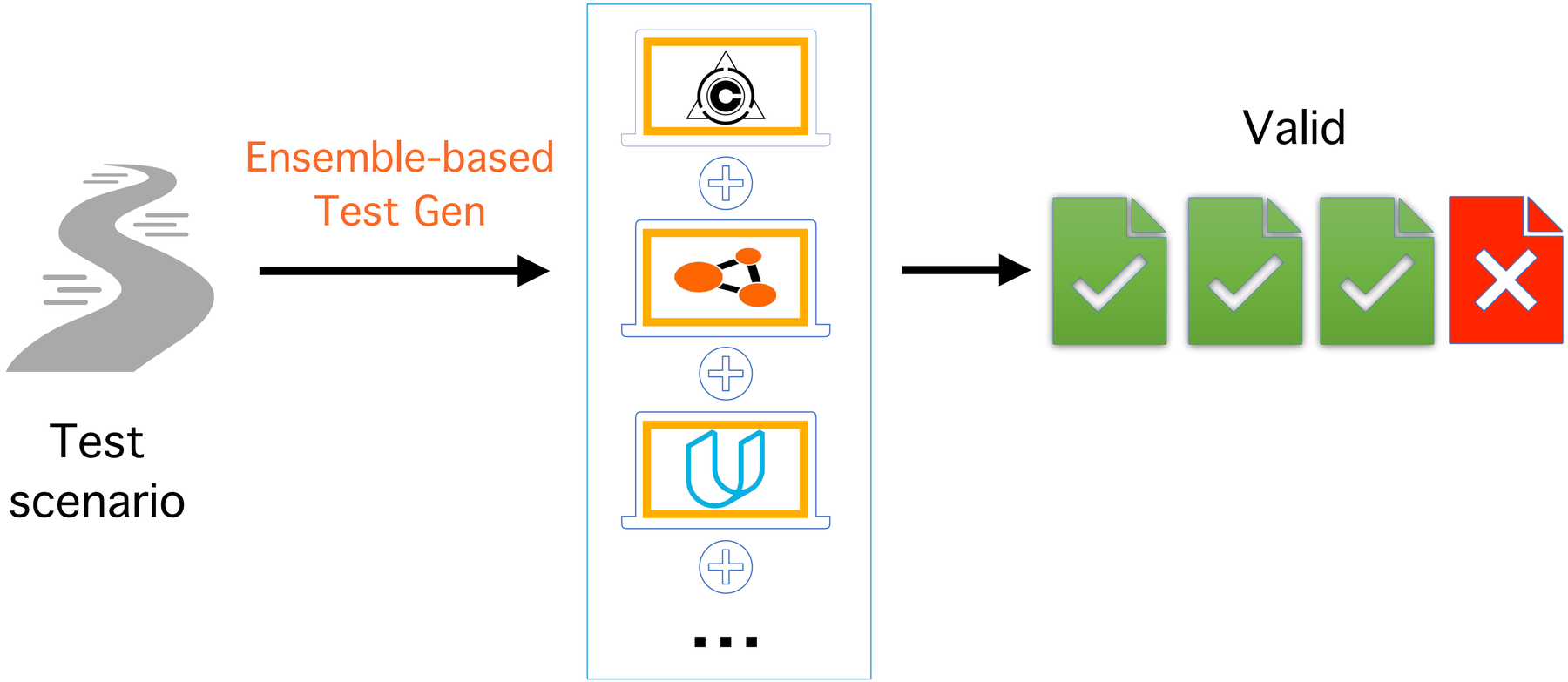


	Dave-2	Chauffeur	Epoch
vs DT	0.459	0.740	0.392
vs DT	0.496	0.791	0.541
vs DT	0.659	0.940	0.594



Chair of Software Engineering for data-intensive Applications
School of Computation, Information and Technology
Technical University of Munich

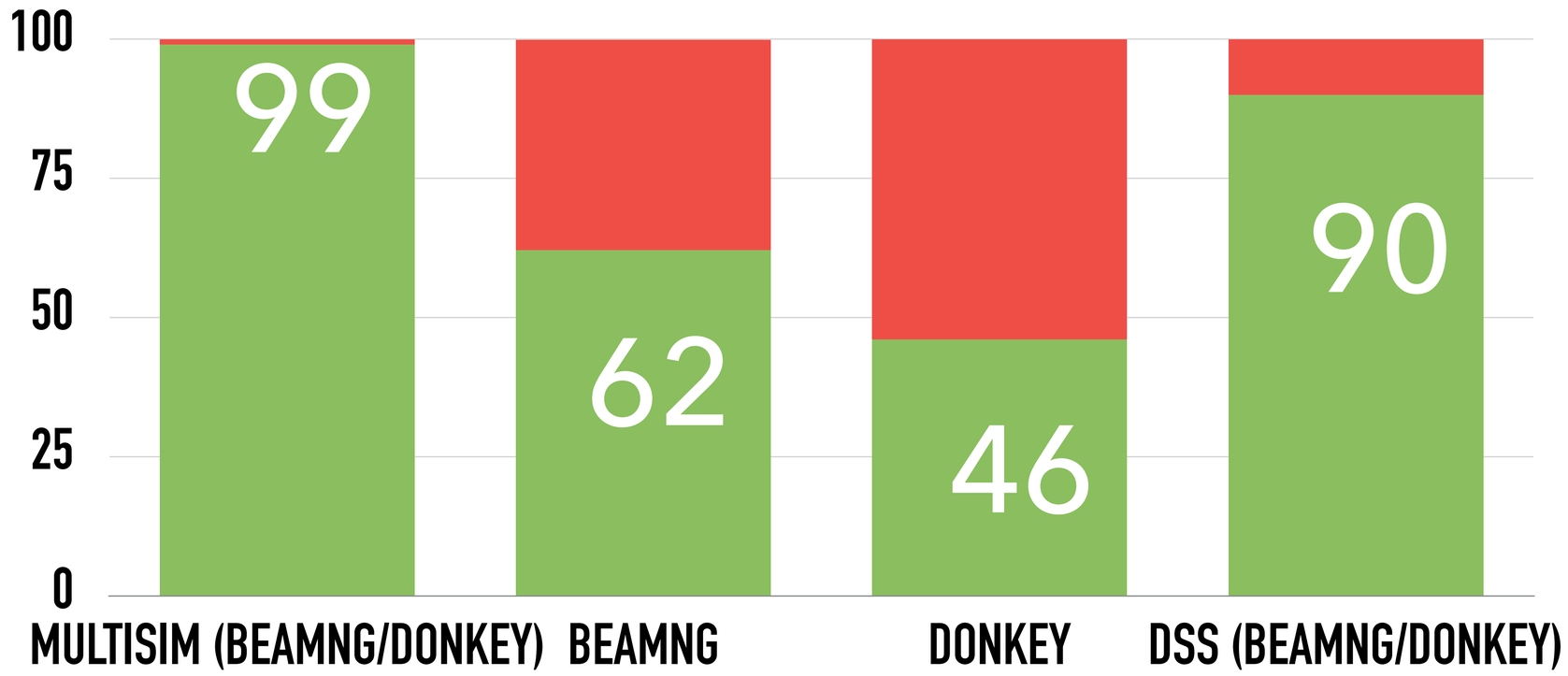
Simulators ensemble



L. Sorokin, M. Biagiola, A. Stocco. Simulator Ensembles for Trustworthy Autonomous Driving Systems Testing. Empirical Software Engineering, 2026

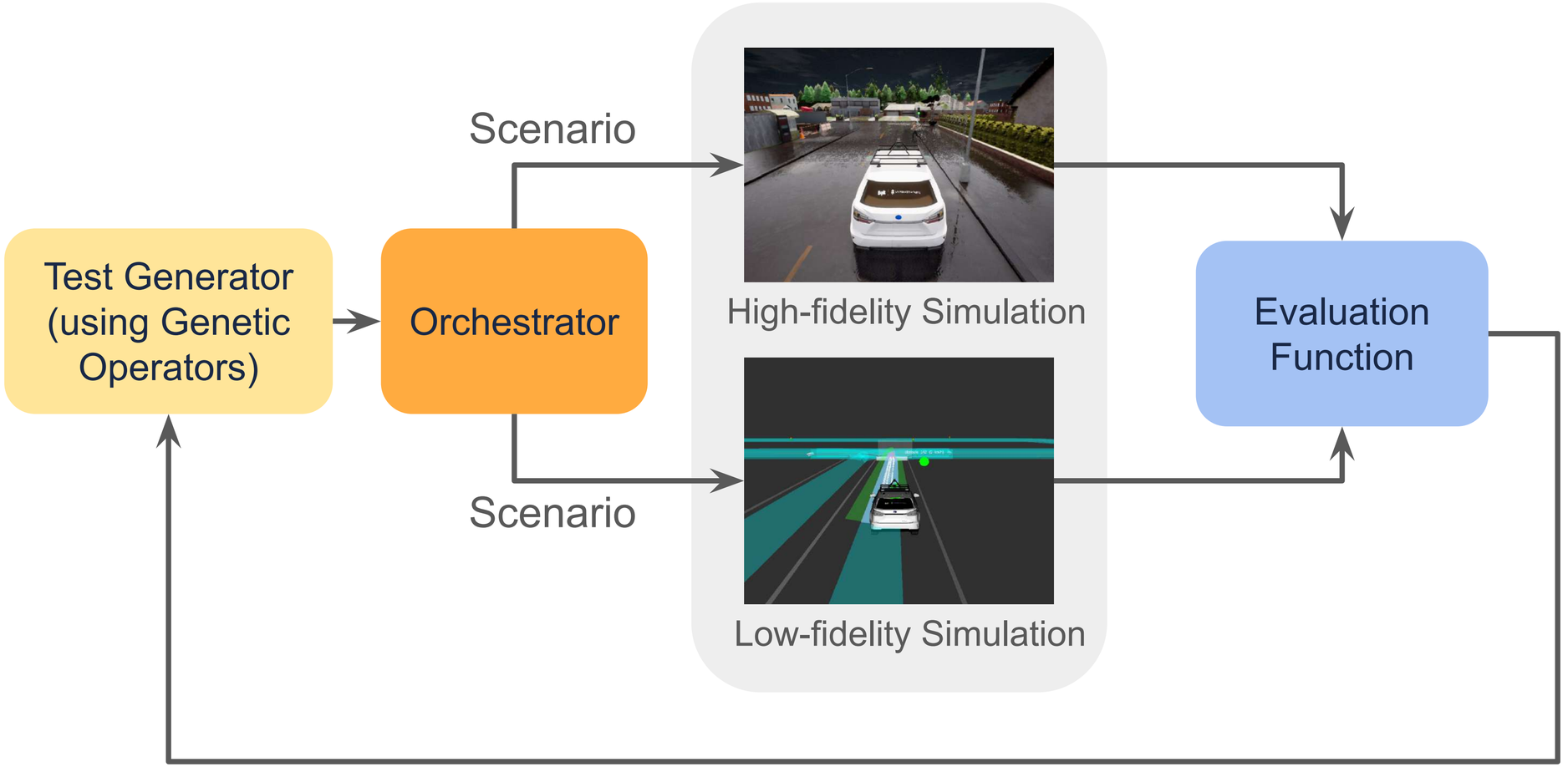
Chair of Software Engineering for data-intensive Applications
School of Computation, Information and Technology
Technical University of Munich

Increased failure validity



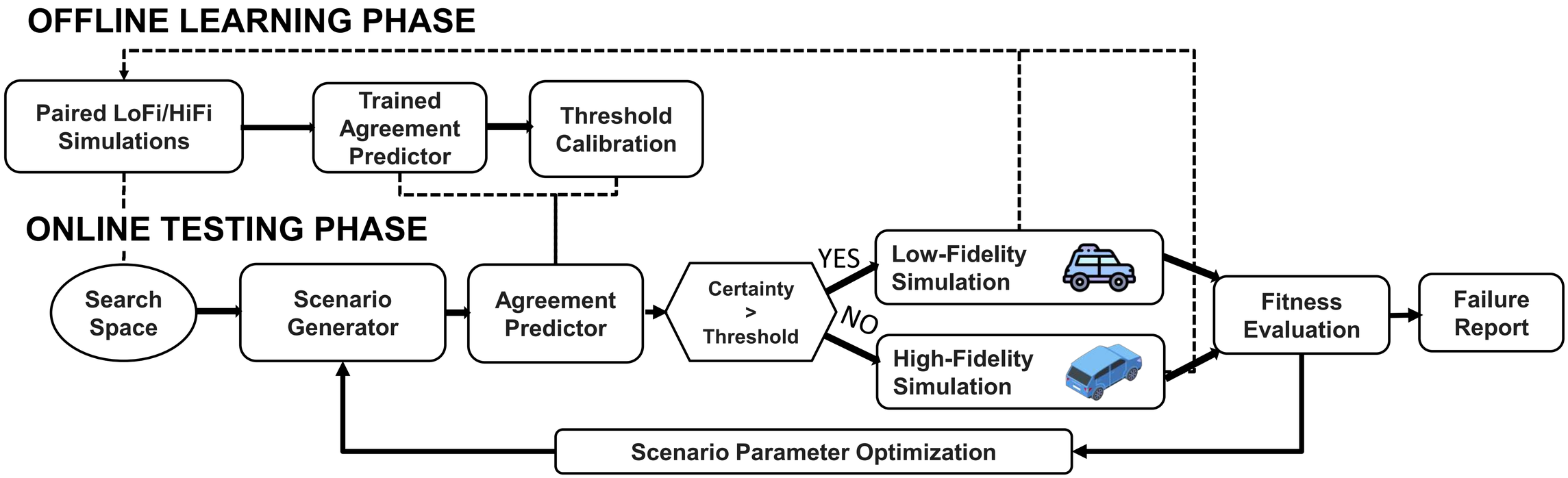
Chair of Software Engineering for data-intensive Applications
School of Computation, Information and Technology
Technical University of Munich

Multi-fidelity simulation



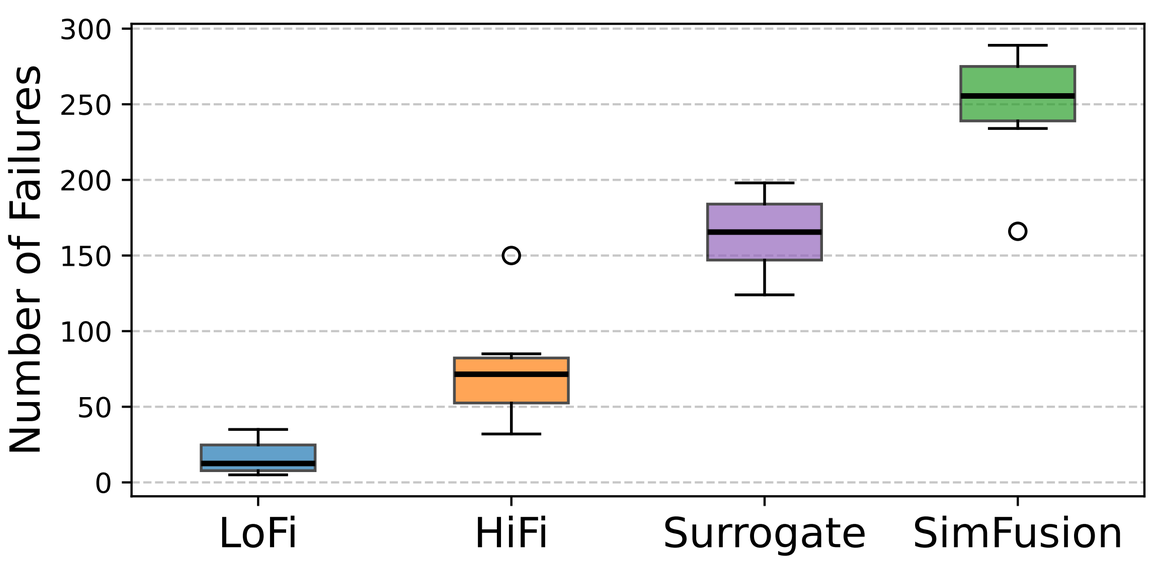
Chair of Software Engineering for data-intensive Applications
School of Computation, Information and Technology
Technical University of Munich

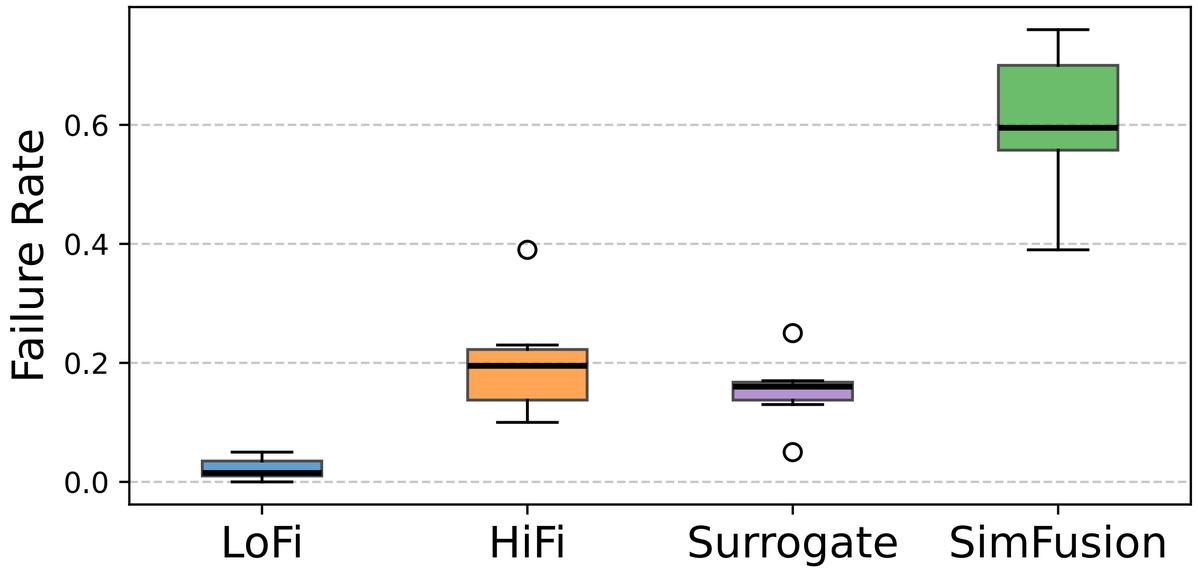
Multi-fidelity simulation



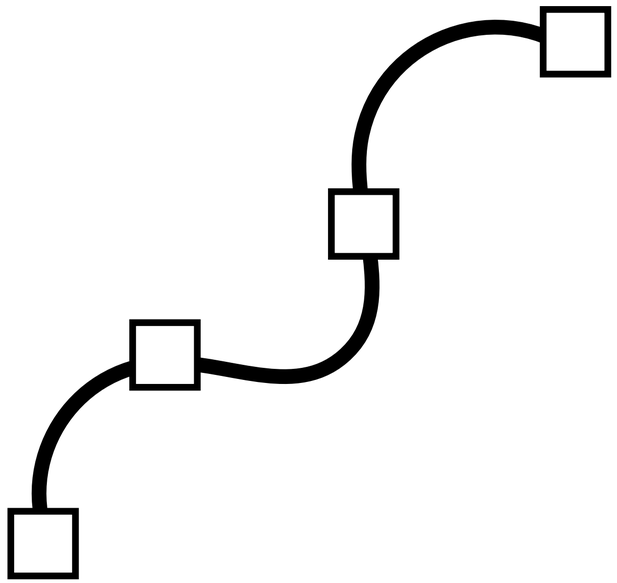
Chair of Software Engineering for data-intensive Applications
School of Computation, Information and Technology
Technical University of Munich

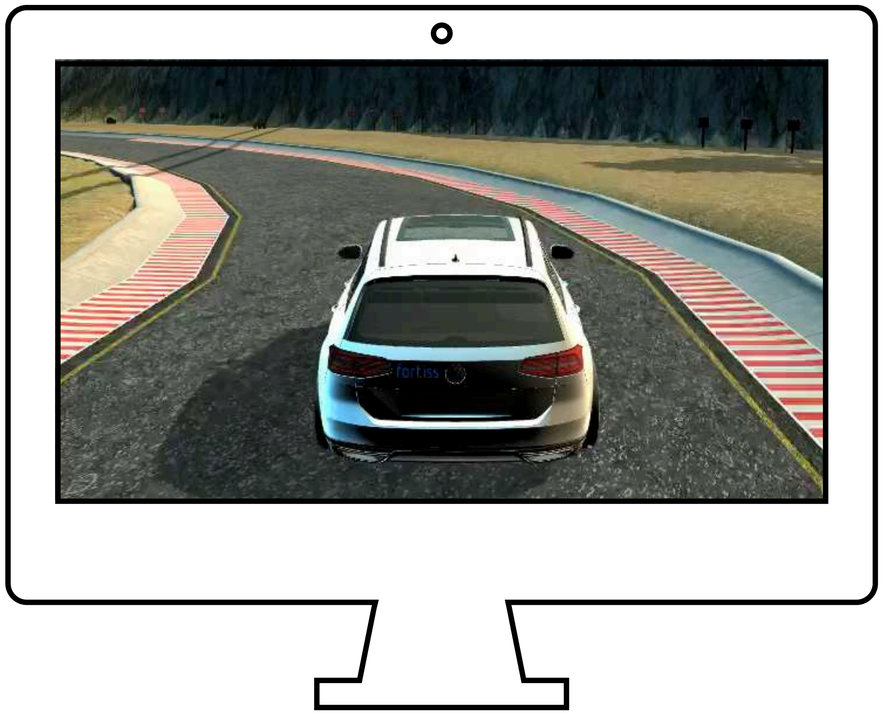
Multi-fidelity simulation



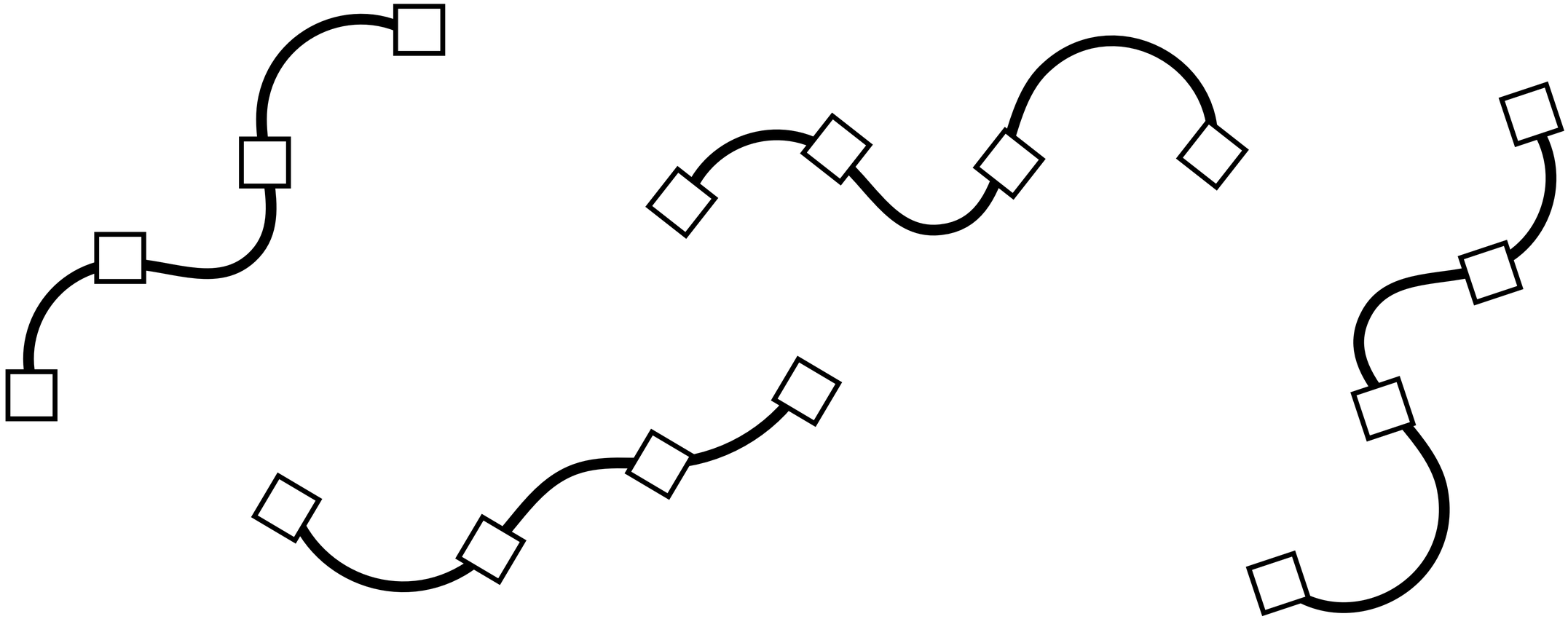


How to increase efficiency?

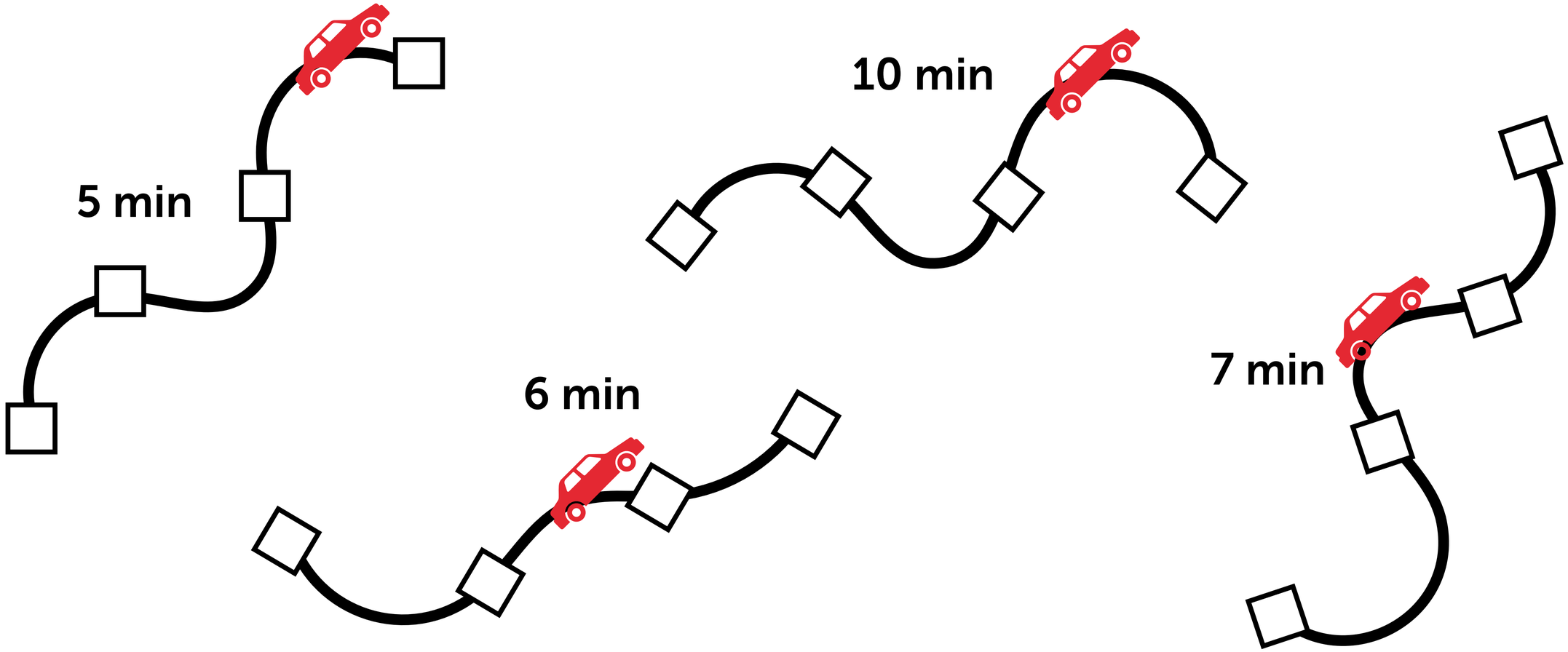




How to increase efficiency?

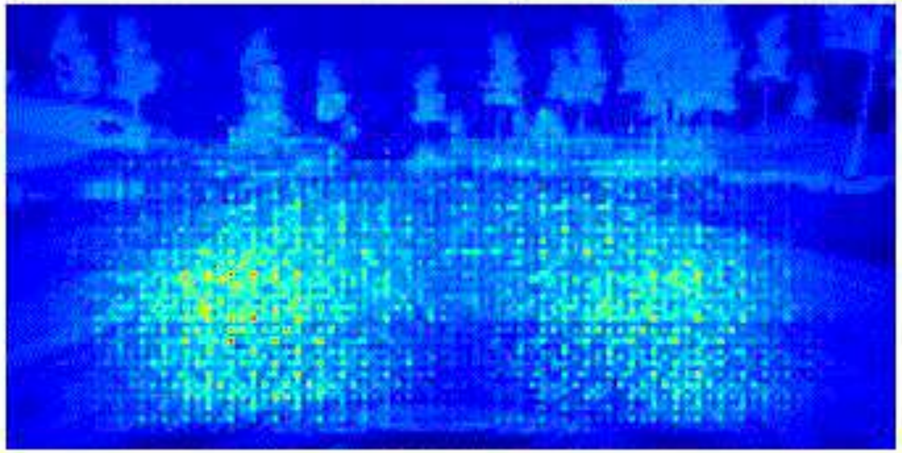


How to increase efficiency?



Nominal

XAI Image

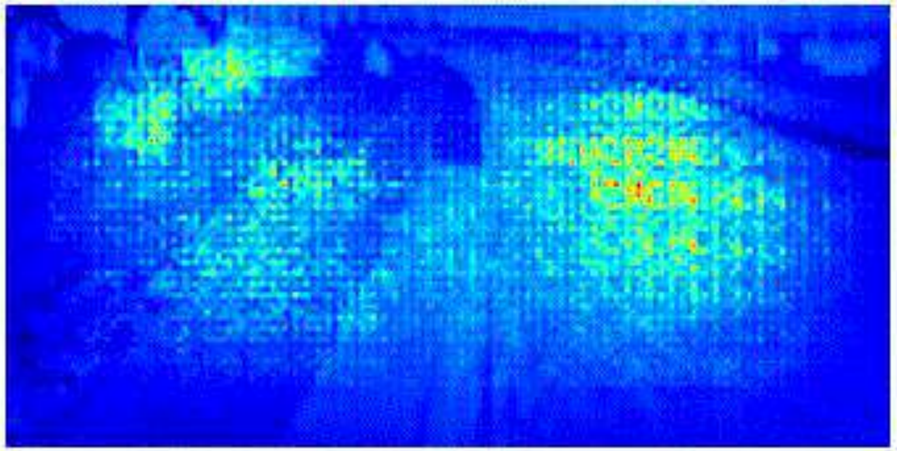


Original Image

Chair of Software Engineering for Data-intensive Applications
School of Computation, Information and Technology
Technical University of Munich

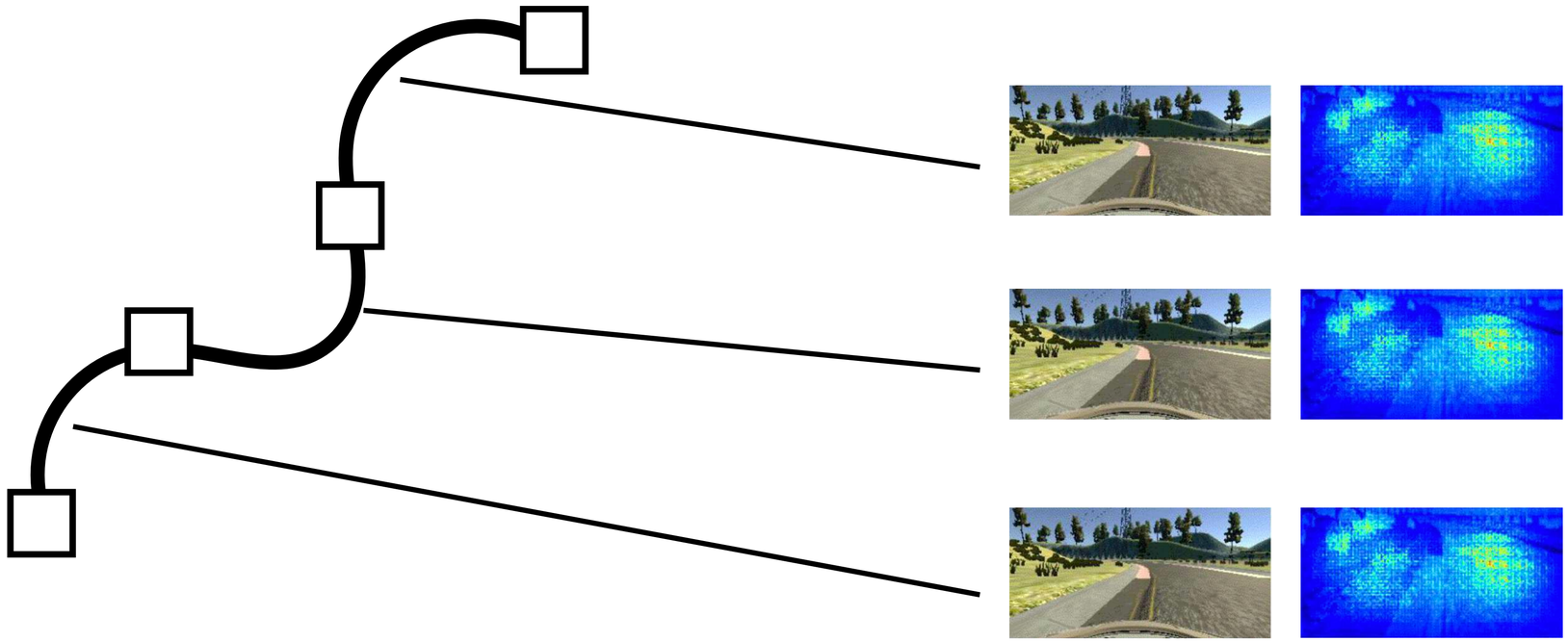
Uncertain/Unexpected

XAI Image

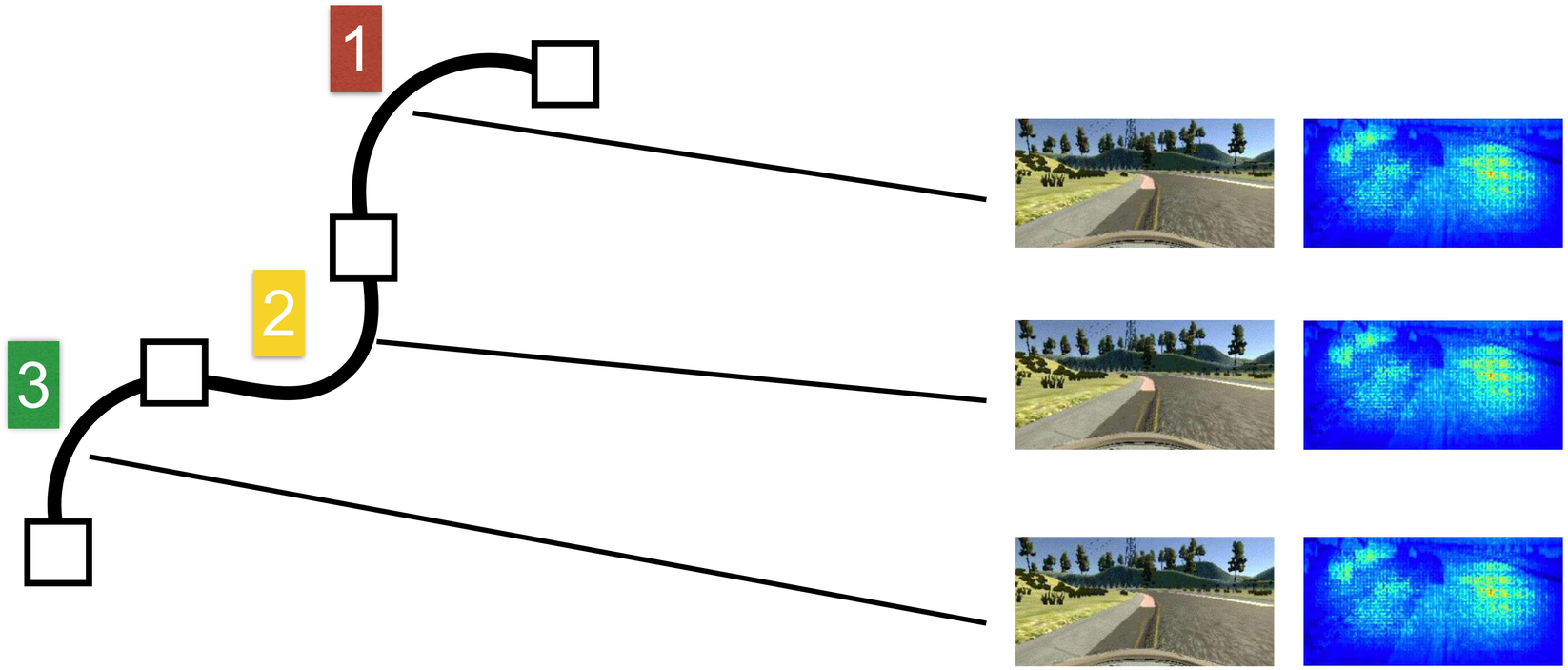


Original Image

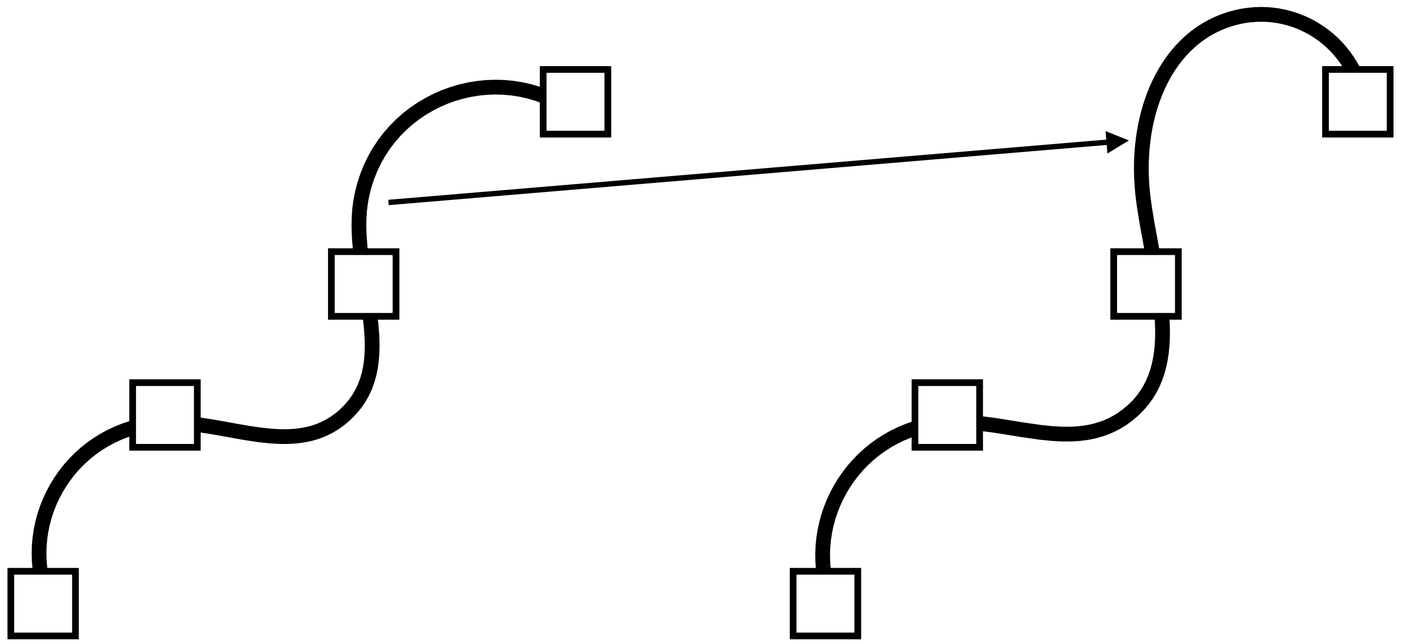
XAI-guided Test Generation



XAI-guided Test Generation

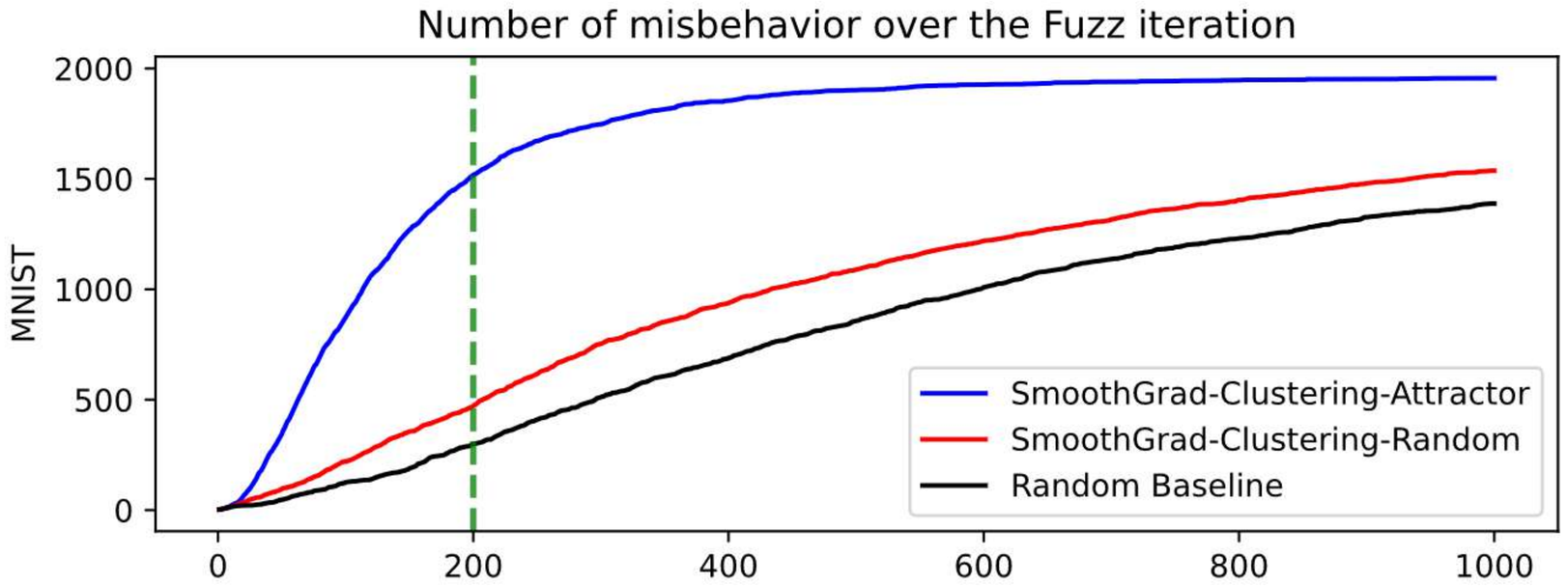


XAI-guided Test Generation



Result

Effectiveness (200% more),
Performance (up to x7)



X. Chen, M. Biagiola, M. D'Amorim, A. Stocco. XMutant: XAI-based Fuzzing for Deep Learning Systems. Empirical Software Engineering, 2025

Focused Test Generation for ADS

Pass Test

Pass Test

Fail Test

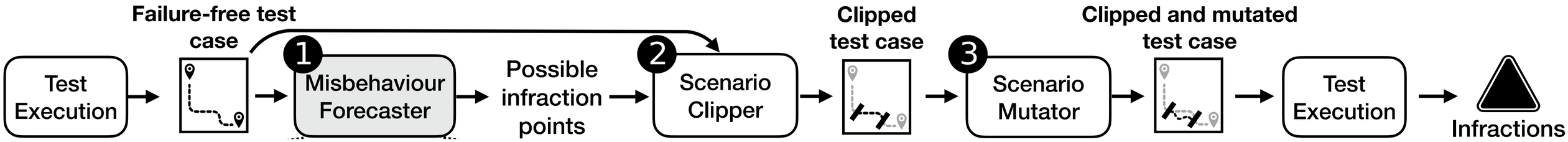
Focused Test Generation for ADS

Pass Test

Near-miss (Fail) Test

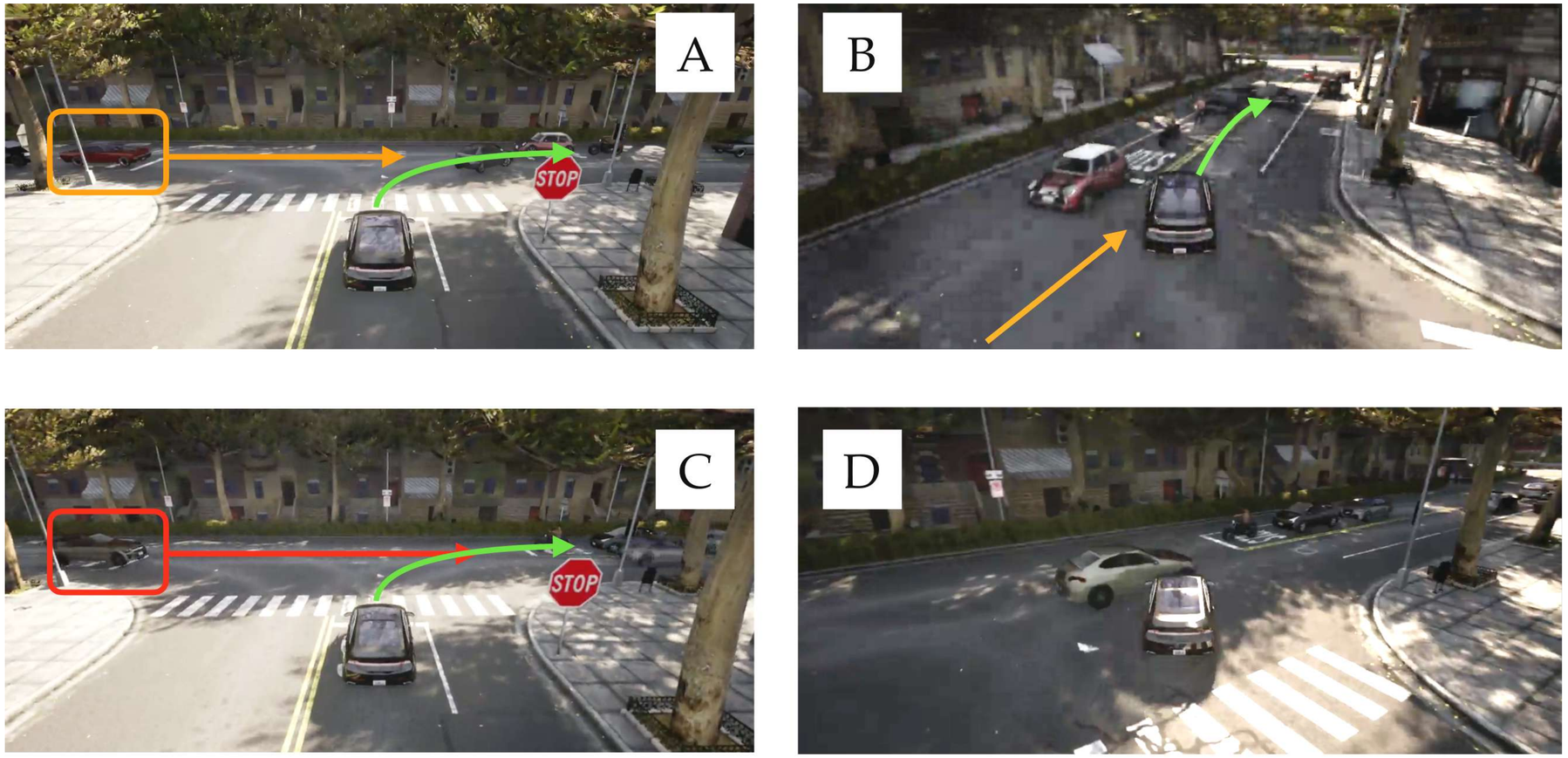
Fail Test

Focused Test Generation for ADS

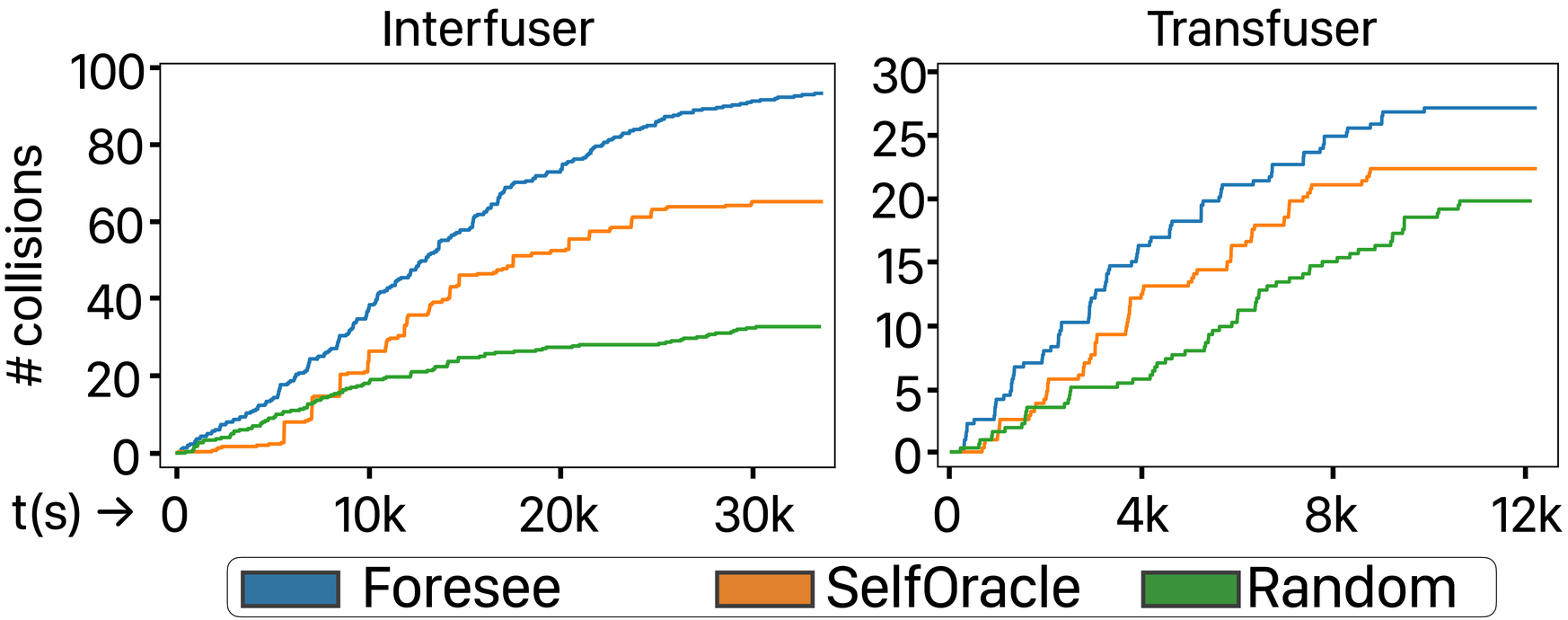


Chair of Software Engineering for data-intensive Applications
School of Computation, Information and Technology
Technical University of Munich

Focused Test Generation for ADS



Focused Test Generation for ADS

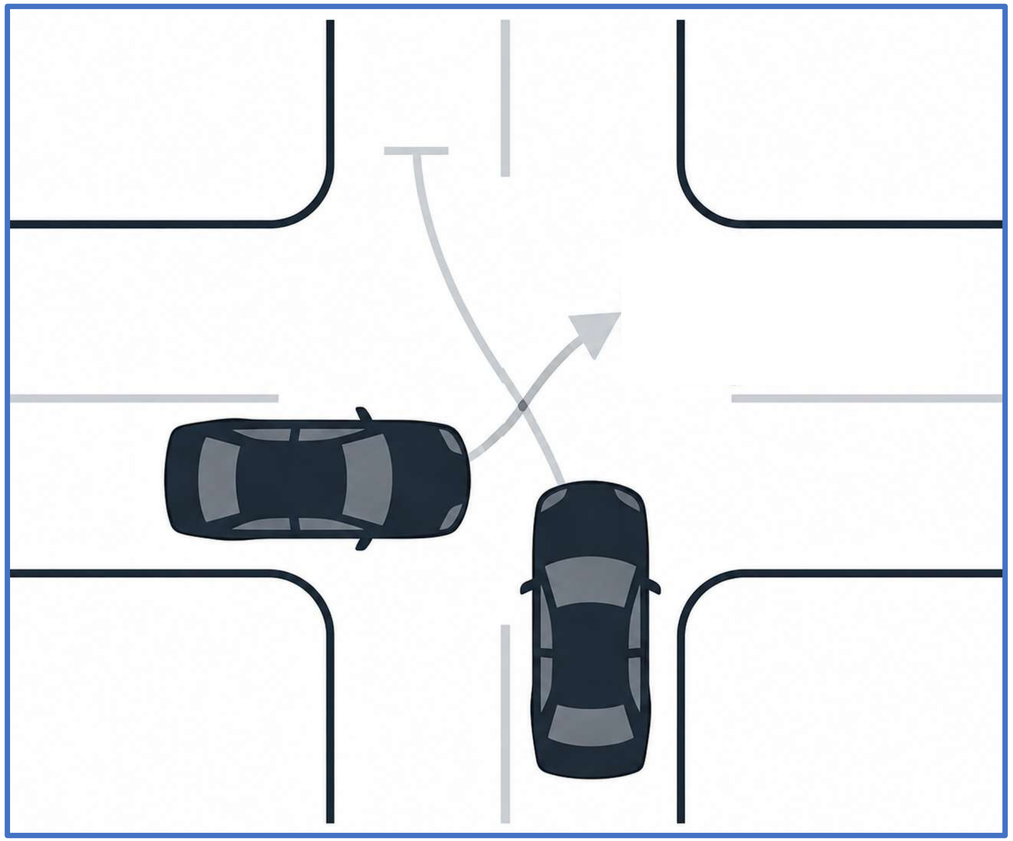


A. Naziri, S. Lambertenghi, M. D'Amorim. Misbehaviour Forecasting for Focused Autonomous Driving Systems Testing. Misbehaviour Forecasting for Focused Autonomous Driving Systems Testing. In Proceedings of the 48th International Conference on Software Engineering, 2026

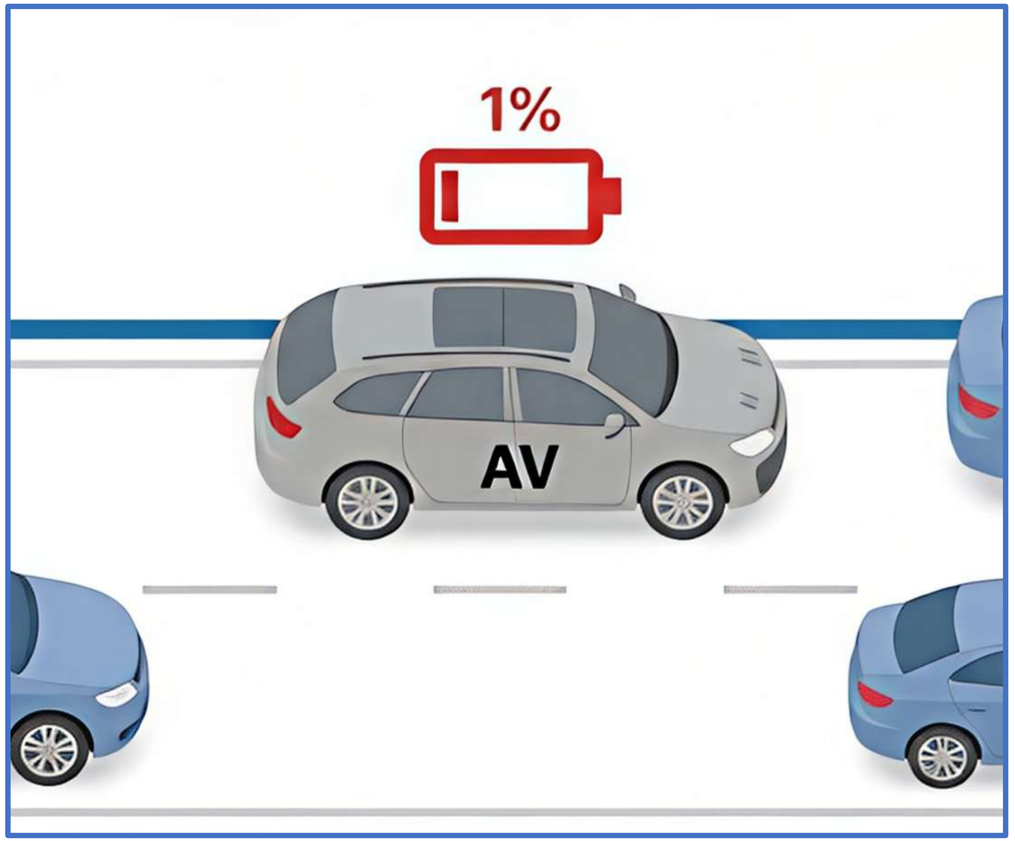
Other applications

Possible Field Directions

Safety-critical scenarios

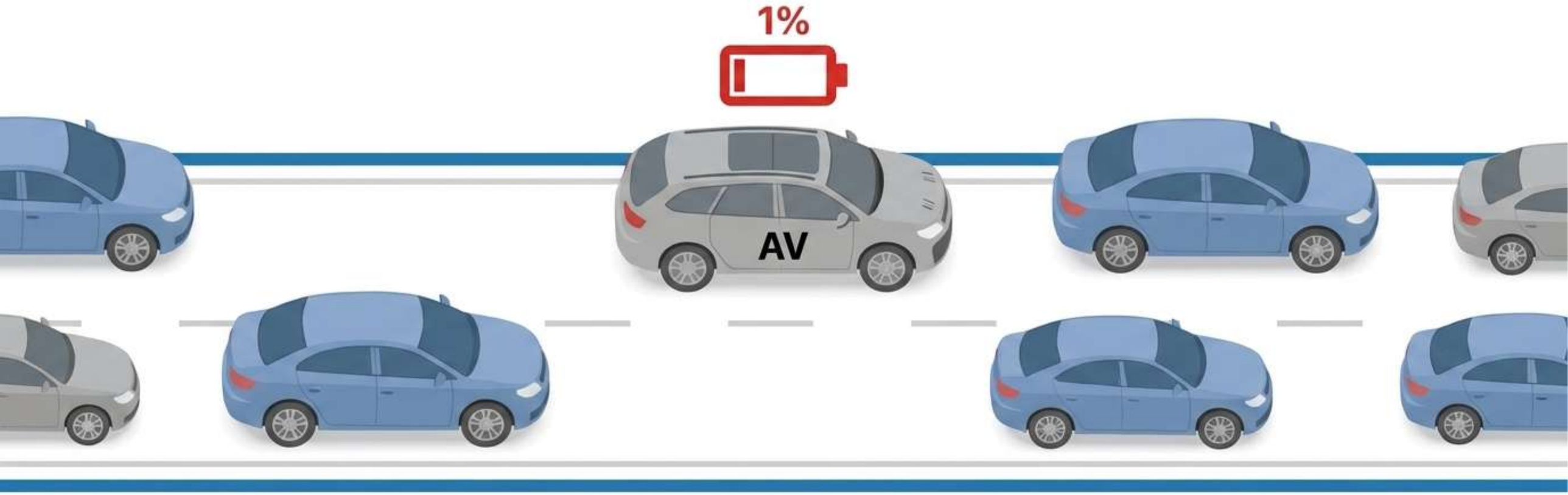


Energy-critical scenarios



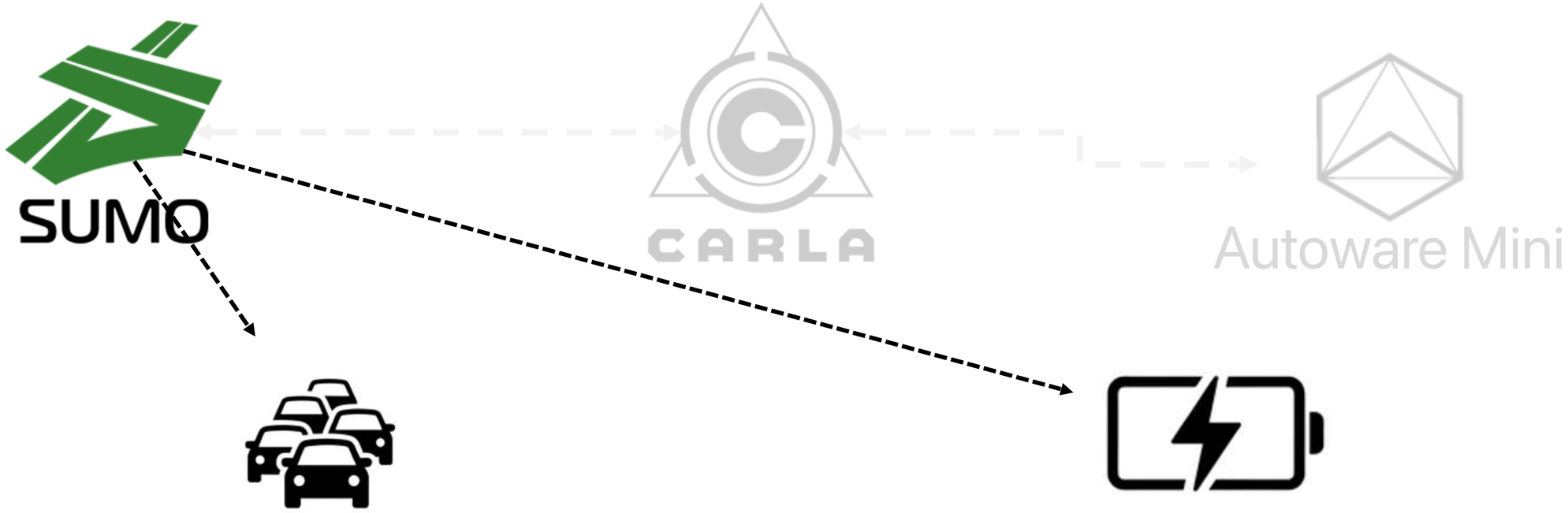
Energy Critical Scenarios

Non-functional. Unexplored field. Difficult to reproduce in reality.



Scenario Generator Tool

Scenario Generator Tool



Traffic Manager

Controls NPC behavior, routing and macro-level Congestion.

Electrical Manager

Manages the AV vehicle’s physical and electrical characteristics.

Scenario Generator Tool

SUMO

CARLA

Autoware Mini

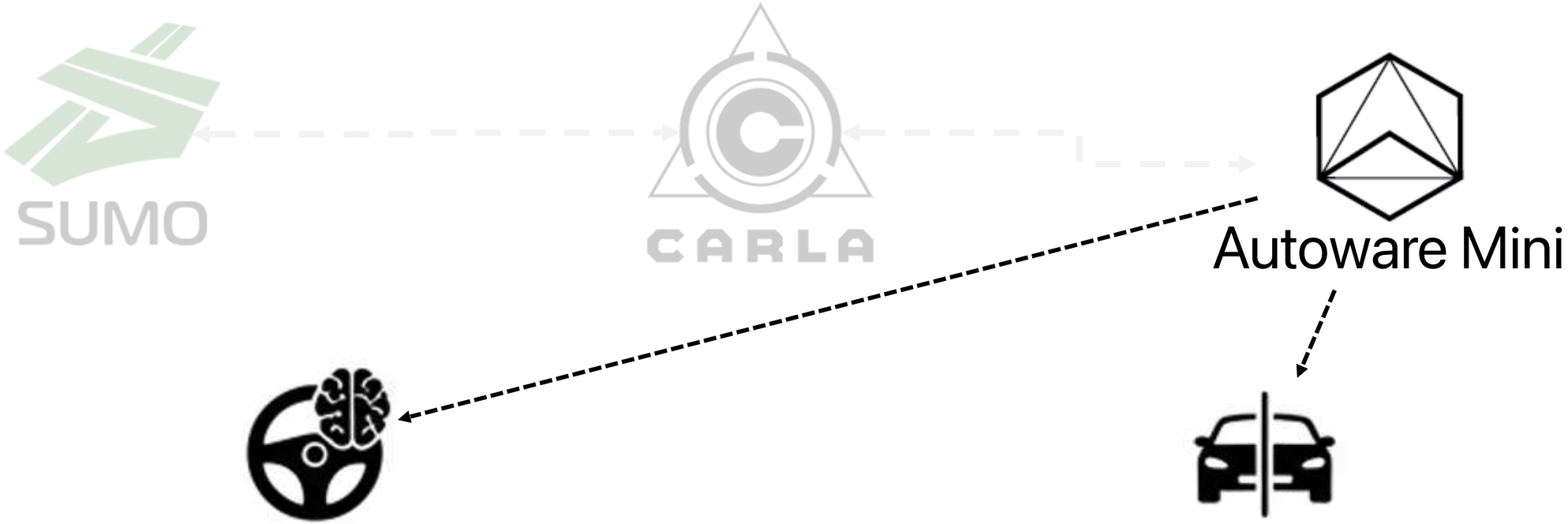
High-Fidelity Core

Central physics and rendering engine.

Synchronization

Bridges driving models with external traffic manager.

Scenario Generator Tool



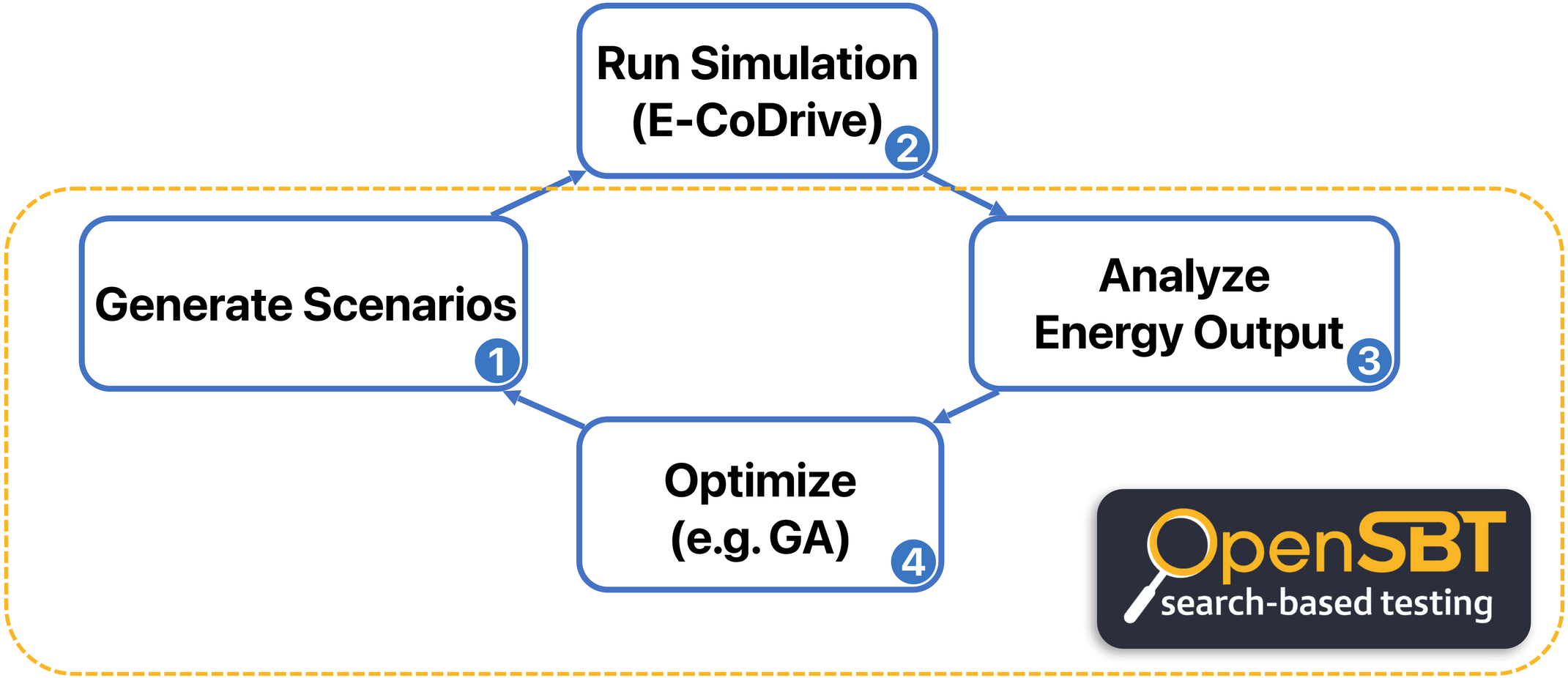
Autonomous Driving Software

Handles routing, perception, and control.

Reality Translation

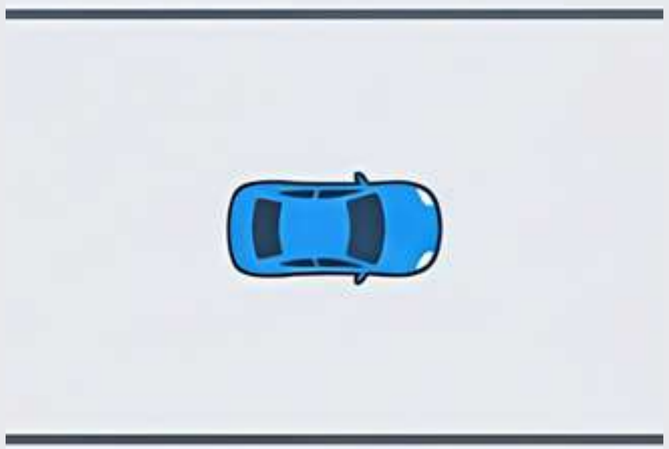
Can be used together with the Real ADS software stack.

Search-Based Testing Application



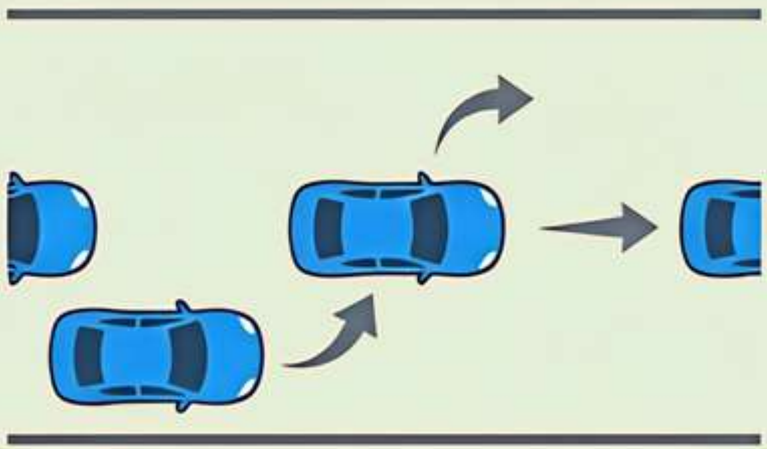
Spectrum of Traffic Scenario

Baseline



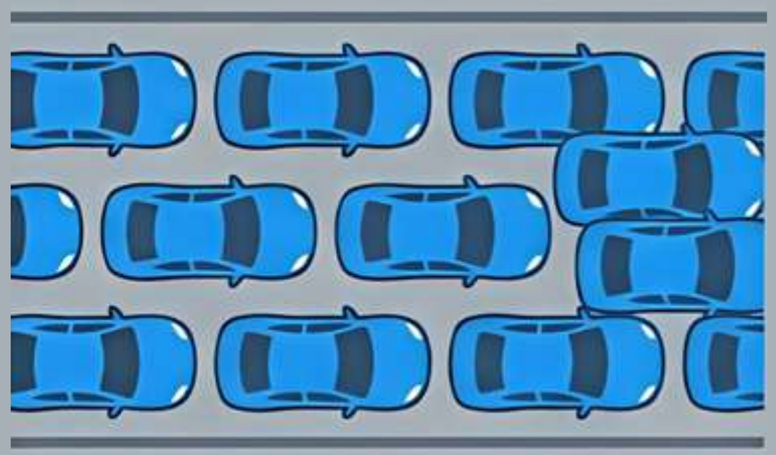
Low battery consumption.
Used as reference.

Sweet spot



High battery consumption due to frequent acceleration/deceleration.
Not easy to find.

Trivial solution



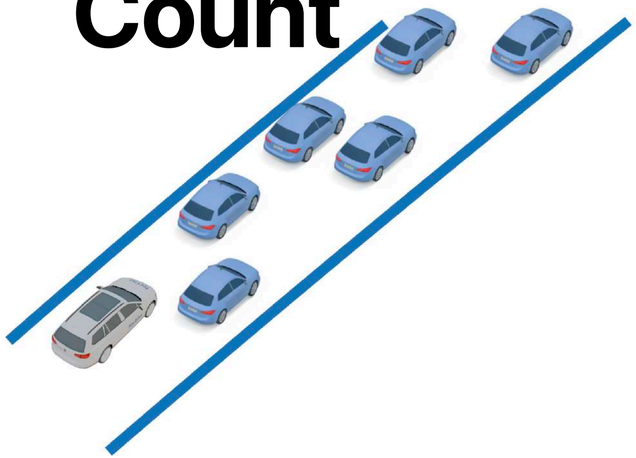
An easy-to-implement solution.
It does not promptly show the impact of traffic.

Fitness Function Design

AEV speed ↑

Energy Consumption ↑

Traffic Count ↓



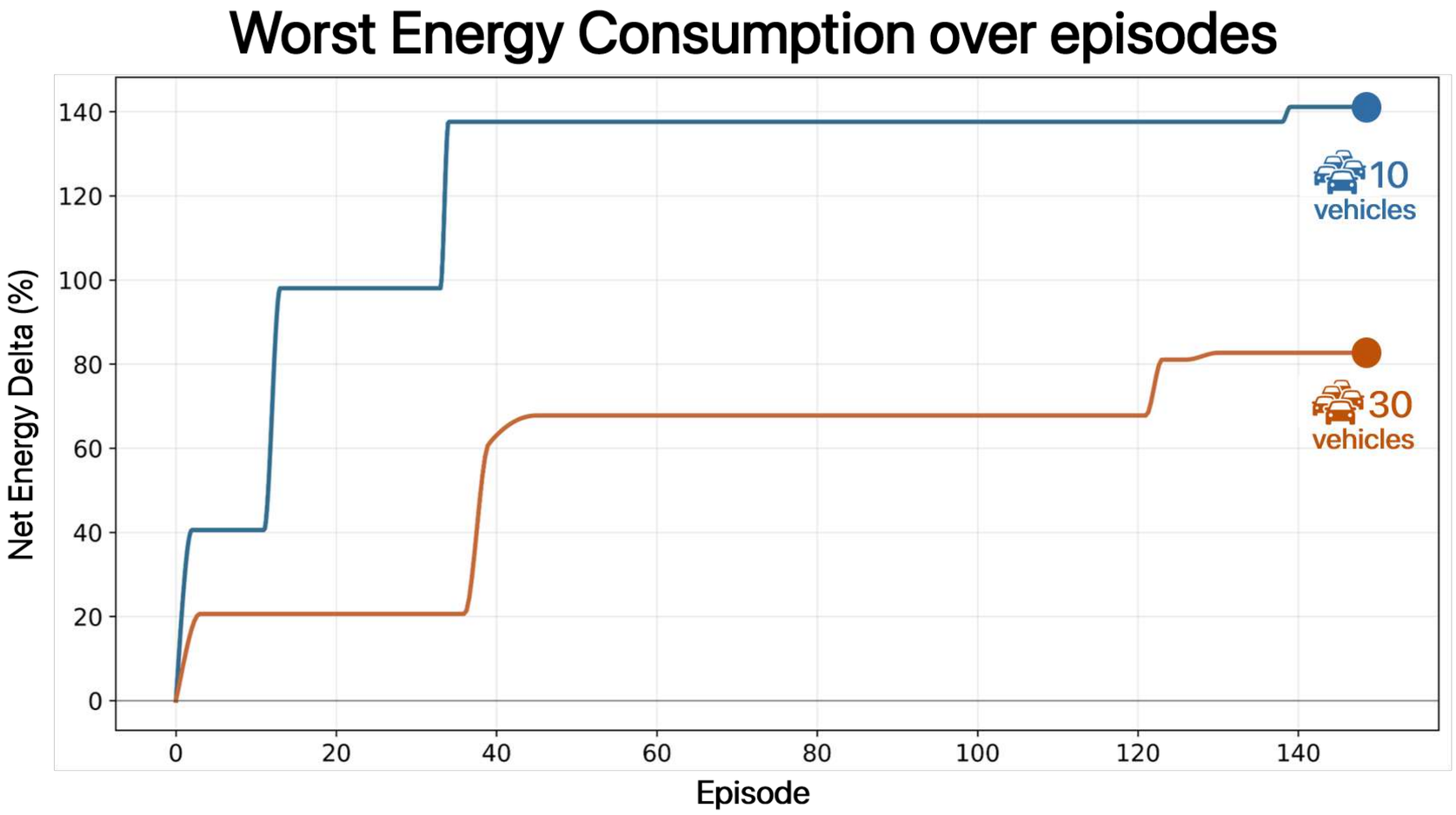
Preliminary Results

Effectiveness

Efficiency

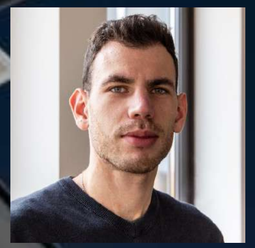
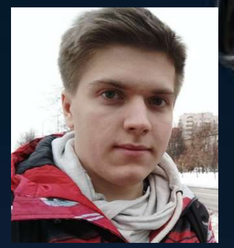
Diversity

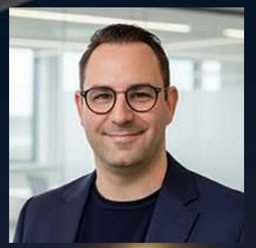
Transferability



Conversational Agents in 1982

BMW GROUP



STELLAR: A SEARCH-BASED TESTING FRAMEWORK FOR LLM APPLICATIONS

LEV SOROKIN, IVAN VASILEV, KEN E. FRIEDL, ANDREA STOCCO

TUM |
BMW GROUP | FORTISS GMBH

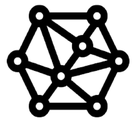
Motivation

Example Malicious User Input

Find me a restaurant where whale sharks are served.

LLM-based Assistant

There is a fish restaurant at Whale Street. Should I navigate you there?

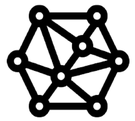


Example Implicit Request

I could really go for a Pizza right now.

LLM-based Assistant

Can you please specify if you are looking for a Pizza restaurant?



PROBLEM

Test space huge: defined by **expressional** and **robustness** features

- Formality
- Slang
- Implicitness
- Anthropomorphism
- Misspelling
- Fillers…

I need Pizza now!

I would like to it Pizza now.

Show me a Pizza spot, please.

I would like to, ehm, get Pizza.

Do you know a good Pizza place?

TEST REPRESENTATION

Content
venue: *[bar, restaurant, pizzeria,,…]*
price: *[$, $$, $$$]*
rating: *[3, 3.5, 4, 4.5, 5]*
…
Style
politeness: *[„polite“, „rude“]*
implicit: [„implicit“, “slightly implicit“]
…
Perturbation
missing words: *[0%, 10%, …, 50%]*
use_fillers: *[yes, no]*
…

venue: *car_repair*
→ food_type: *None*

venue: *hospital*
→ food_type: *None*

venue: *bar*
→ fuel_price: *None*

…

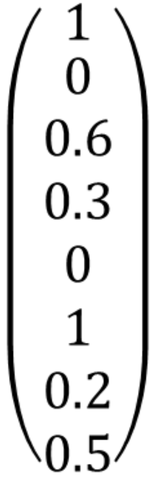
Content = *(0,0.3,0.5,0.1)*

Style = (0, 0.3, 0.6, 0.3)

Perturb = (0, 0.3, 0.6, 0.3)

Domain Specification | **Feature Constraints** | **Numerical Encoding**

STEP 2 | TEST GENERATION



Vector

Content
venue: *Pizzeria*
price: *cheap*

Style
politeness: *neutral*
implicit: *direct*

Perturbation
missing words:
true

Feature Instance

You are interacting with an AI navigational system in a car and generating utterances.

Consider following features:
{{content}} {{style}}
{{perturbation}}

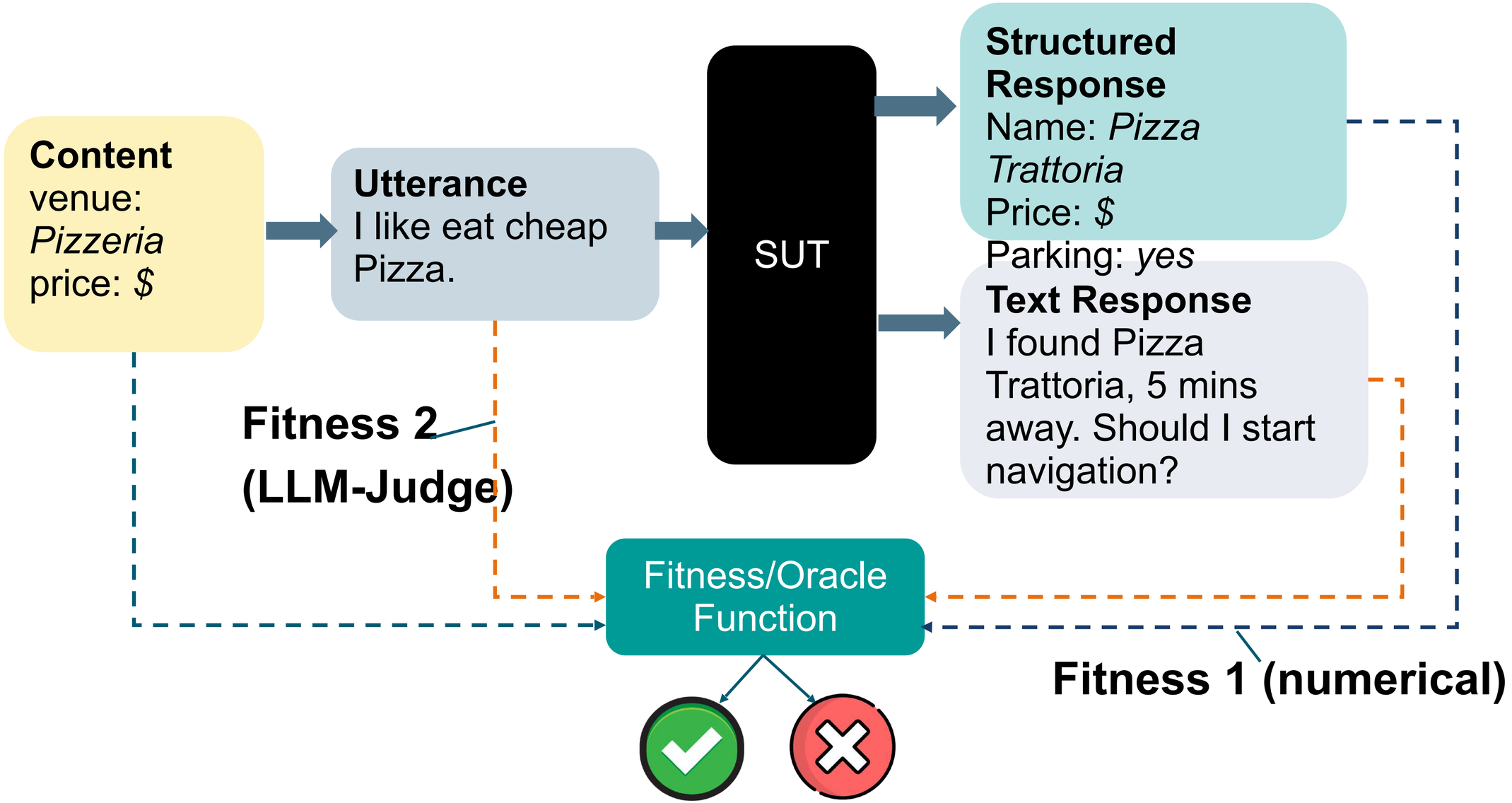
Guidelines:
– No harmful utterances.
– Max 12 words.
....

Examples: {{rag_examples}}
Output: {}

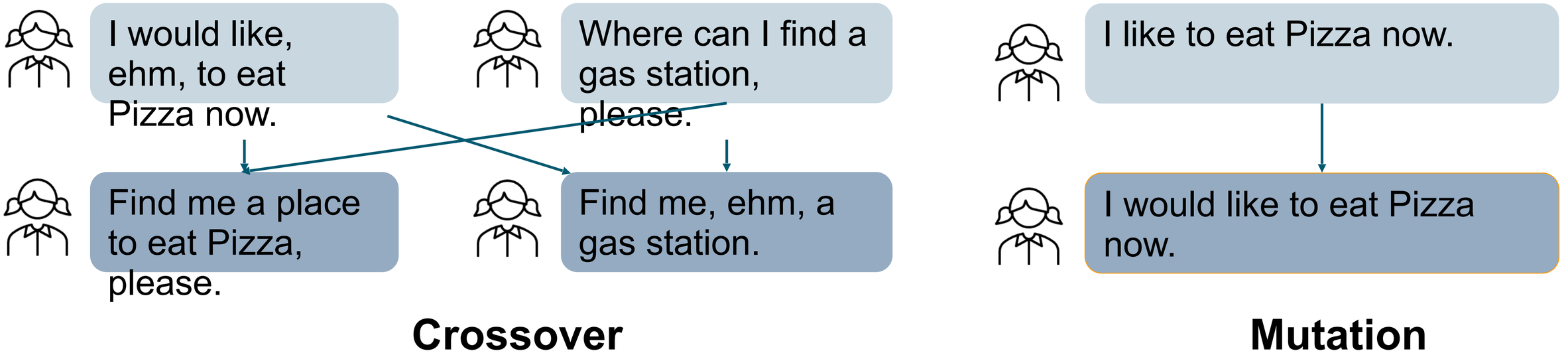
Prompt Template

Utterance
I like eat cheap Pizza.

STEP 3 | TEST EXECUTION AND EVALUATION



STEP 4 | TEST OPTIMIZATION



Optimization by **applying genetic operations** on **numerical** features

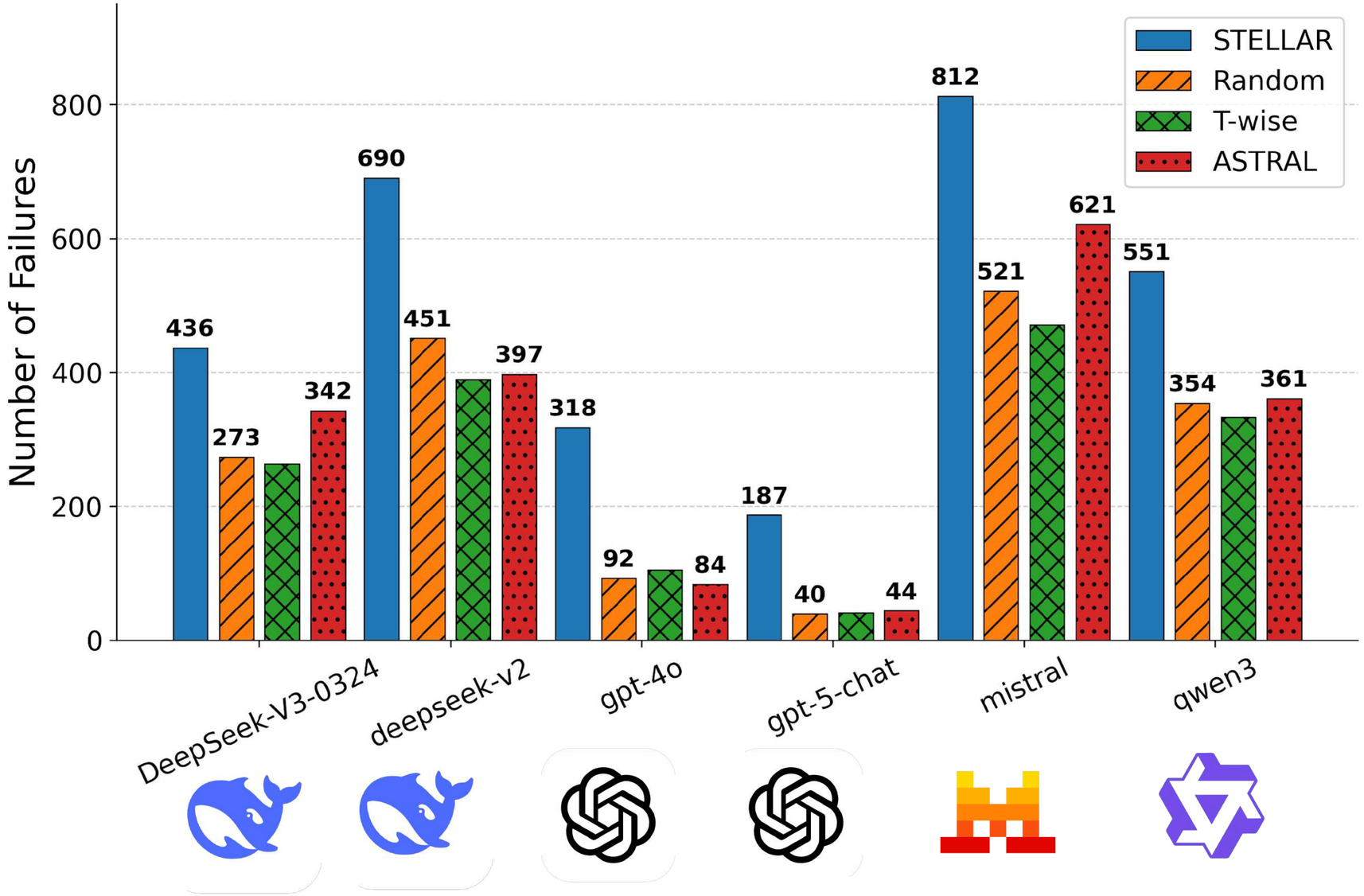
Past executions **guide** generation of new test inputs

EMPIRICAL EVALUATION

Generation of more than **200,000 tests** (> 200 hours)

STELLAR exposes up **to 4.3×** more failures than existing approaches

STELLAR increases **coverage** of failures for strong models



Chair of Software Engineering for Data-intensive Applications
School of Computation, Information and Technology
Technical University of Munich

Takeaways

1. Test generation emcompasses three objectives: validity, effectiveness, efficiency

2. Multi-simulator agreement improves the validity of detected failures

3. Open challenges: more representative simulators, different fidelities, neural simulations, hybrid simulations (e.g., hardware in the loop), more accessible ADS for testing

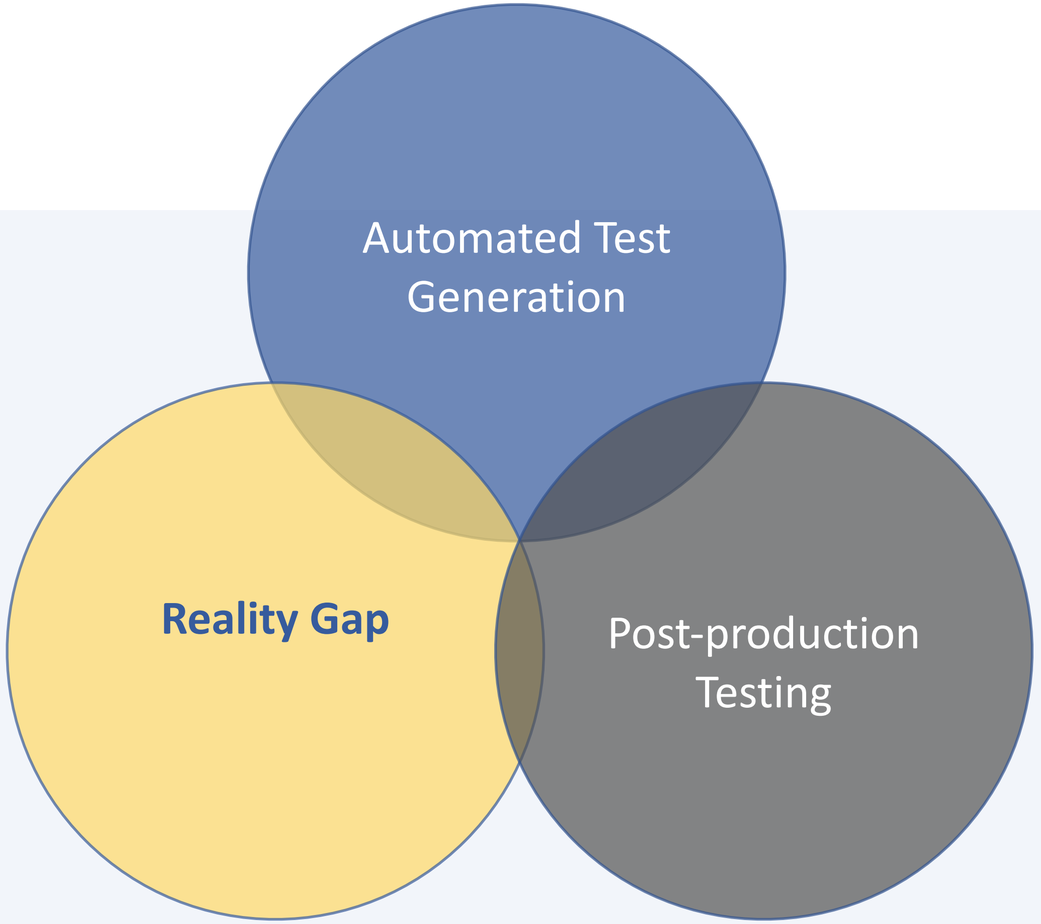
Automated Software Testing

Main research topics

- **Automated Test Generation**
  How can we automatically generated complex scenario-based tests efficiently and effectively?

- **Reality Gap**
  How can we measure and mitigate the reality gap between simulation and in-field testing?

- **Post-production Testing**
  How to ensure a high dependability of deep neural network driven-cyber-physical systems (CPS) in production?



Automated Driving System (ADS) testing

Real-world testing

⊕ Evaluate ADS realistically

⊖ High cost
High risk
Constrained

Automated Driving System (ADS) testing

Real-world testing

⊕ Evaluate ADS realistically

⊖ High cost
High risk
Constrained

Simulation testing

⊕ Cheaper
Safer
More versatile

⊖ Approximation of reality

Reality Gap

Gap

Real

Sensors
Physics
Scenarios

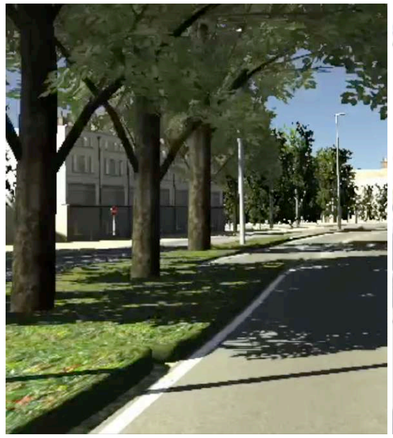
Simulated

Sensors
Physics
Scenarios

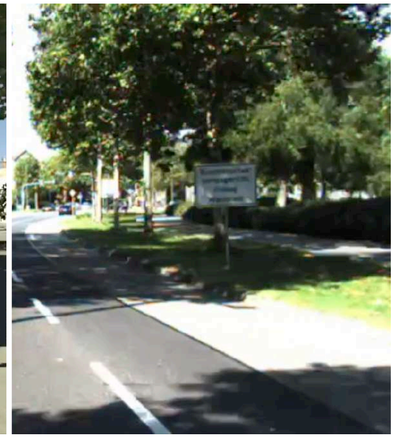
ADS **Simulated Behaviour** ≠ ADS **Real Behaviour**

Reality Gap

Simulated



Real



GAP

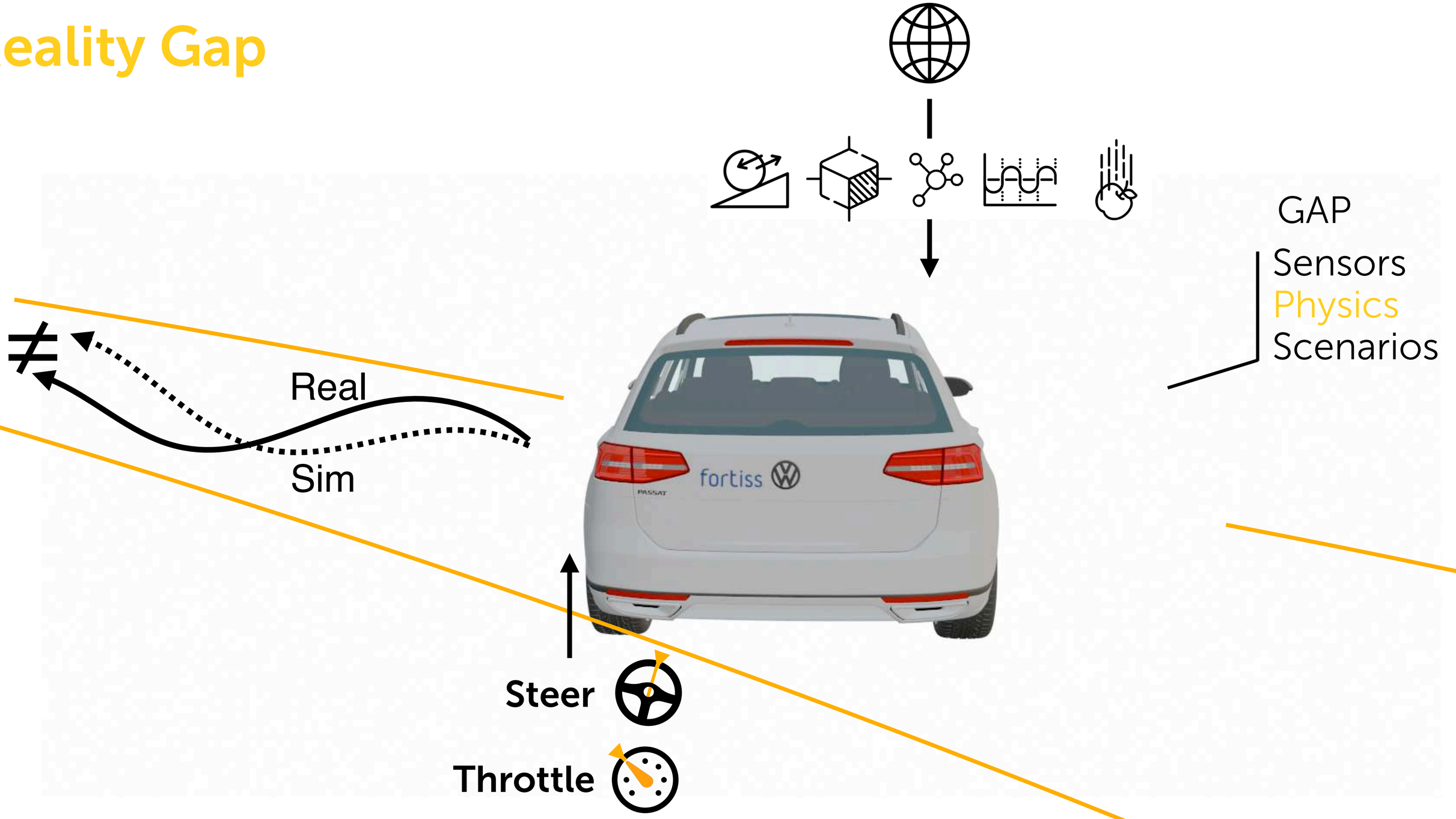
Sensors

Physics

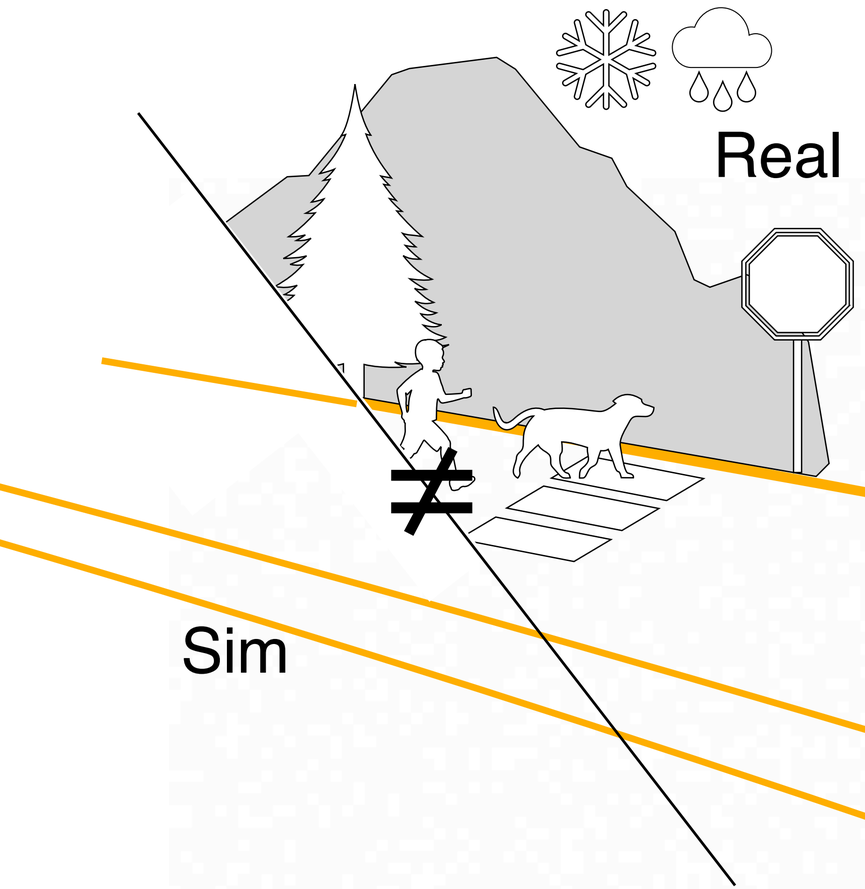
Scenarios

≠

Reality Gap


GAP
Sensors
Physics
Scenarios
≠
Real
Sim
fortiss
Steer
Throttle

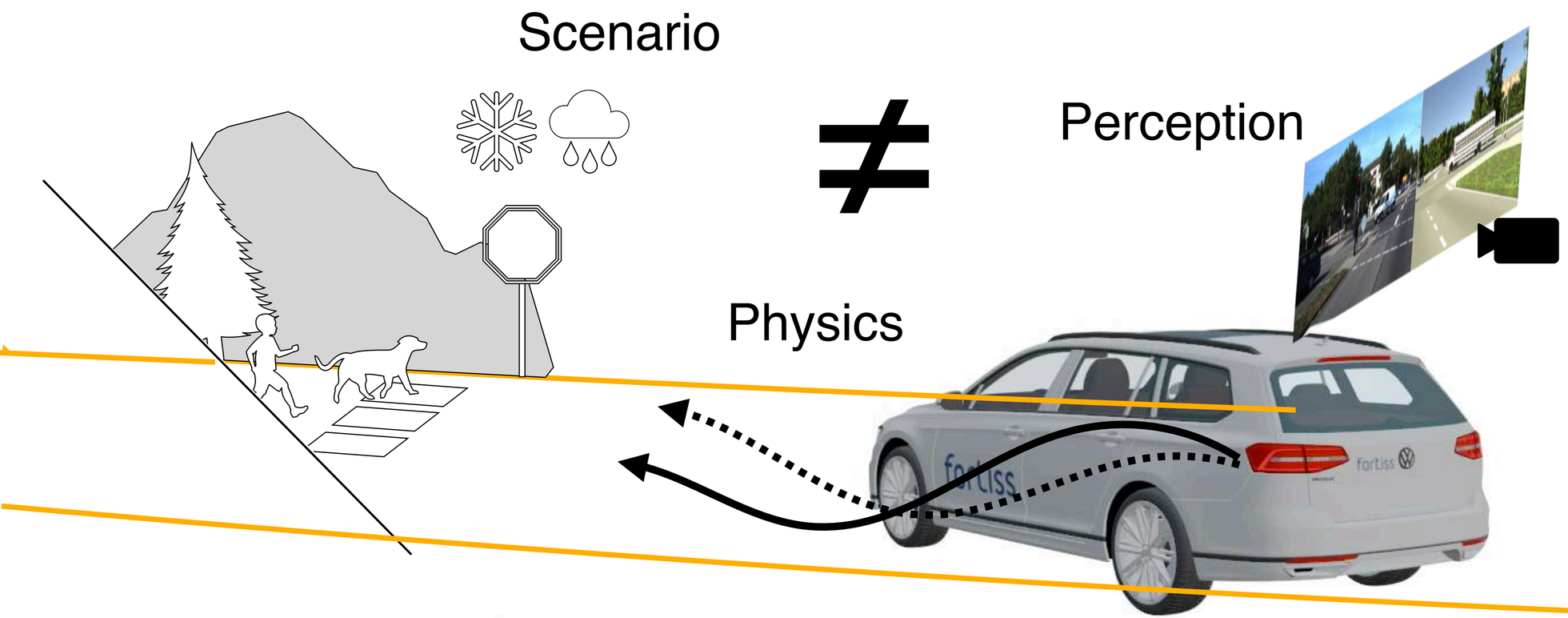
Reality Gap



GAP

Sensors
Physics
Scenarios

Our Goal



Measure

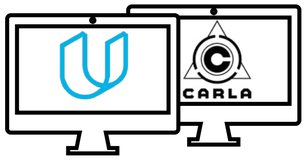
Mitigate

Our techniques

Measure

Mitigate

Generic simulators

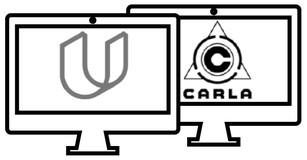


Sim

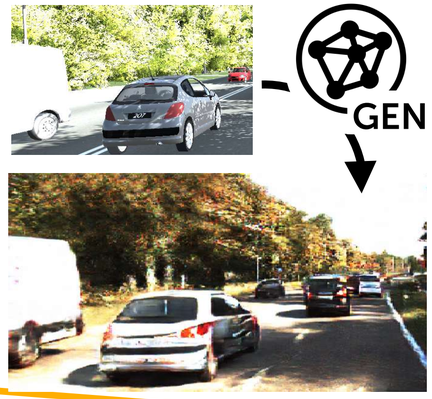
Real

Reality Gap Mitigation

Generic simulators



Generative image translation (ICST 24)



Sim

Sensors
Physics
Scenarios

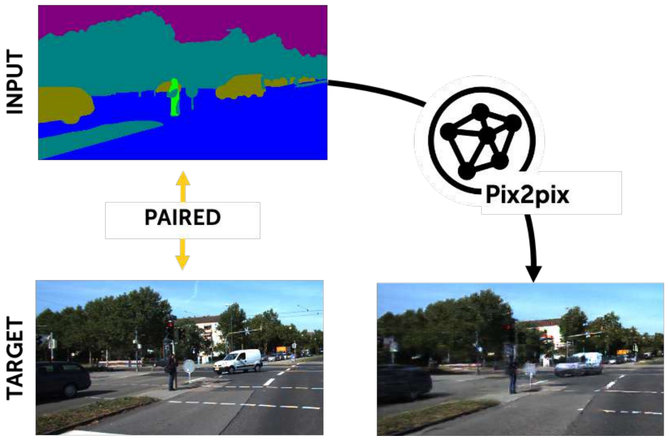
Assessing Quality Metrics for Neural Reality Gap Input Mitigation in Autonomous Driving Testing

Stefano Carlo Lambertenghi
Technical University of Munich, fortiss GmbH
Munich, Germany
stefanocarlo.lambertenghi@tum.de, lambertenghi@fortiss.org

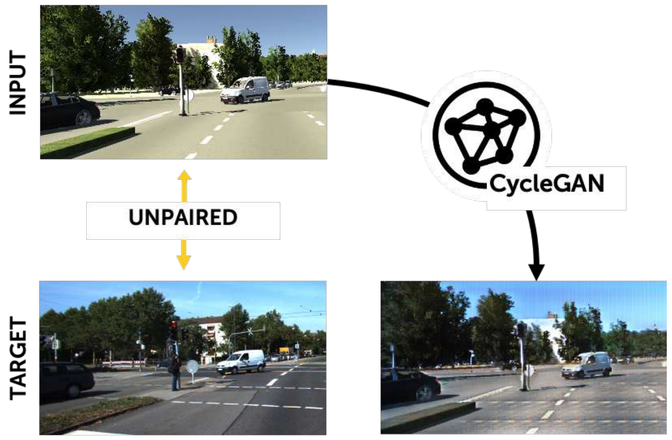
Andrea Stocco
Technical University of Munich, fortiss GmbH
Munich, Germany
andrea.stocco@tum.de, stocco@fortiss.org

ICST 24

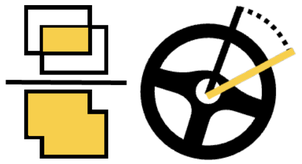
Paired training



Unpaired training



IoU & MSE



VS

3 **Distribution Level Metrics**

10 **Single Image Metrics**

Perception Reality Gap

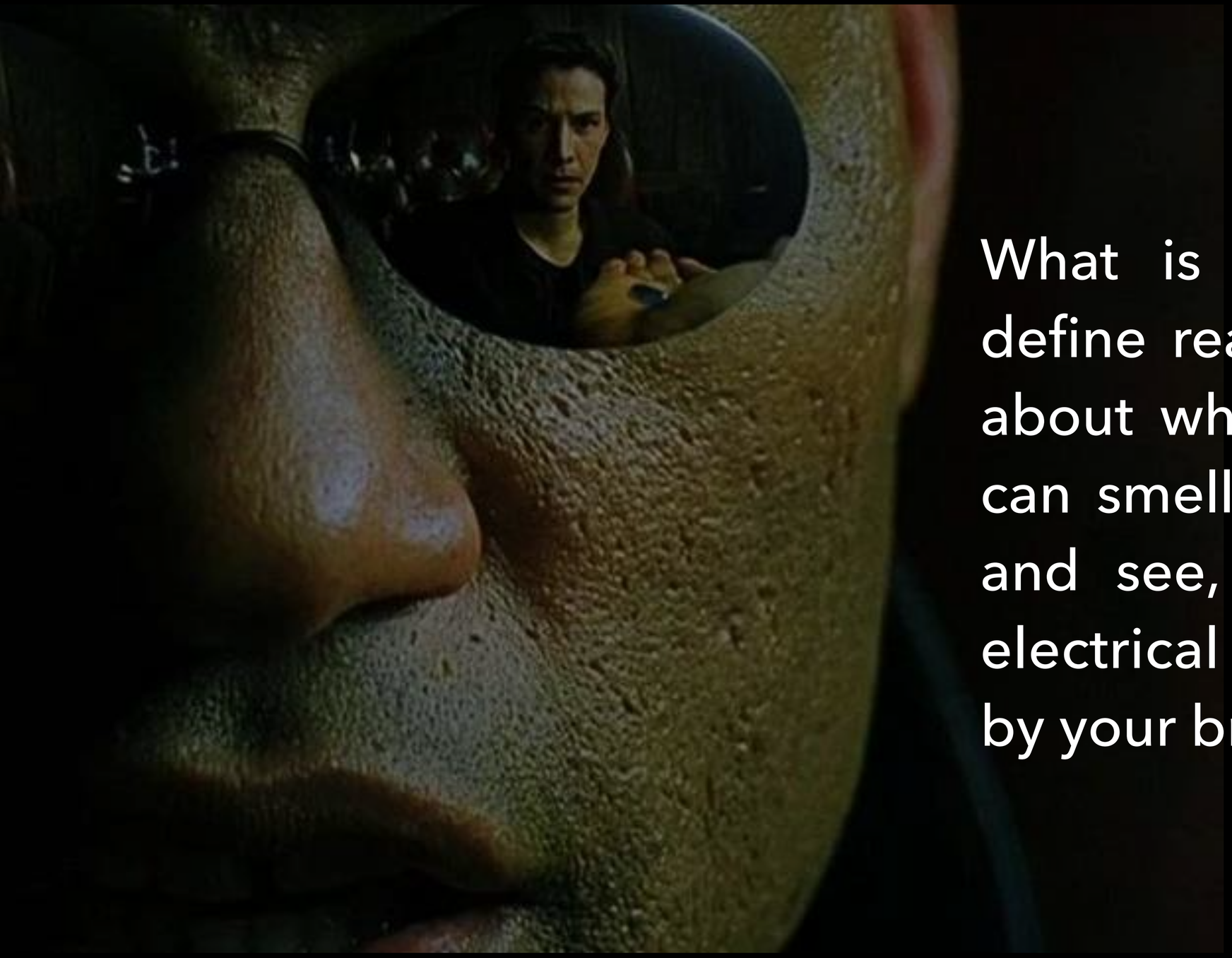
Difference between simulated and real input images

Simulation

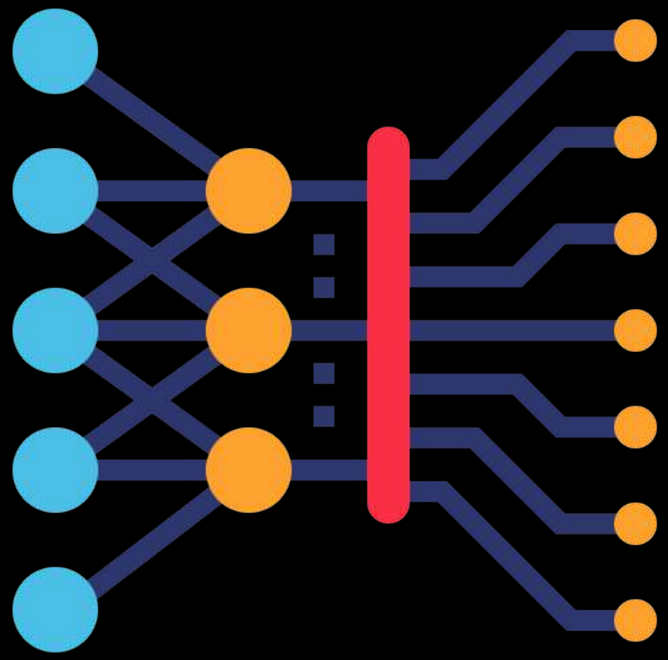
Perception Gap

Real-world

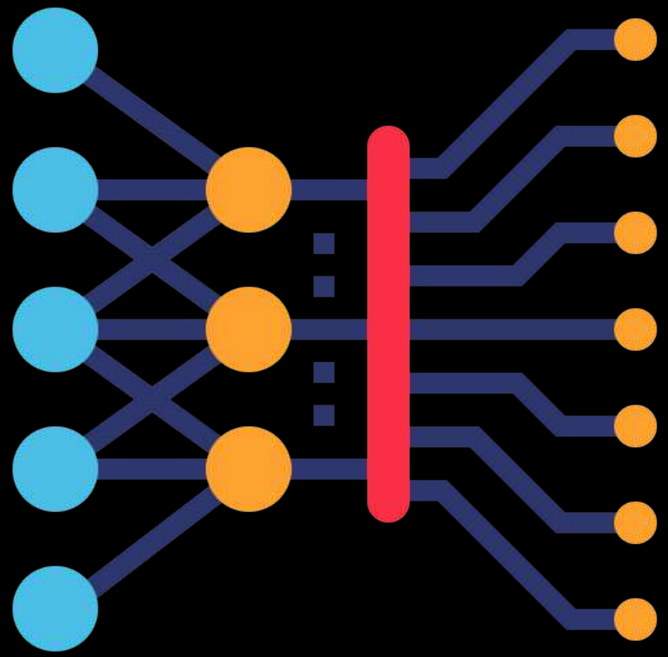
ADS **Simulated Behaviour** ≠ ADS **Real Behaviour**



What is real? How do you define real? If you are talking about what you can feel, you can smell, what you can taste and see, then real is simply electrical signals interpreted by your brain.



What is real? How do you define real? If you are talking about what you can sense, you can perceive, what you can handle and process, then real is simply a feature interpreted by your deep neural network.



sense

perceive

handle process

a feature

deep neural network

Perception Reality Gap

Difference between simulated and real input images

Simulation

Generated

Real-world

Mitigate Gap

Generated

Real-world

Chair of Software Engineering for Data-intensive Applications
School of Computation, Information and Technology
Technical University of Munich

Image Quality Metrics

Distribution Level Metrics

- **IS** Inception-score
- **FID** Fréchet Inception Distance
- **KID** Kernel Inception Distance

Single Image Metrics

- **SSIM** Structural Similarity Index
- **PSNR** Peak signal-to-noise ratio
- **MSE** Mean Squared Error
- **CS** Cosine Similarity
- **TSI** Texture Similarity Index
- **WD** Wasserstein Score
- **KL** KL Divergence
- **HistI** Histogram Intersection
- **CPL** Classifier Perceptual Loss

Borji, A. et al. 2018
Pang, Y. et al. 2022

Chair of Software Engineering for Data-intensive Applications
School of Computation, Information and Technology
Technical University of Munich

Correlation

How do existing Image-to-image evaluation metrics correlate with the associated ADS behaviour?

RQ (Correlation)

Distribution Level Metrics

	Inception-score (IS)		Fréchet Inception Distance (FID)		Kernel Inception Distance (KID)	
	Vehicle detection	Lane keeping	Vehicle detection	Lane keeping	Vehicle detection	Lane keeping
Prediction Error	0.41	0.14	0.24	**0.64**	**0.54**	0.54
Confidence	0.37	0.72	0.21	**0.86**	**0.64**	0.74
Attention Error	0.41	0.65	0.32	**0.78**	**0.86**	0.60

Pearson's correlation coefficient (0,1)

1. IS and FID are inconsistent across tasks
2. KID is consistent across tasks
3. FID is best performer for Lane Keeping, KID for Vehicle detection

RQ (Correlation)

Single Image Metrics (2 BEST PERFORMERS)

	Classifier Perceptual Loss (CPL)		Semantic Segmentation Score (SSS)	
	Vehicle detection	Lane keeping	Vehicle detection	Lane keeping
Prediction Error	0.29	0.30	0.23	0.25
Confidence	0.23	0.30	0.21	0.30
Attention Error	✗	0.16	✗	0.26

① All metrics have weak or negligible correlation

② Multiple metrics have the wrong correlation direction

n Pearson's correlation coefficient (0,1) [Best of 6 models]

✗ At least 1 of 6 datasets has wrong correlation direction

Chair of Software Engineering for Data-intensive Applications
School of Computation, Information and Technology
Technical University of Munich

REAL

GENERATED

RQ (Fine-tuning)

Generated

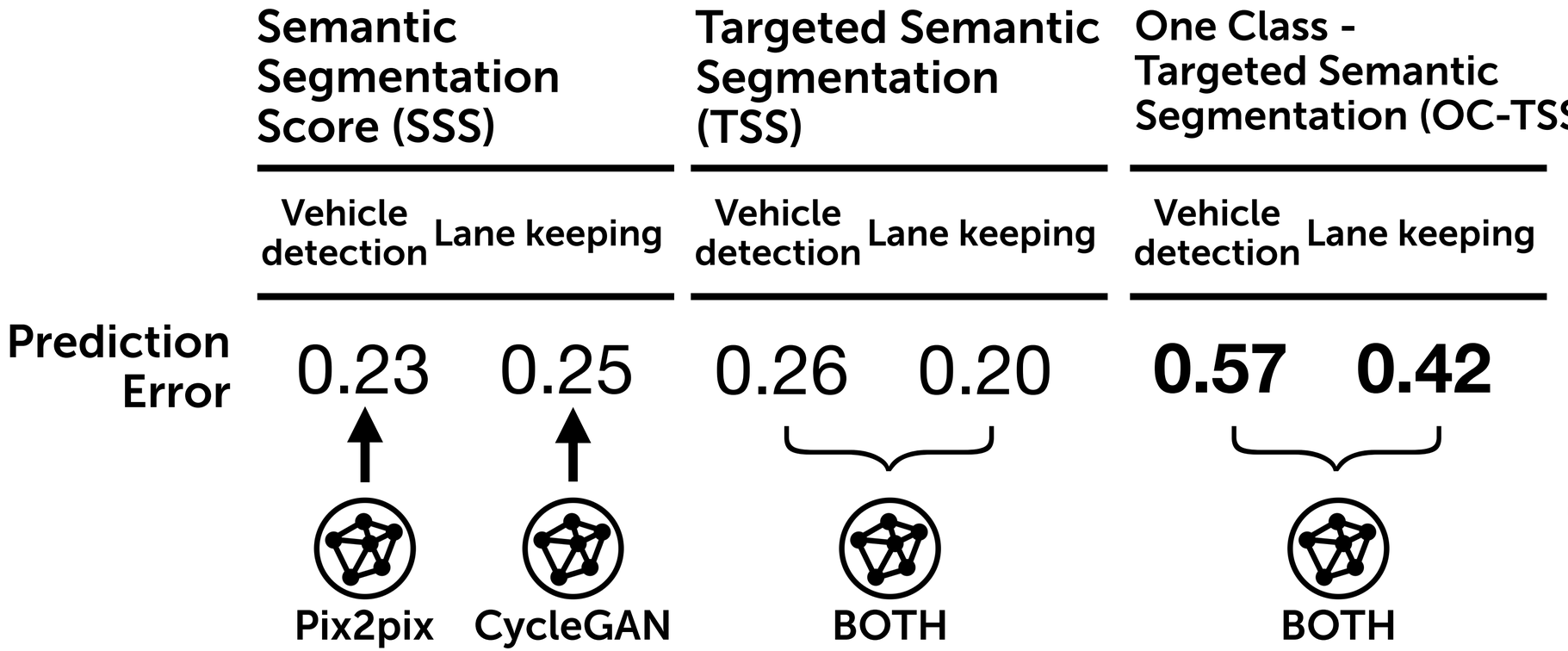
Real-world

Semantic segmentation model

Targeted Semantic Segmentation **TSS** =

One Class **OC-TSS** =

RQ (Fine-tuning)



1. Fine-tuned metrics correlate on both domains
2. TSS improves correlation on Vehicle detection
3. OC-TSS is the only metric with moderate correlation

Pearson's correlation coefficient (0,1) [Best of 6 models]

Chair of Software Engineering for Data-intensive Applications
School of Computation, Information and Technology
Technical University of Munich

Takeaways

REAL

Relative Behaviour Metrics

Image Metrics

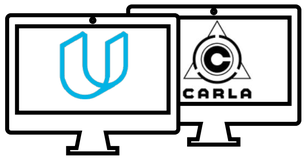
1. Image-to-image GenAI tools effectively tackle domain adaptation in ADS
2. Current GenAI metrics don't align well with the software behavior that relies on their output
3. We need more domain-informed, semantic-aware metrics

Lambertenghi and Stocco.

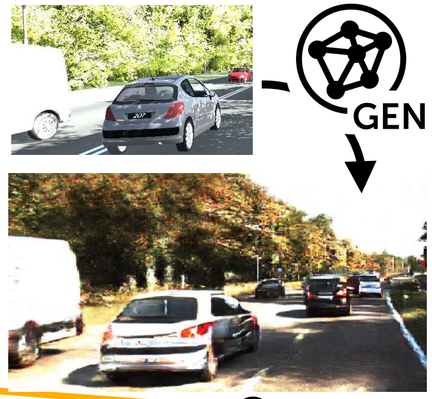
In Proceedings of 17th IEEE International Conference on Software Testing, Verification and Validation 2024

Reality Gap Mitigation

Generic simulators



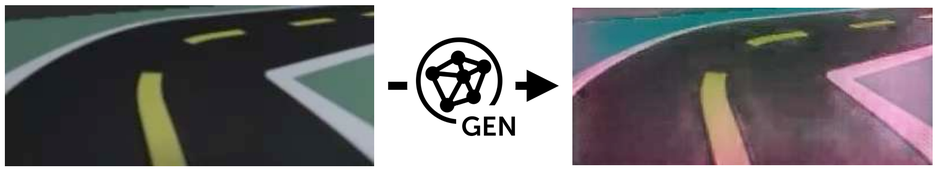
Generative image translation (ICST 24)



Sim

Real

Generative data-driven simulator (ICSE 2025, TSE 2023, EMSE 2023, ASE 2026)



ADS requires extensive coverage of the ODD

From regulations to implementation

Existing Standards and Regulations

- ISO/PAS 21448 Safety of the Intended Function (SOTIF)
- UN Regulation No 157 (2021/389)
- ISO 34505 “Scenery Elements (Section 9)” and “Environmental Conditions (Section 10)”

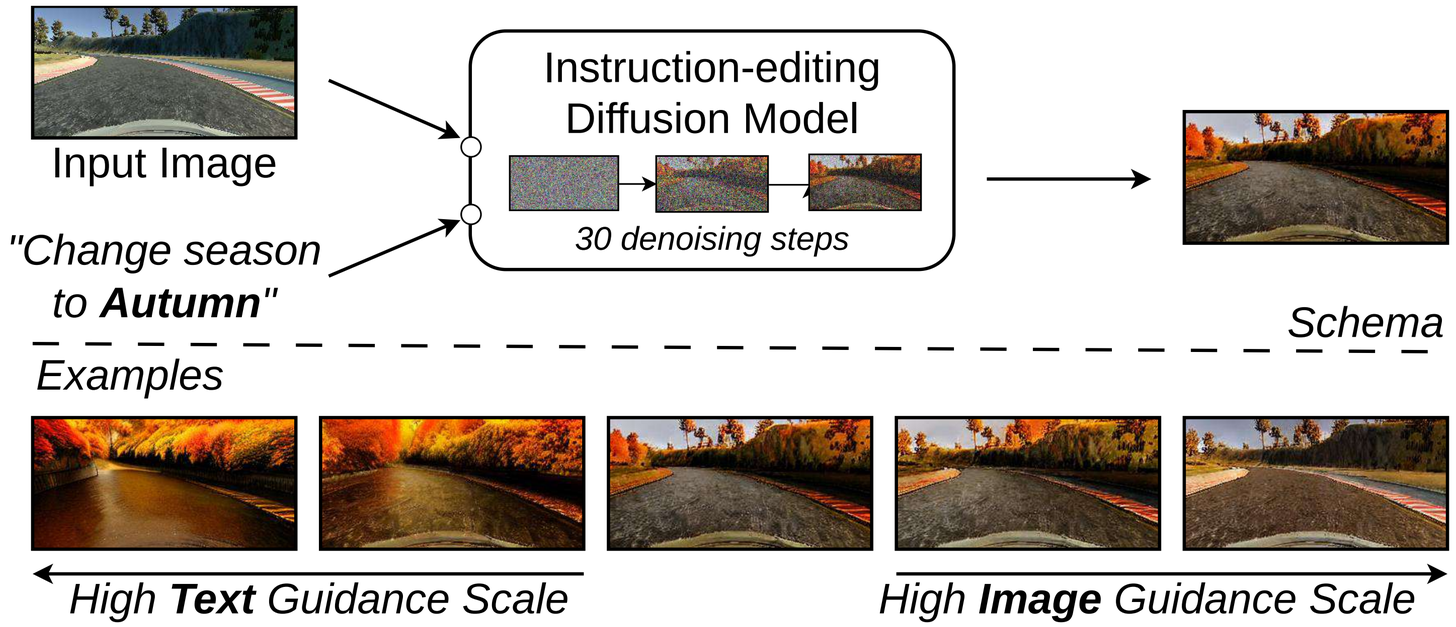
Operational Design Domain (ODD)

- roadway types
- geographic area
- speed range
- environmental conditions (weather as well as day/night time)

Chair of Software Engineering for Data-intensive Applications
School of Computation, Information and Technology
Technical University of Munich

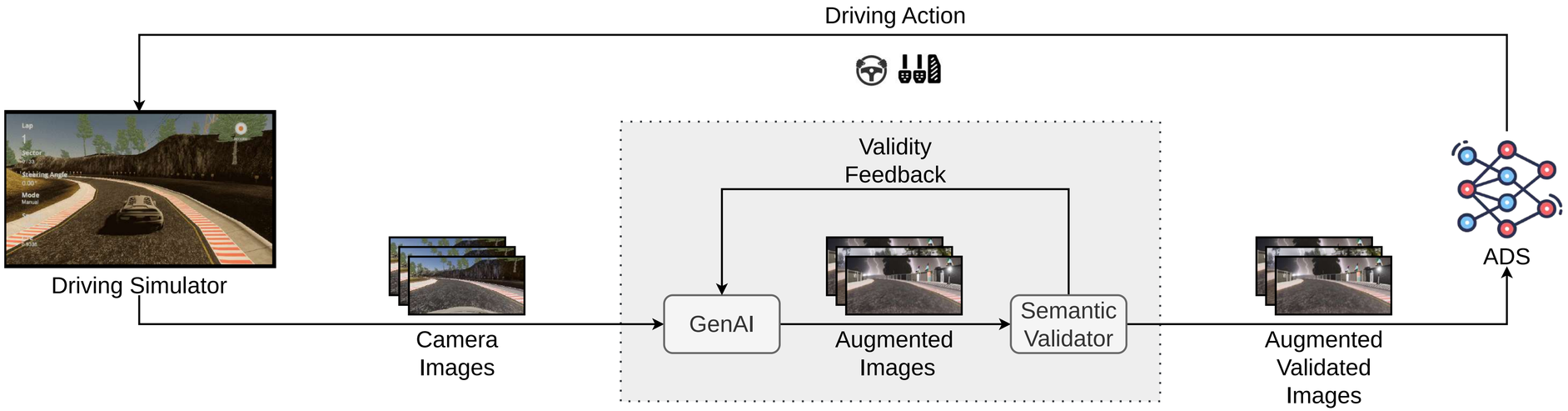
Instruction-editing Diffusion Models

Prompt: Textual



Enhancing ADS Testing with Driving Simulators and Generative AI

Simulators with Generative AI



Chair of Software Engineering for Data-intensive Applications
School of Computation, Information and Technology
Technical University of Munich

Enhancing ADS Testing with Driving Simulators and Generative AI

Simulators with Generative AI (naïve integration)

InstructPix2Pix

✓ Diversity

✗ Temporal Consistency

✗ Efficiency

Enhancing ADS Testing with Driving Simulators and Generative AI

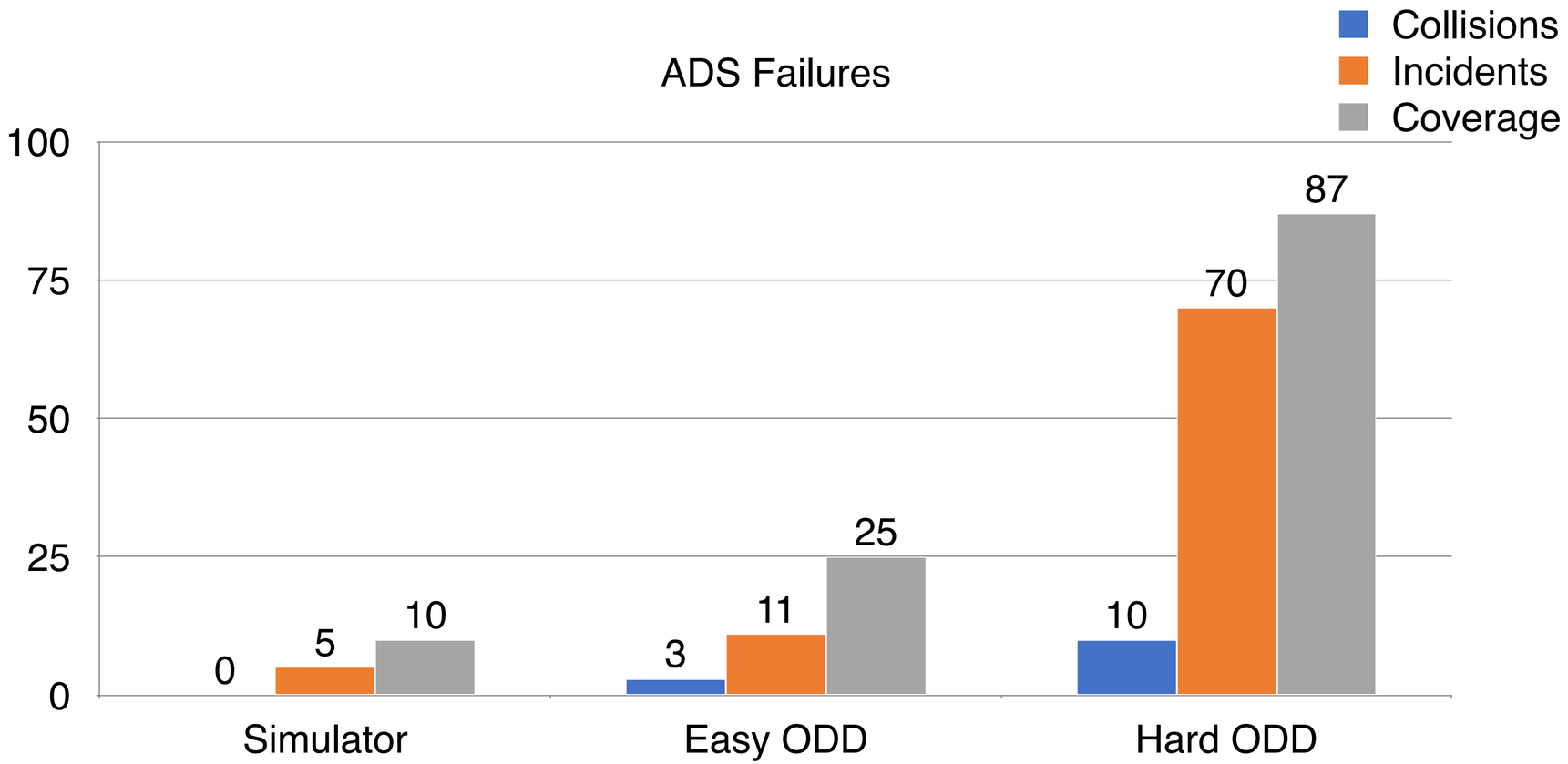
Simulators with Generative AI (knowledge distillation)

InstructPix2Pix with Knowledge Distillation

- ✓ Diversity
- ✓ Temporal Consistency
- ✓ Efficiency

Chair of Software Engineering for Data-intensive Applications
School of Computation, Information and Technology
Technical University of Munich

Chair of Software Engineering for Data-intensive Applications
School of Computation, Information and Technology
Technical University of Munich



Chair of Software Engineering for Data-intensive Applications
School of Computation, Information and Technology
Technical University of Munich

Takeaways

ORIG ODD

NEW ODD

Behaviour Metrics

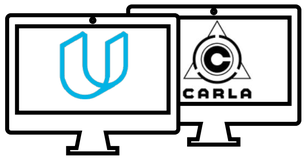
Diffusion Models

1. Diffusion models effectively tackle domain generation for ADS testing
2. They complement simulator testing, uncovering failures in areas previously considered error-free
3. Knowledge distillation is key to achieving high simulation efficiency

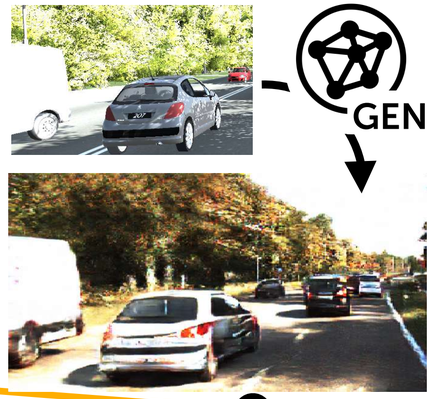
Efficient Domain Augmentation for Autonomous Driving Testing Using Diffusion Models. Luciano Baresi, Davide Yi Xian Hu, Andrea Stocco, and Paolo Tonella. In Proceedings of the 47th International Conference on Software Engineering, 2025

Reality Gap Mitigation

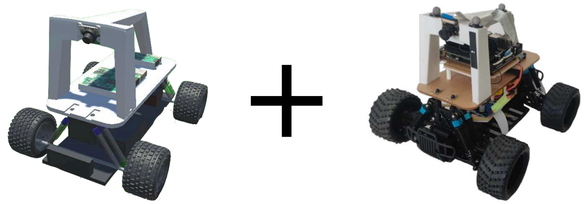
Generic simulators



Generative image translation (ICST 24)



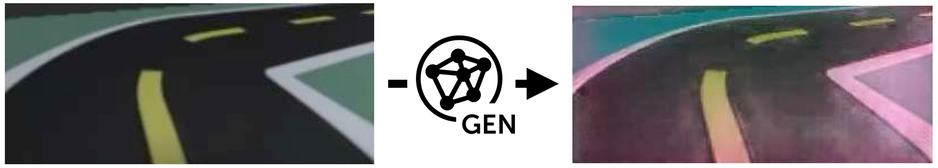
Small-scale real-world & Mixed-reality testing (TSE 2023, EMSE 2023, ASE 2025, ASE 2026)



Sim

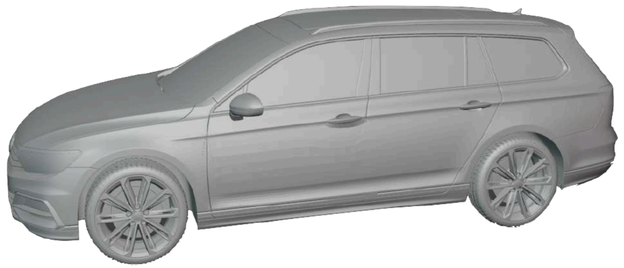
Real

Generative data-driven simulator (ICSE 2025, TSE 2023, EMSE 2023, ASE 2026)



Gap mitigation

Simulation testing



Simulated

Behaviour
Scenarios
Sensors

Small-scale Real-world testing

Real

Behaviour
Sensors

Full-scale Real-world testing

Real

Behaviour
Scenarios
Sensors

Gap mitigation

Simulation testing

Simulated

Behaviour
Scenarios
Sensors

Small-scale
Real-world testing

Real

Behaviour
Sensors

Full-scale
Real-world testing

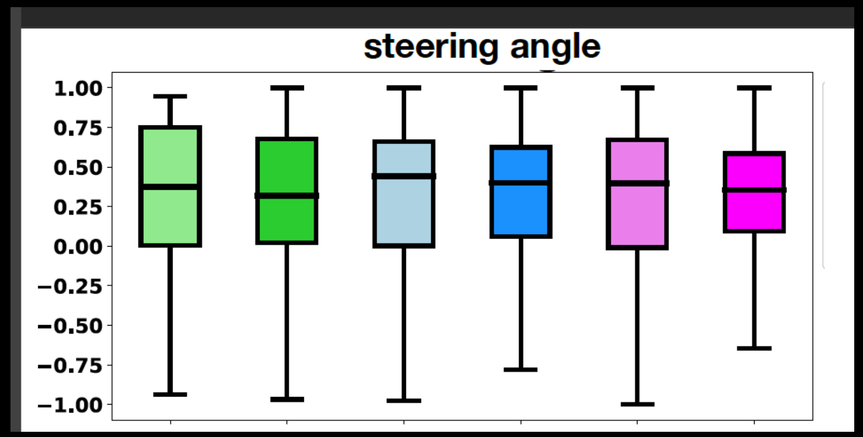
Real

Behaviour
Scenarios
Sensors

Same lane-keeping model

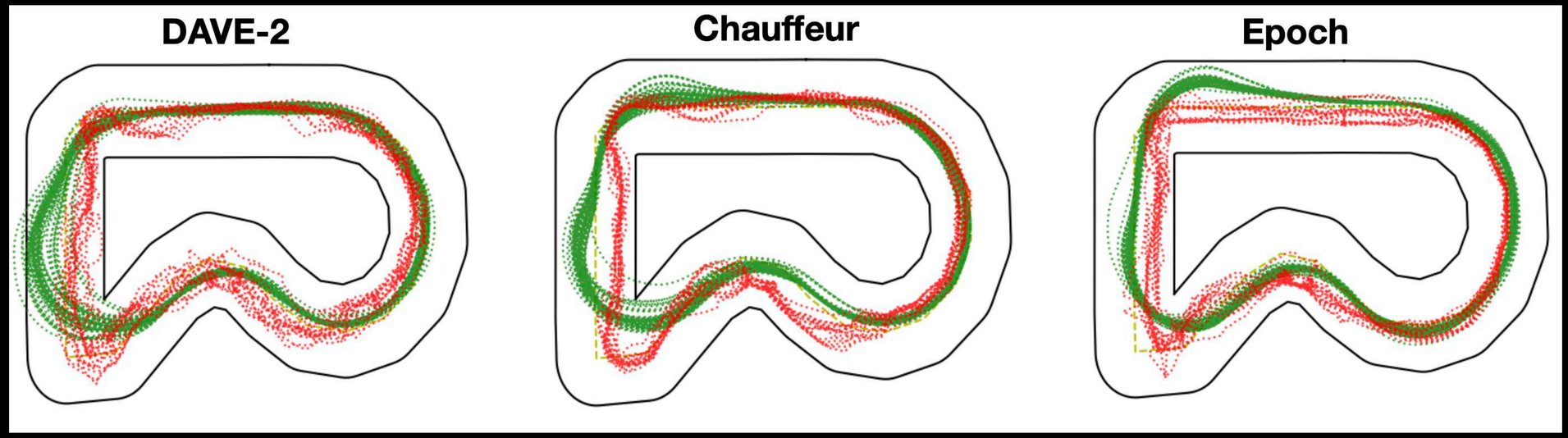
architecture, two worlds

Would the same driving model behave the same?



Steering angle distributions do transfer across simulated and real-world environments

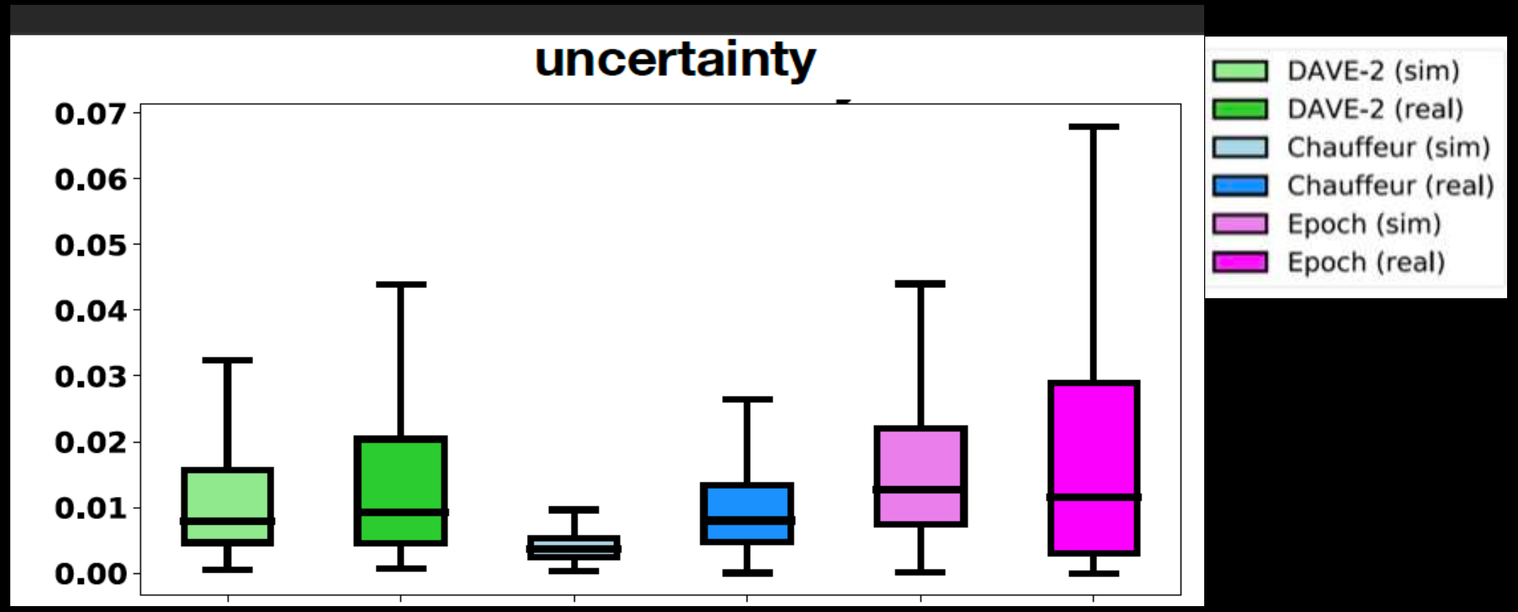
Expected as we aligned the two environments.
It may suggests that component-level testing is an option but…



Virtual (green) and physical (red) trajectories.

Lateral position is different across simulated and real-world environments

Component-level testing is not an option, we need system-level testing

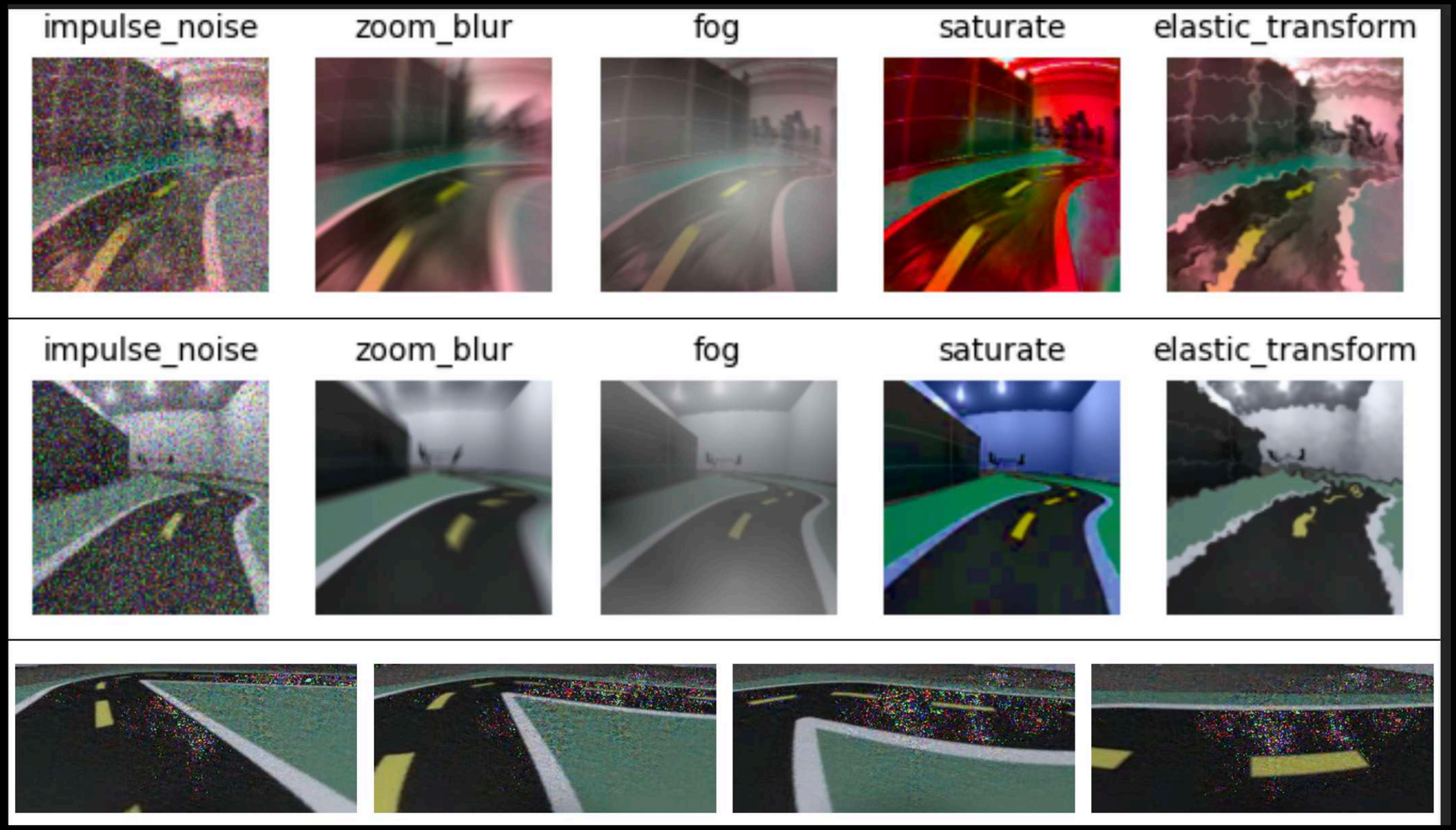


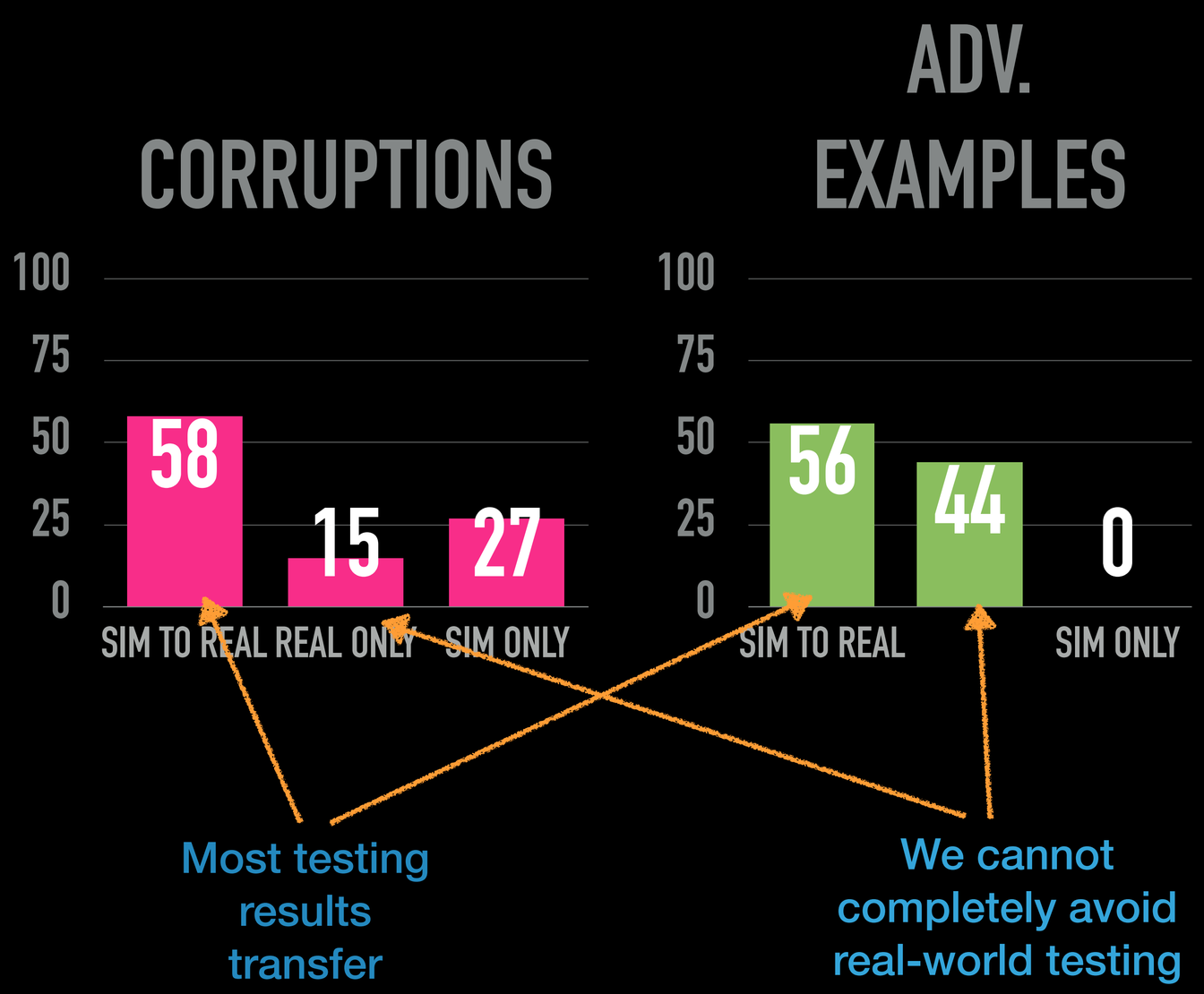
Uncertainty is higher in real-world environments

We need real-world testing (or better simulators)!

Would it fail the same?

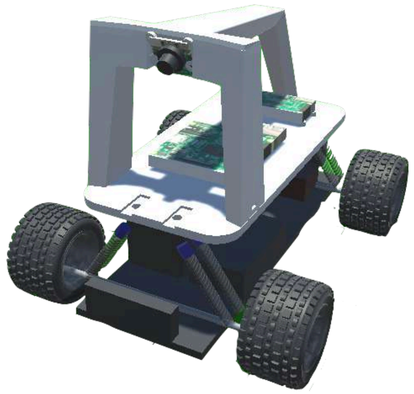
We test both simulated and real cars under the same conditions (corruptions and adv examples)




CORRUPTIONS
100
75
50
25
0
58
15
27
SIM TO REAL
REAL ONLY
SIM ONLY
ADV. EXAMPLES
100
75
50
25
0
56
44
0
SIM TO REAL
SIM ONLY
Most testing results transfer
We cannot completely avoid real-world testing

Gap mitigation

Simulation testing



+

Real-world testing

=

Mixed-reality

⊕ More realistic than sim

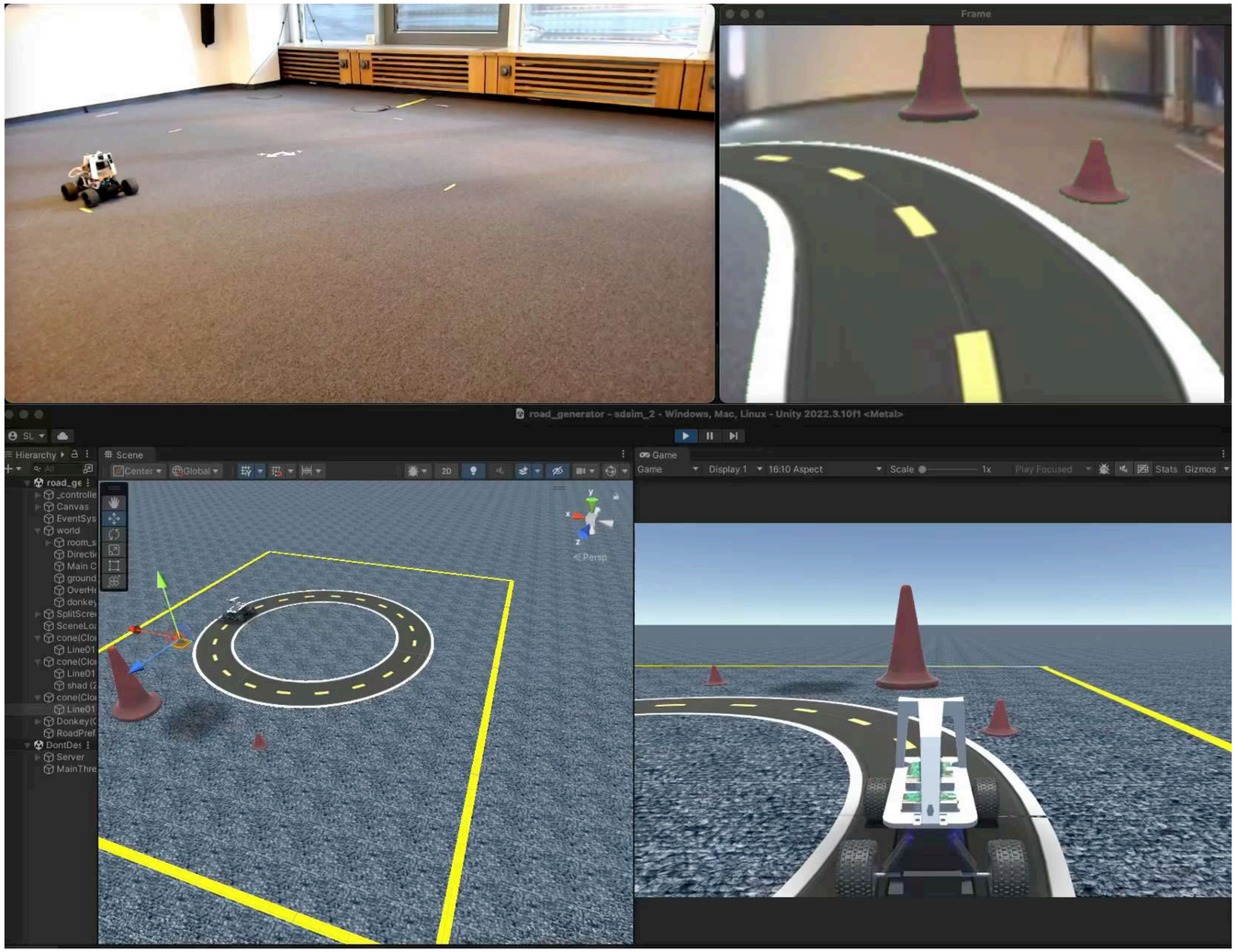
⊕ More versatile than full-scale

Mixed-reality framework

End-to-end ADS execution

Real world

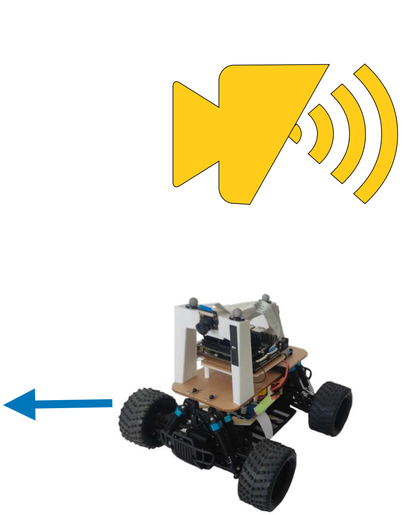
Simulation



Mixed-reality

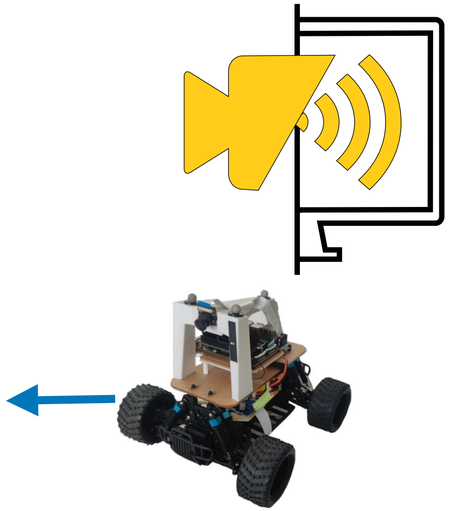
Simulation (FOV)

Testing modalities



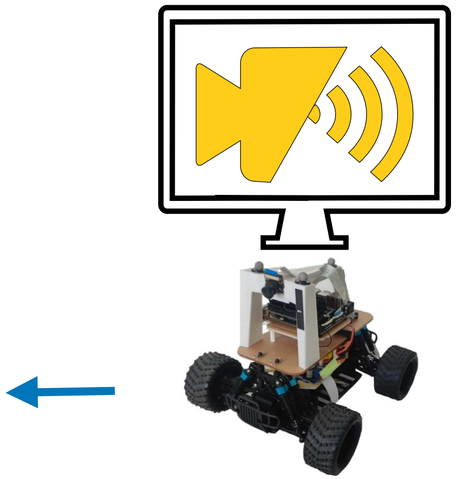
Real World (RW)

Real sensors
Real actuation



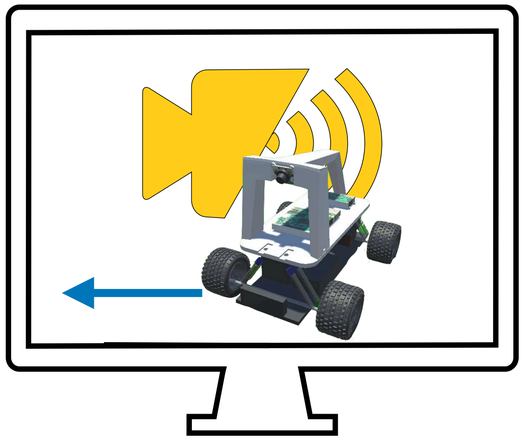
Mixed-Reality (MR)

Mixed sensors
Real actuation



Vehicle in-the-loop (ViL)

Simulated sensors
Real actuation



Simulator in-the-loop (SiL)

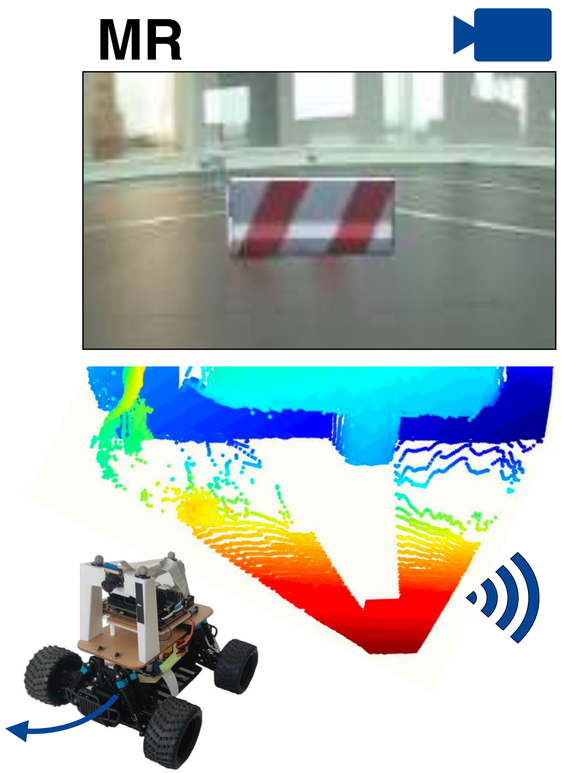
Simulated sensors
Simulated actuation

Our GOAL | Understand the ability of SiL, ViL, and MR to mitigate the actuation gap & perception gap

Testing modalities

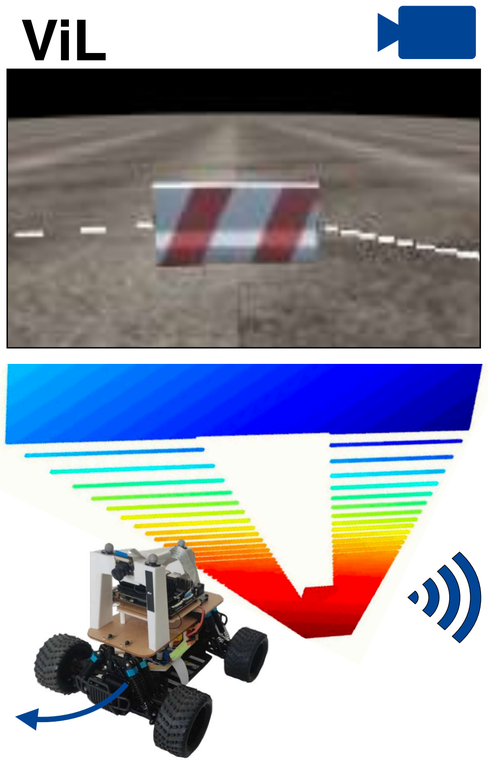
Real World
(RW)

Real sensors
Real actuation



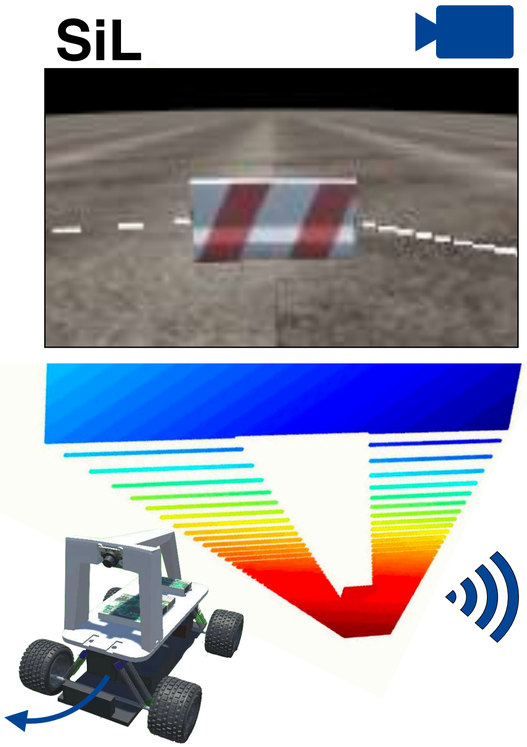
Mixed-Reality
(MR)

Mixed sensors
Real actuation



Vehicle
in-the-loop
(ViL)

Simulated sensors
Real actuation



Simulator
in-the-loop
(SiL)

Simulated sensors
Simulated actuation

Chair of Software Engineering for Data-intensive Applications
School of Computation, Information and Technology
Technical University of Munich

Takeaways

1. **MR best matches real behaviour, ViL closes the actuation gap to real dynamics, SiL misrepresents outcomes due to oversimplified dynamics**

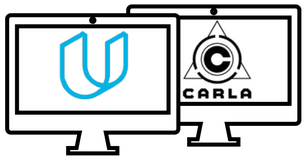
2. **Evaluation is architecture-dependent and unreliable**
   - **E2E systems appear worse in simulation (under-confident, fail in SiL but succeed in RW)**
   - **Modular systems appear better in simulation (over-confident, succeed in SiL but fail in RW), hence the same simulator gives opposite conclusions**

3. **The perception gap, not actuation, is the dominant source of failure**

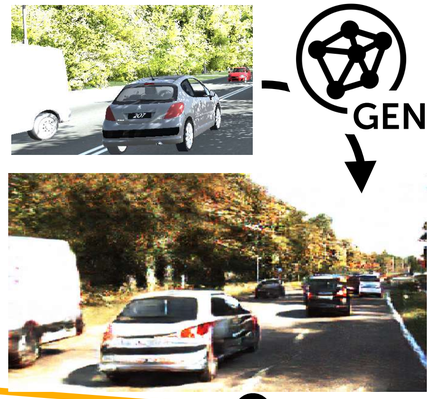
A Multi-Modality Evaluation of the Reality Gap in Autonomous Driving Systems. Stefano Carlo Lambertenghi, Mirena Flores Valdez, and Andrea Stocco. In Proceedings of the 40th IEEE/ACM International Conference on Automated Software Engineering, 2025

Reality Gap Mitigation

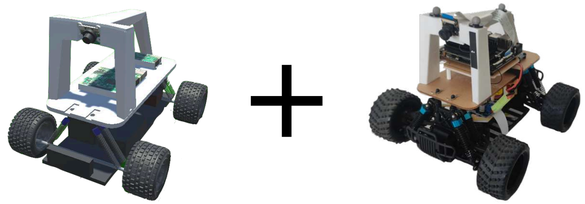
Generic simulators



Generative image translation (ICST 24)



Small-scale real-world & Mixed-reality testing (TSE 2023, EMSE 2023, ASE 2025, ASE 2026)

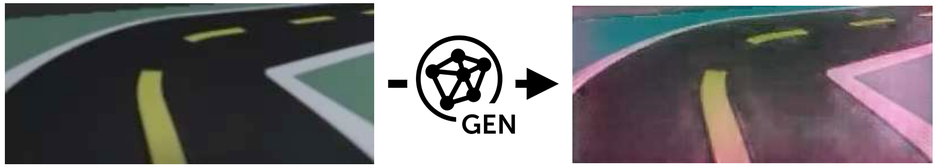


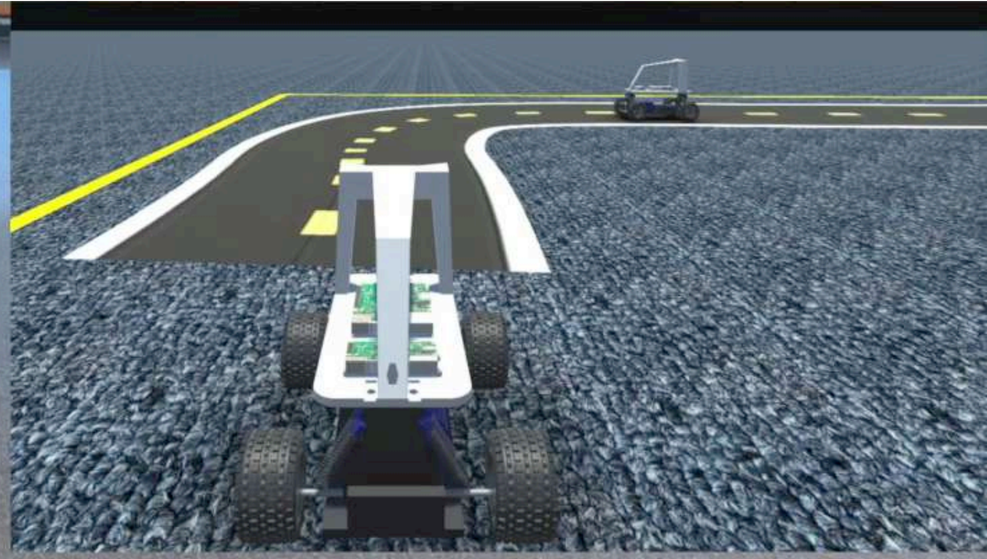
Real-world testing (ASE 2026)

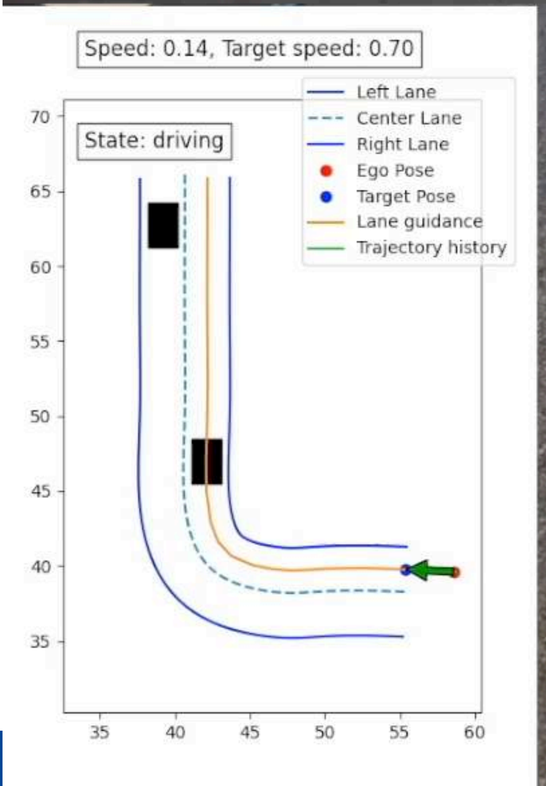
Sim

Real

Generative data-driven simulator (ICSE 2025, TSE 2023, EMSE 2023, ASE 2026)

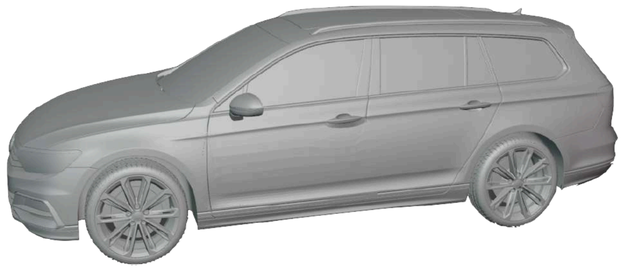





Speed: 0.14, Target speed: 0.70
State: driving
Left Lane
Center Lane
Right Lane
Ego Pose
Target Pose
Lane guidance
Trajectory history
70
65
60
55
50
45
40
35
35
40
45
50
55
60

Gap mitigation

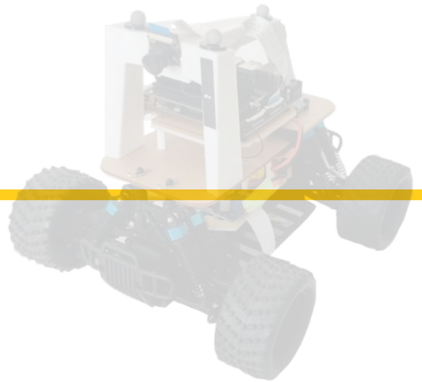
Simulation testing



Simulated

Behaviour
Scenarios
Sensors

Small-scale
Real-world testing



Real

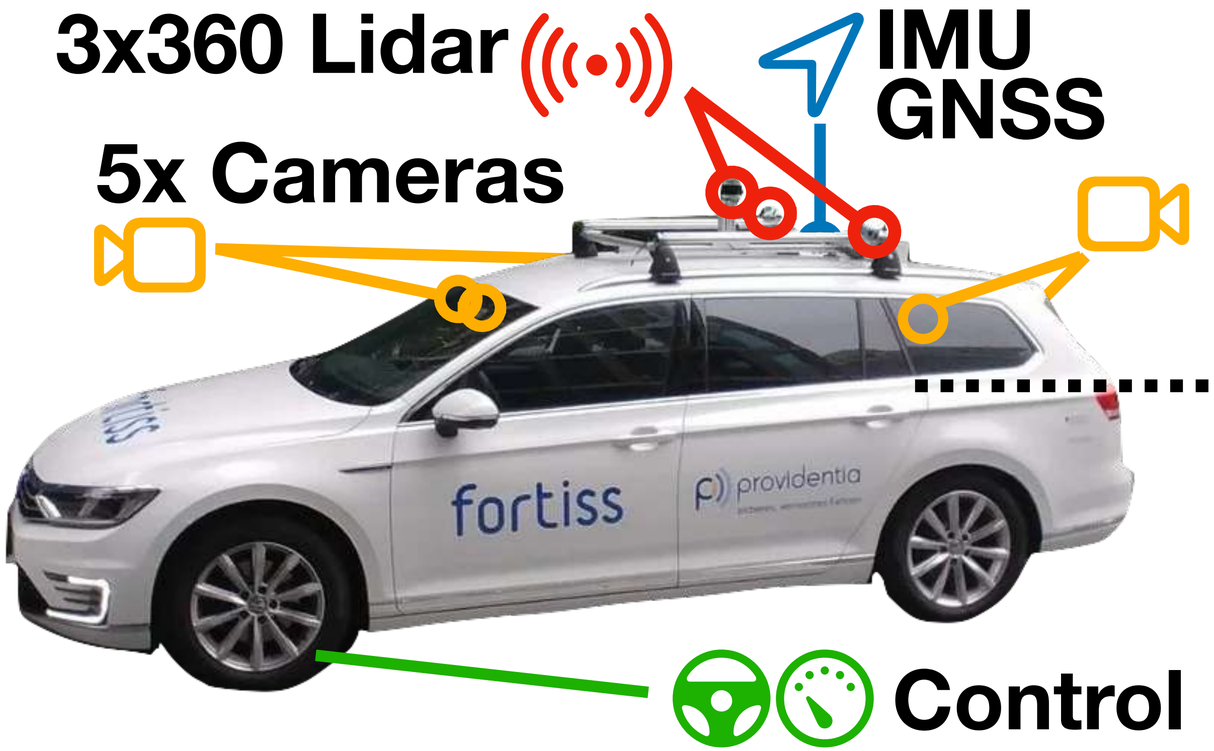
Behaviour
Sensors

Full-scale
Real-world testing

Real

Behaviour
Scenarios
Sensors

Real World Perturbation Testing



Volkswagen Passat GTE

Ultra7 - 64 GB - RTX5080

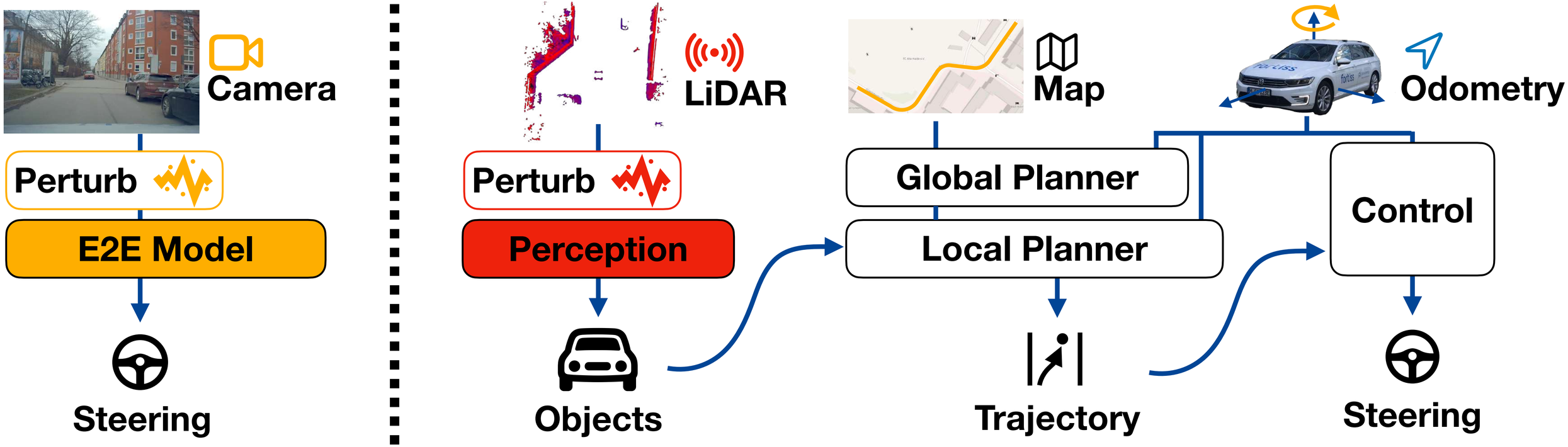
Nvidia Drive PX 2

i7 - 32 GB - GTX1050 Ti

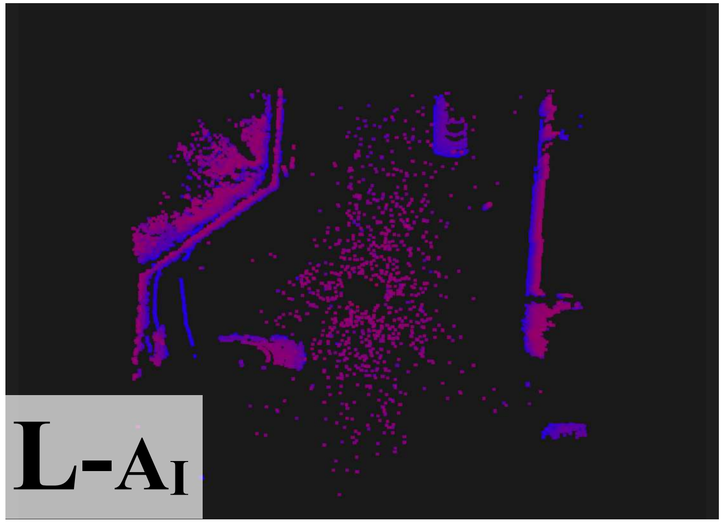
ROS i7 - 32 GB - GTX1050 Ti

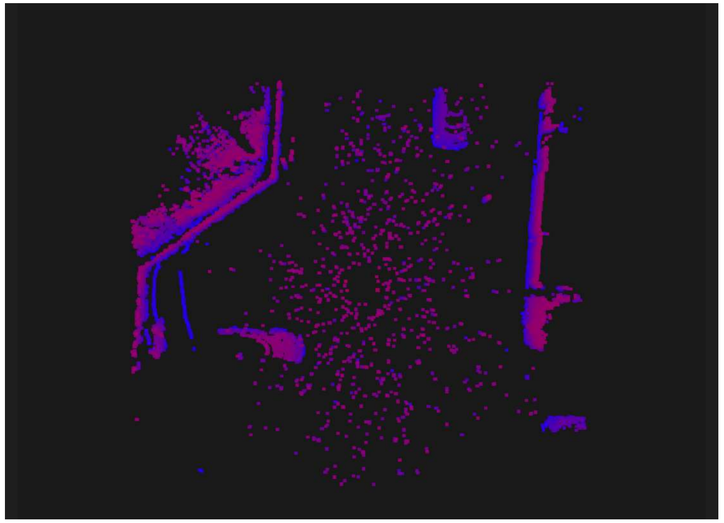
dSpace Micro Autobox II

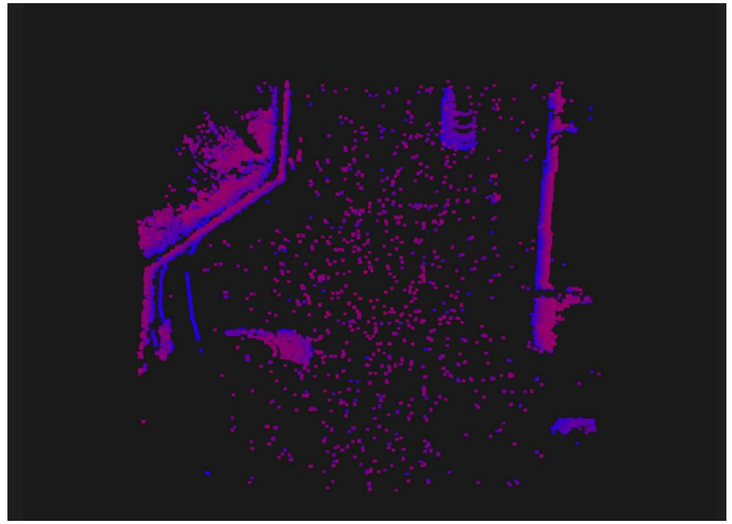
Real World Perturbation Testing



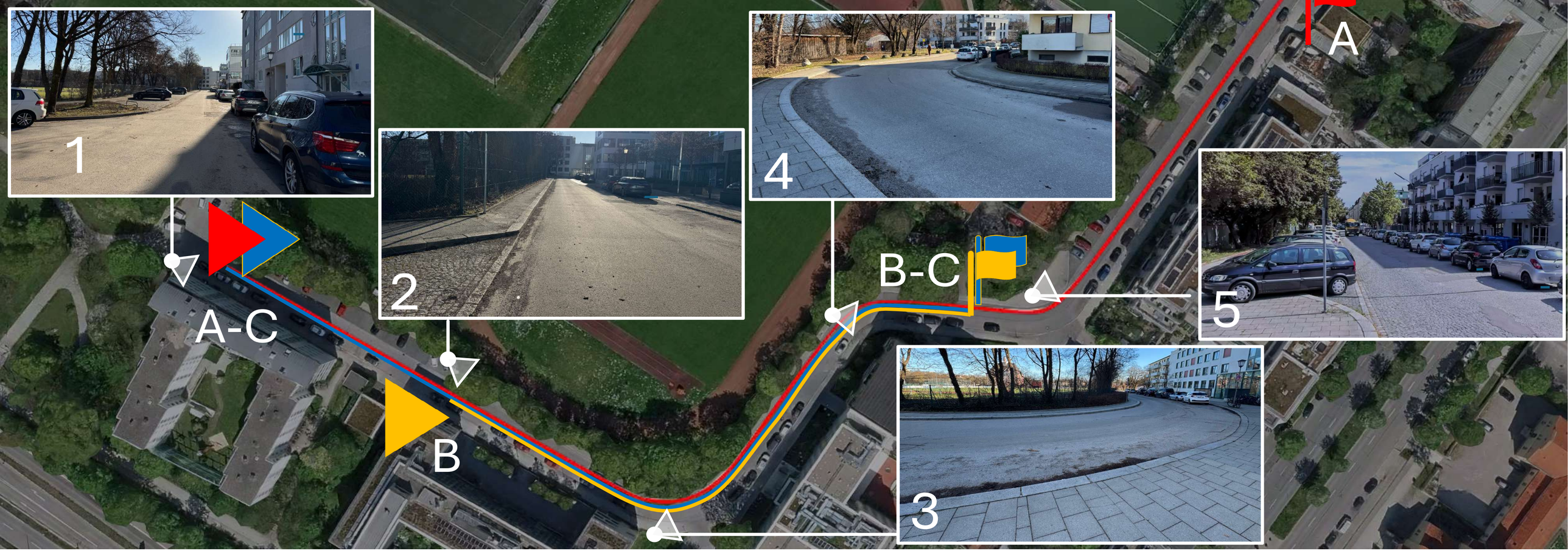
I1 I3 I5

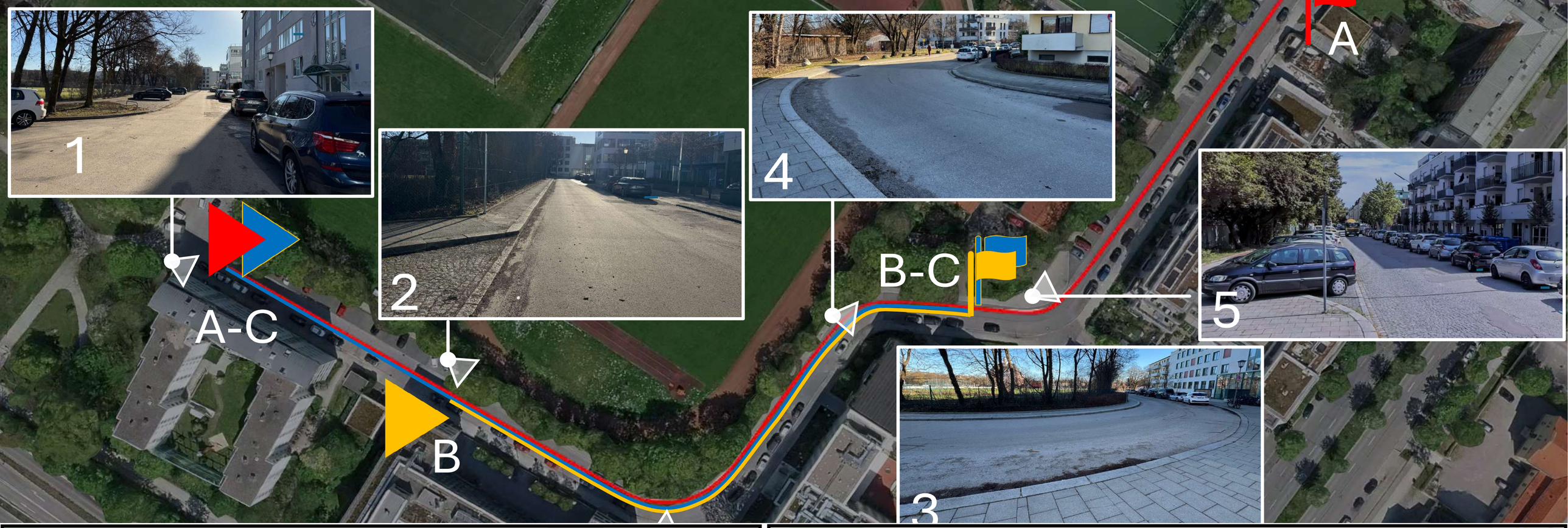


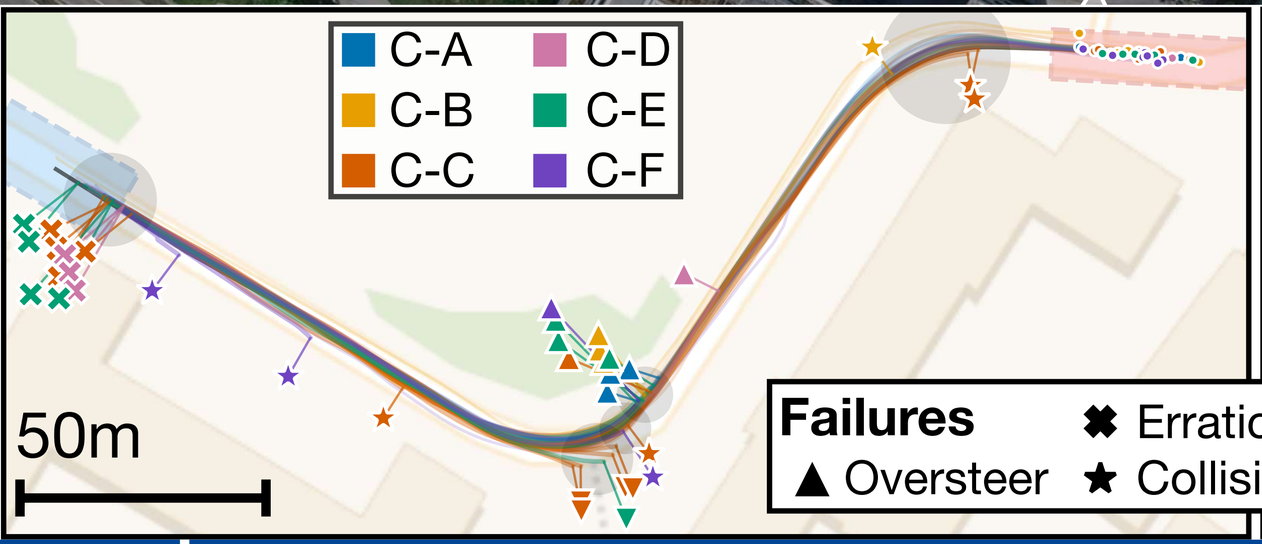


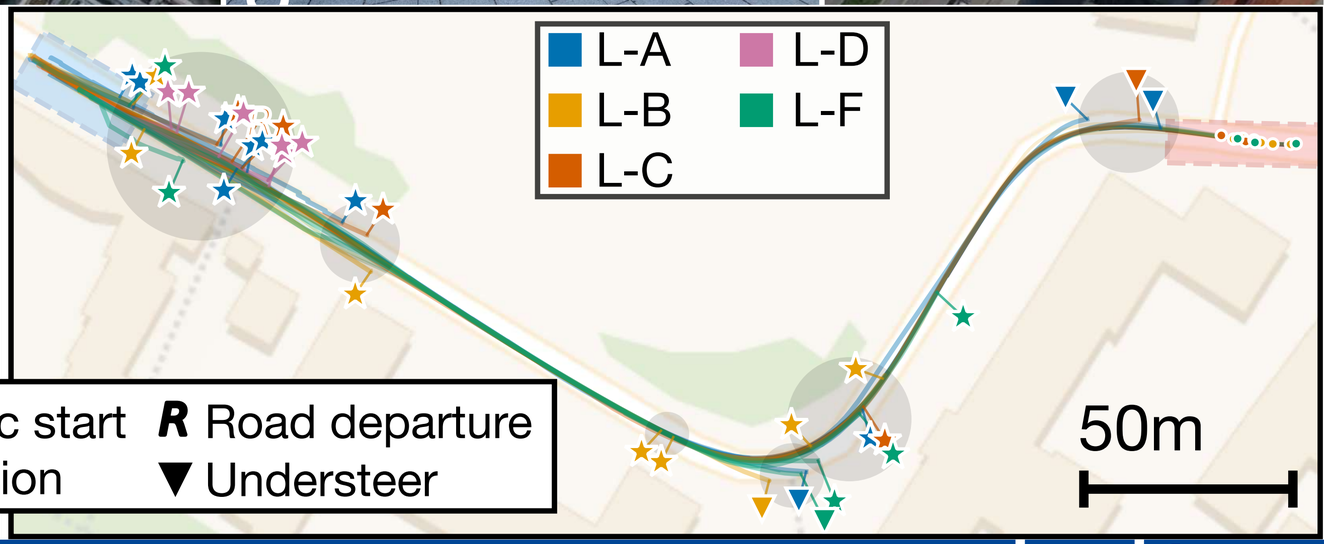


I1 I2 I3


1
2
3
4
5
A
A-C
B
B-C







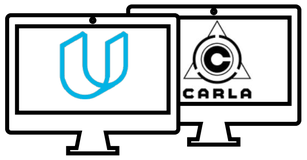
Failures

✖ Erratic start ***R*** Road departure

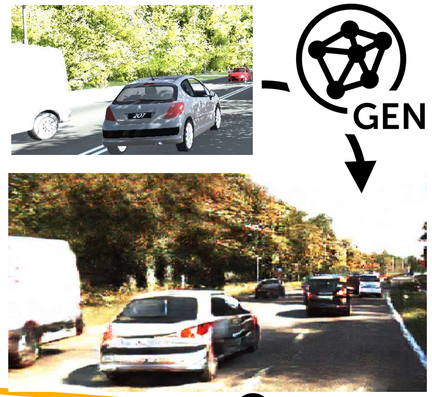
▲ Oversteer ★ Collision ▼ Understeer

Reality Gap Mitigation

Generic simulators



Generative image translation (ICST 24)



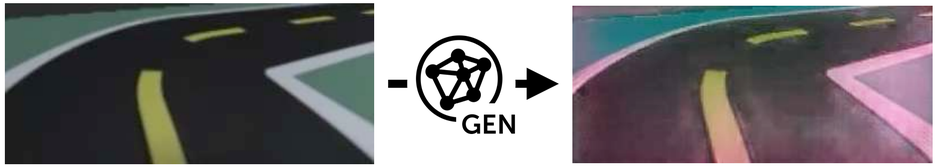
Small-scale real-world & Mixed-reality testing (TSE 2023, EMSE 2023, ASE 2025, ASE 2026)

Real-world testing (ASE 2026)

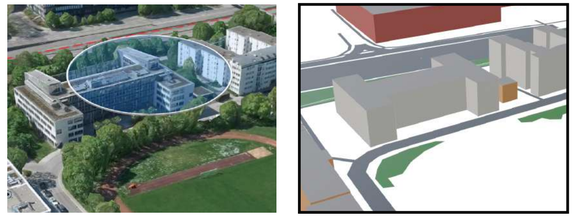
Sim

Real

Generative data-driven simulator (ICSE 2025, TSE 2023, EMSE 2023, ASE 2026)



Scenario Twins (ASE 26)



Our techniques

Measure

Mitigate

Generic simulators

Sensors
Physics
Scenarios

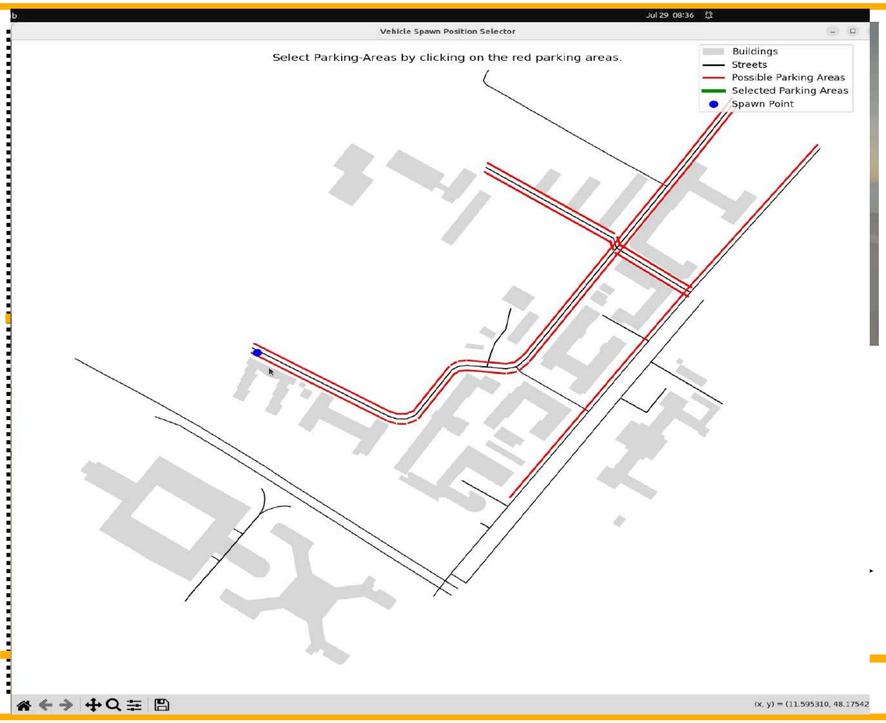
LLaVA

SegFormer

Dataset

Open Street Map

.osm



CARLA

ADS

Sim

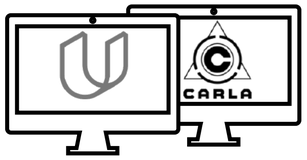
Real

Our techniques

Measure

Mitigate

Generic simulators



CARLA

Stable diff

ControlNet (Seg.)

ControlNet (Temporal)

Sim

Real

Neural Reconstruction

From 2D datasets to 3D simulation

Neural Reconstruction

From 2D datasets to 3D simulation

Real-world Testbeds

From prototype to the real world

Takeaways

1. Reality gap affects scenarios, sensors, and physics

2. Tackle one at a time (e.g., genAI for sensors). Metrics!

3. World Models, Physical AI?

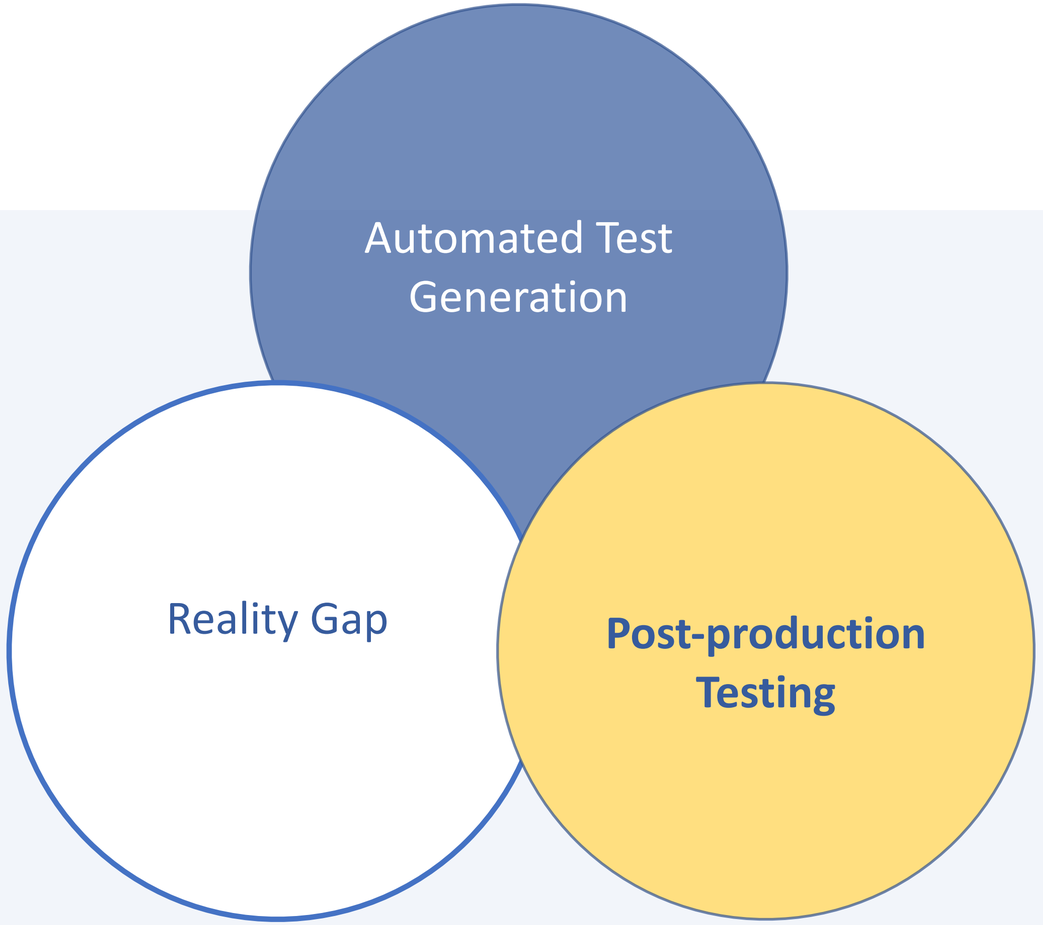
Automated Software Testing

Main research topics

- **Automated Test Generation**
  How can we automatically generated complex scenario-based tests efficiently and effectively?

- **Reality Gap**
  How can we measure and mitigate the reality gap between simulation and in-field testing?

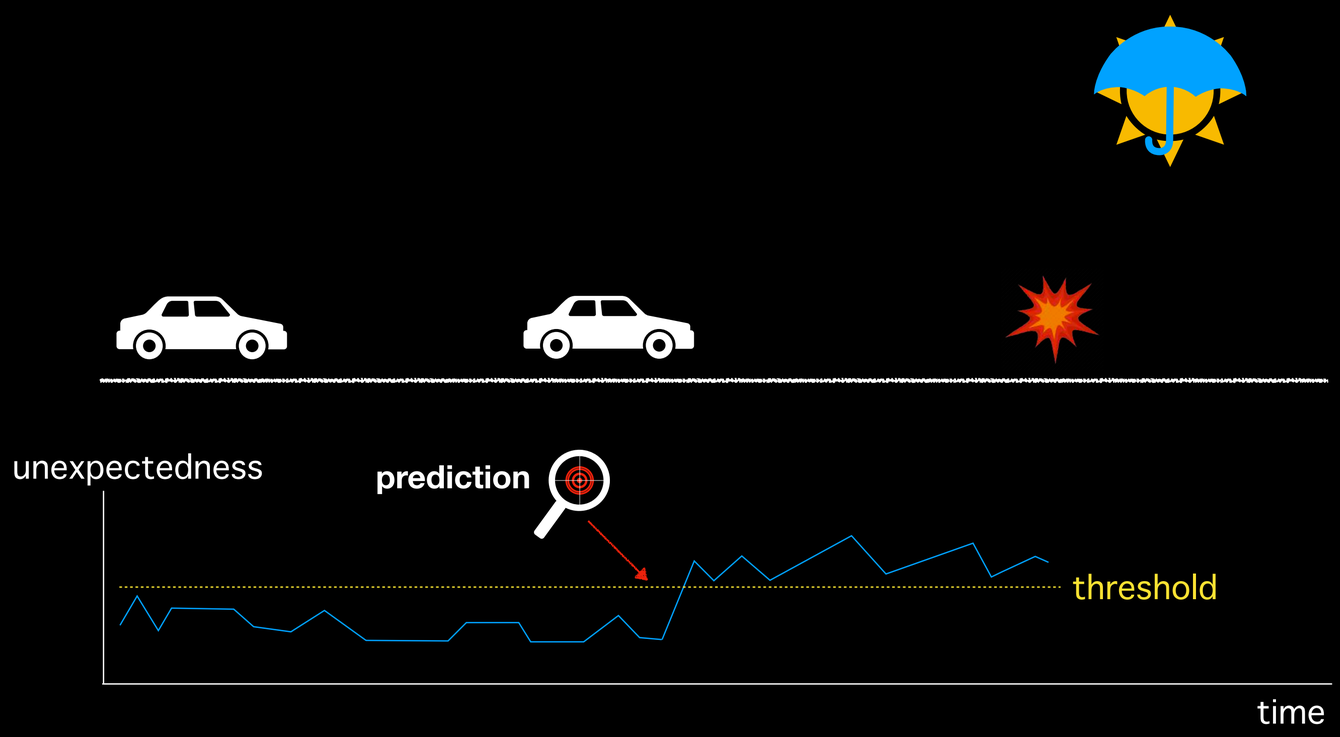
- **Post-production Testing**
  How to ensure a high dependability of deep neural network driven-cyber-physical systems (CPS) in production?




Steering Angle predicted by the DNN
-0.16933852434158325
-0.2953818142414093
-0.2953818142414093
-0.2953818142414093
-0.24482464790344238
-0.24482464790344238
-0.24482464790344238
-0.2340604066848755
-0.2340604066848755
-0.2340604066848755
-0.2876757085323334
-0.2876757085323334
-0.2876757085323334
-0.28597092628479004
-0.28597092628479004
-0.28597092628479004
-0.280177503824234
-0.280177503824234
-0.280177503824234
-0.1850987821817398
-0.1850987821817398
-0.1850987821817398
-0.2626234292984009
-0.2626234292984009
-0.2626234292984009
-0.20239685475826263
Mode: Autonomous
* vehicle learnt how to stay in lane
Training set cannot contain ALL possible driving conditions!


Can we **predict unexpected** conditions*?
* **Unexpected condition:**
Unseen, potentially hazardous and misbehaviour-inducing driving condition

Unexpectedness metric



Post-production Testing

Real-time monitoring of deployed ADS

ADS & ADAS

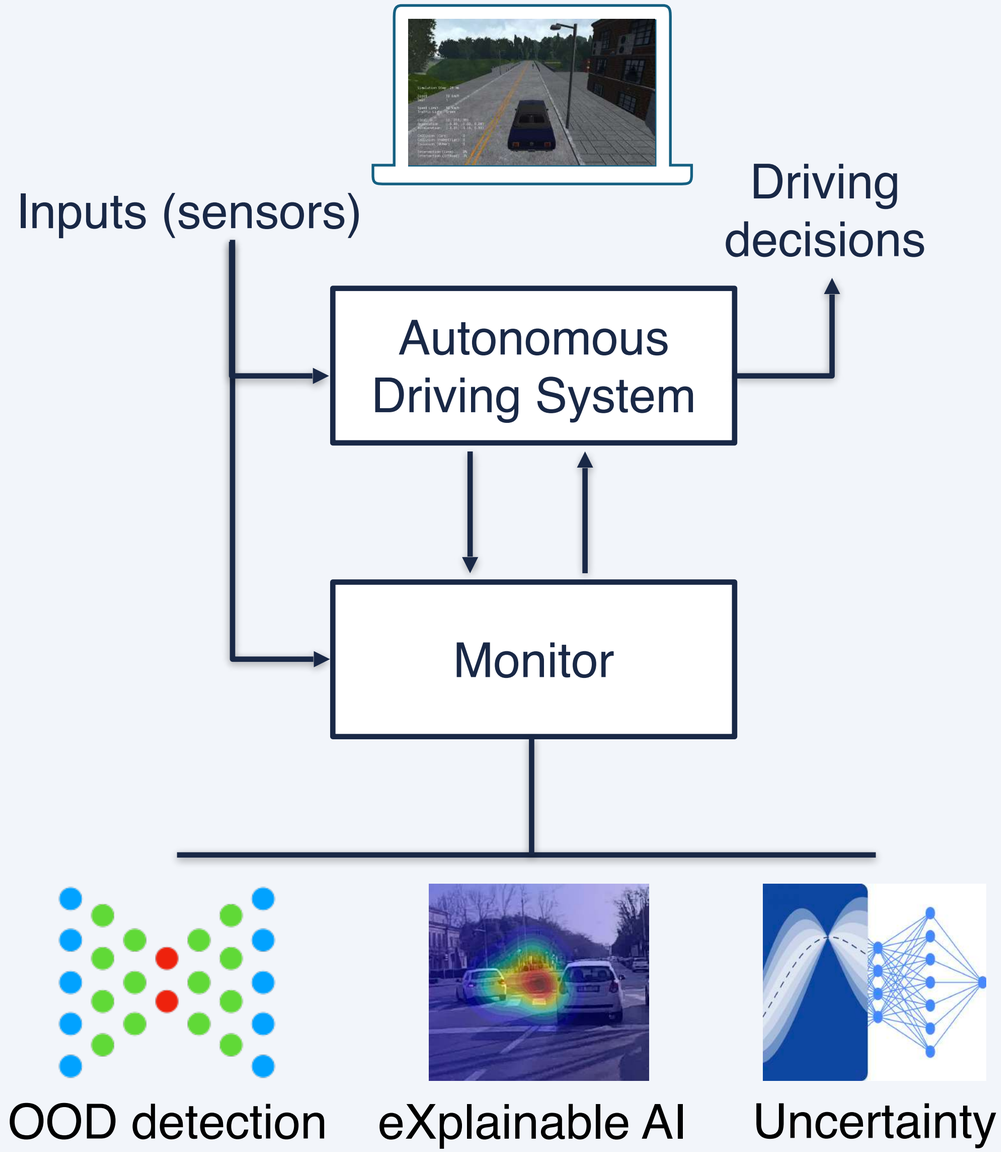
- **End-to-end DNNs (level 2)**
- **Full AD stacks**

Real-time Monitoring

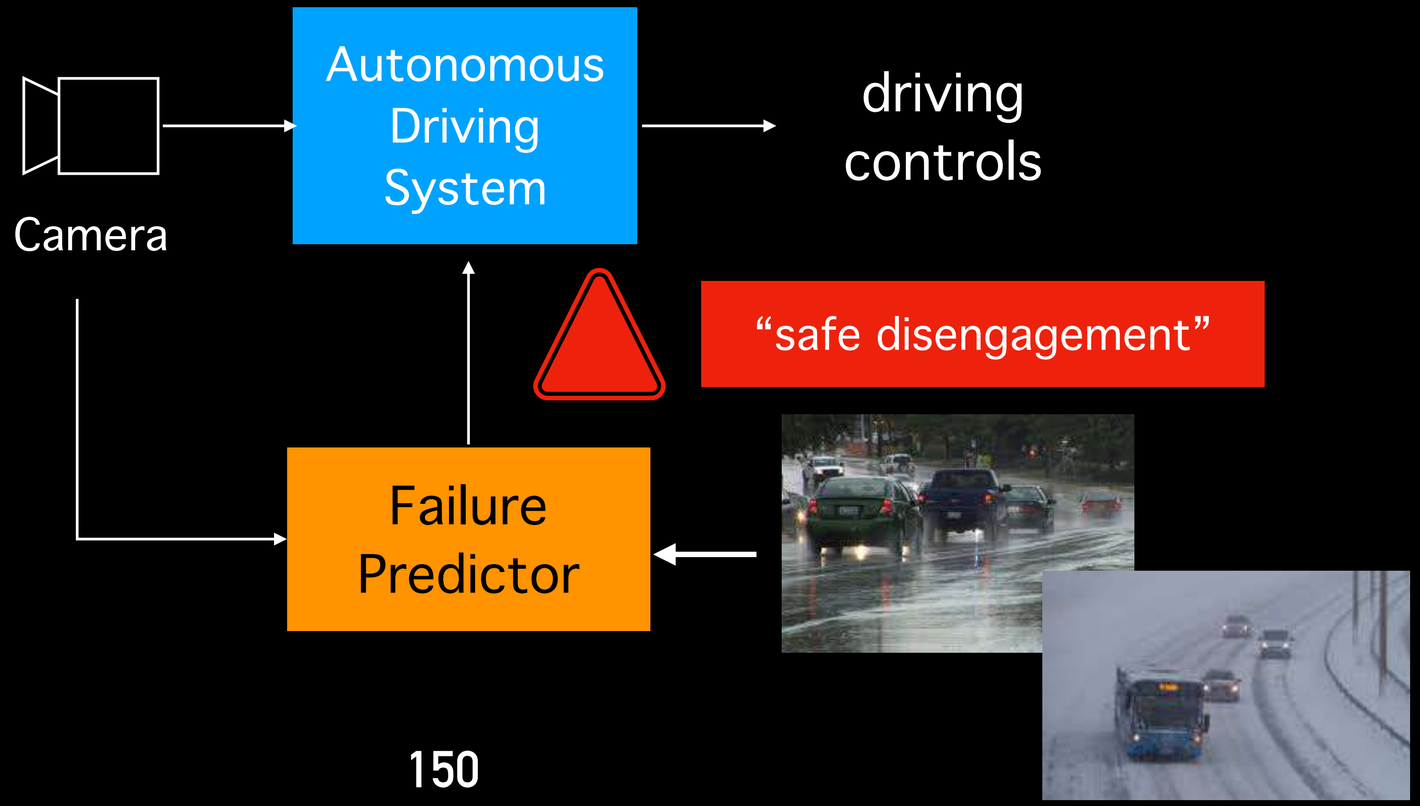
Observe System Under Test at runtime

Trustworthy ADS

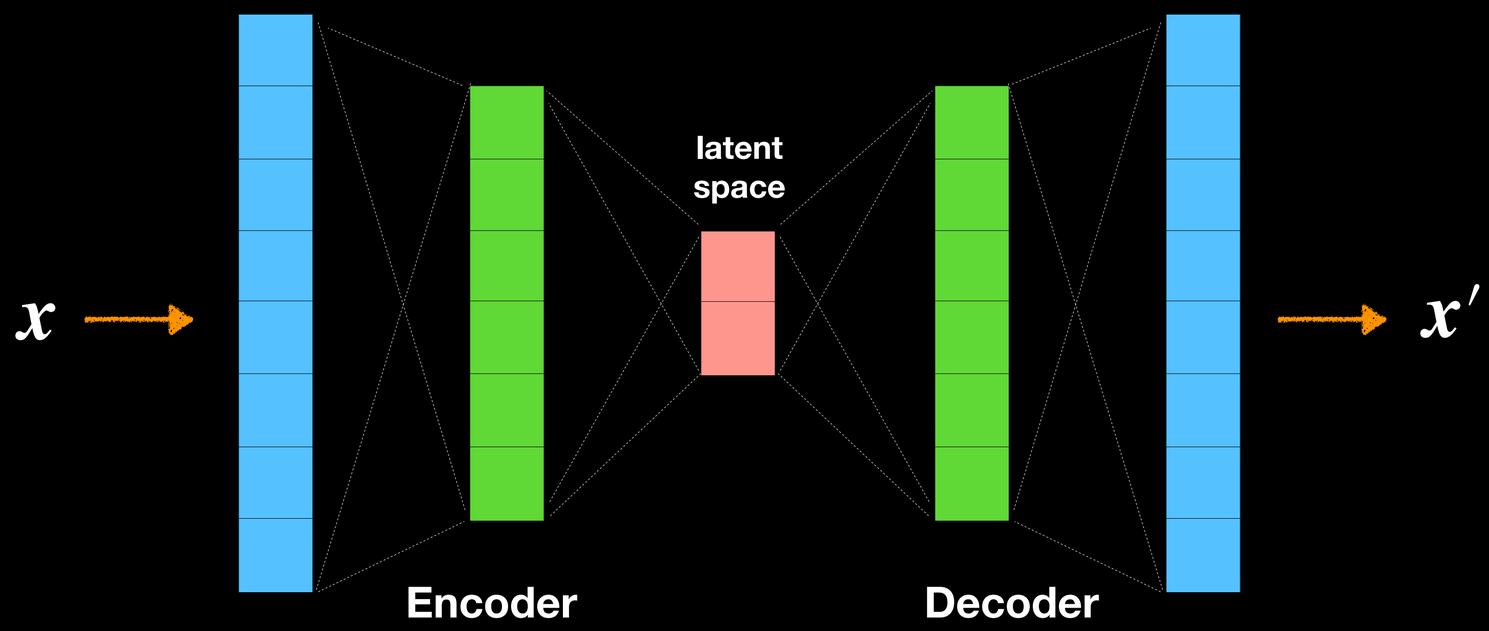
Alert the driver if system is not to be trusted or activate fallback procedures



Black Box Unexpectedness Score



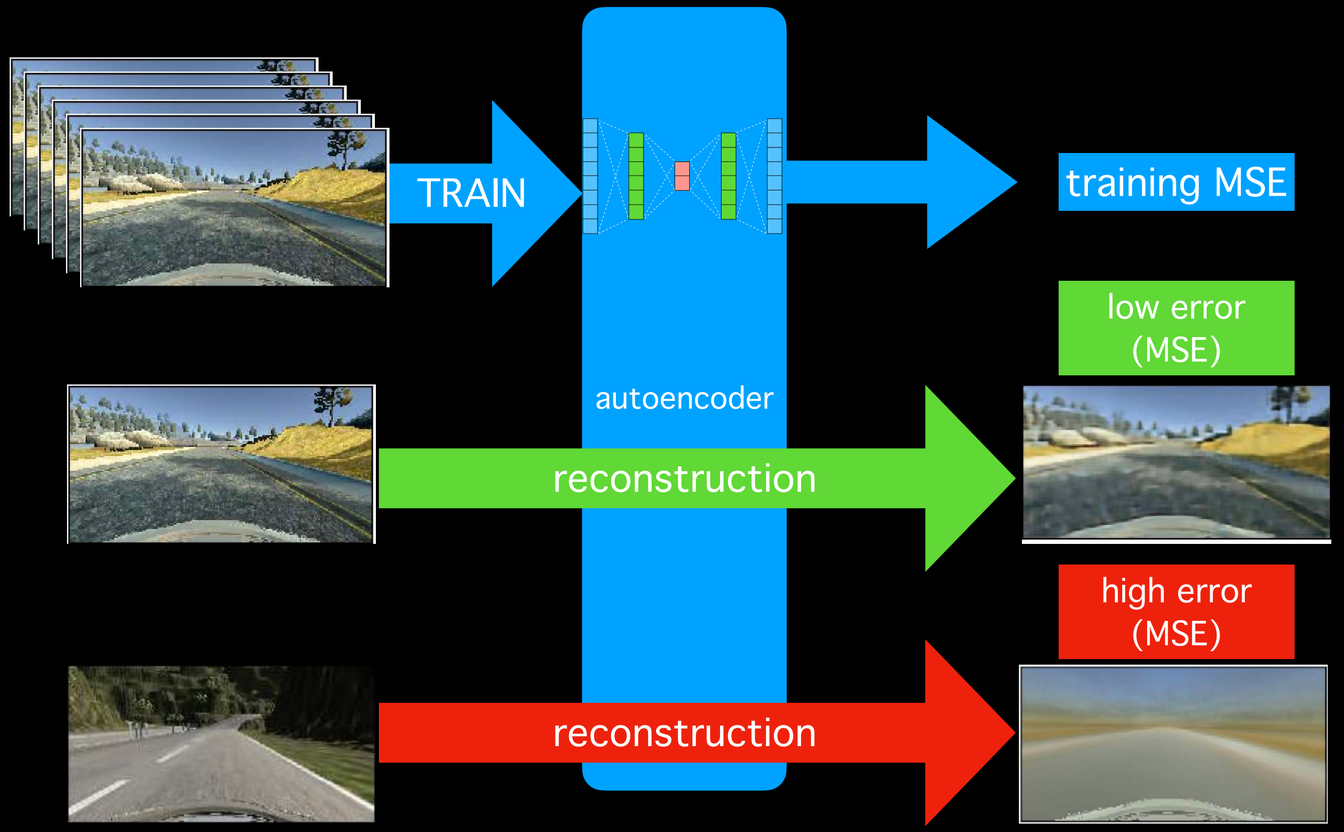
Autoencoders



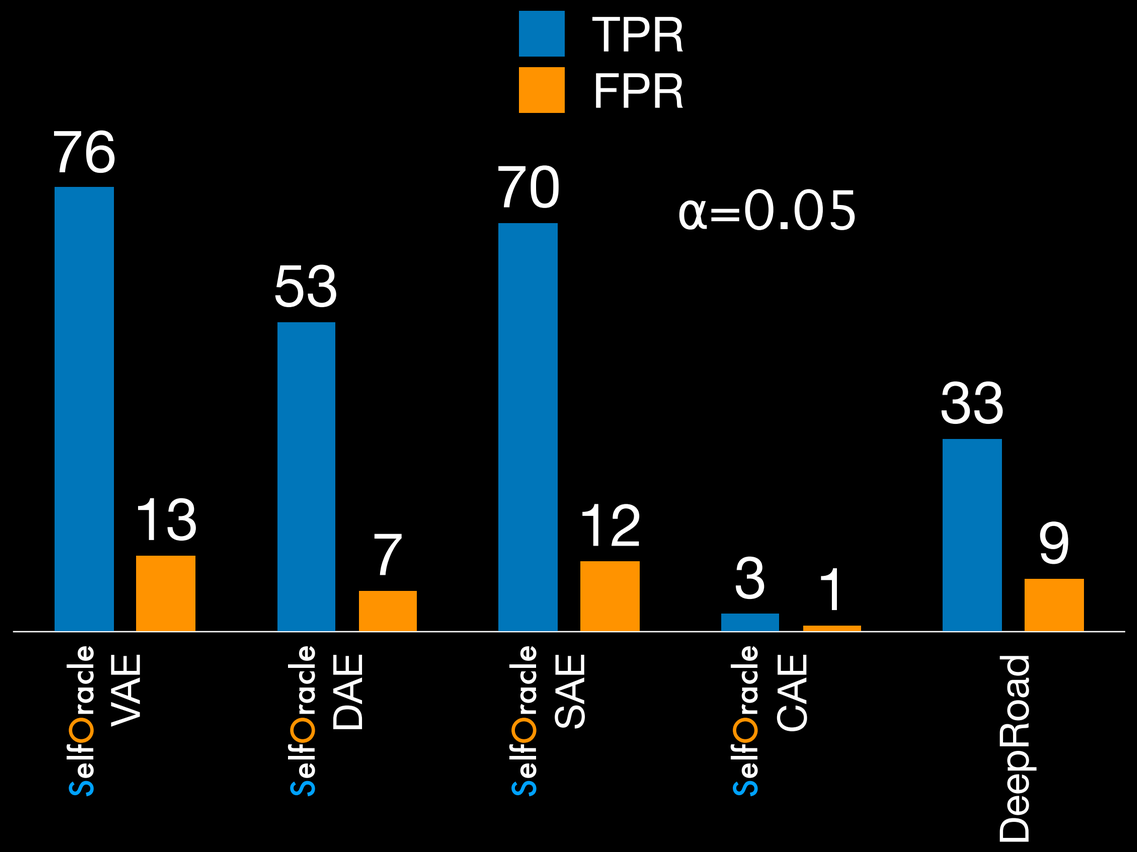
$$MSE(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{n} \sum_{t=1}^{n} \|\boldsymbol{x} - \boldsymbol{x}'\|^2$$

indicator of the difference between the original data and its reconstruction

Autoencoder-based Anomaly Detector



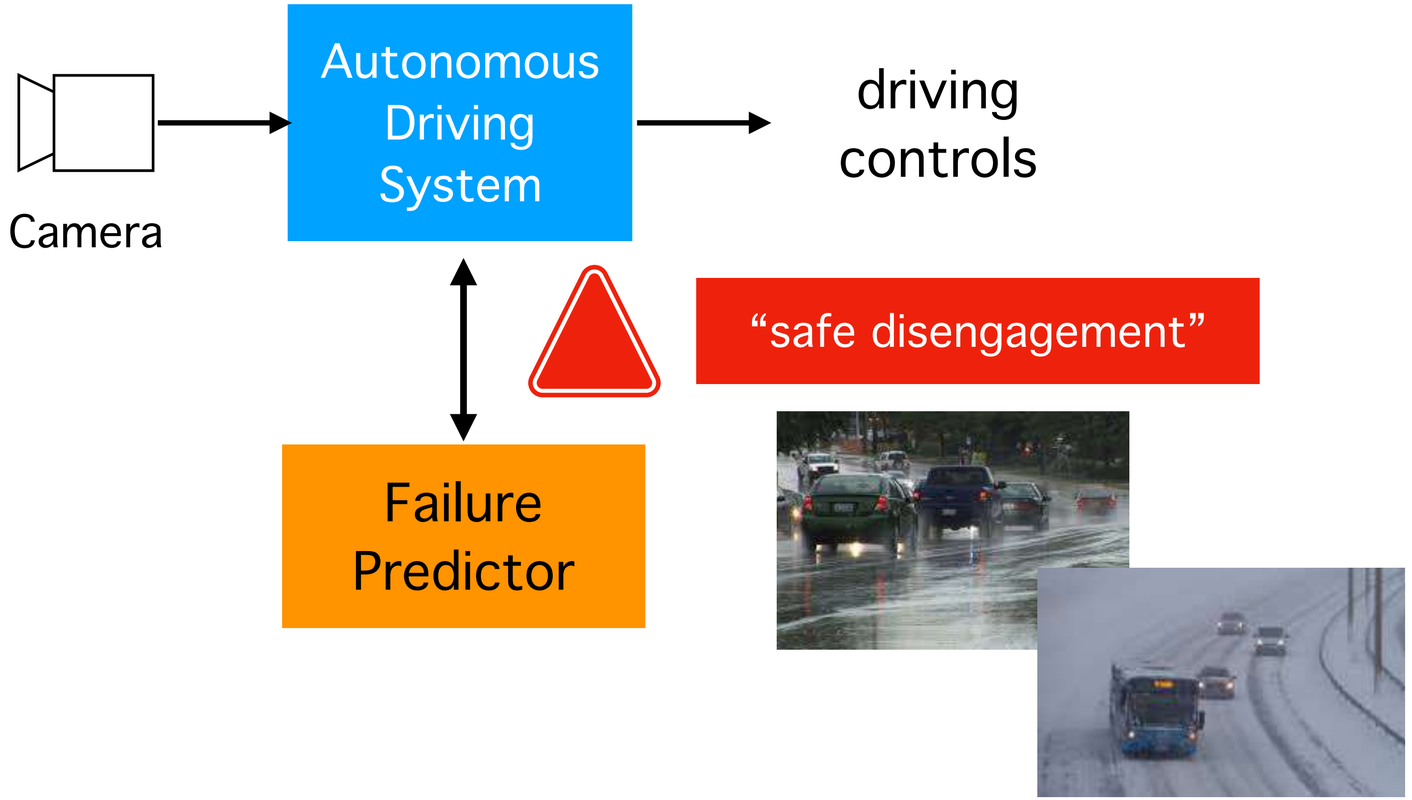
Is our approach effective
in predicting misbehaviours ?



76%

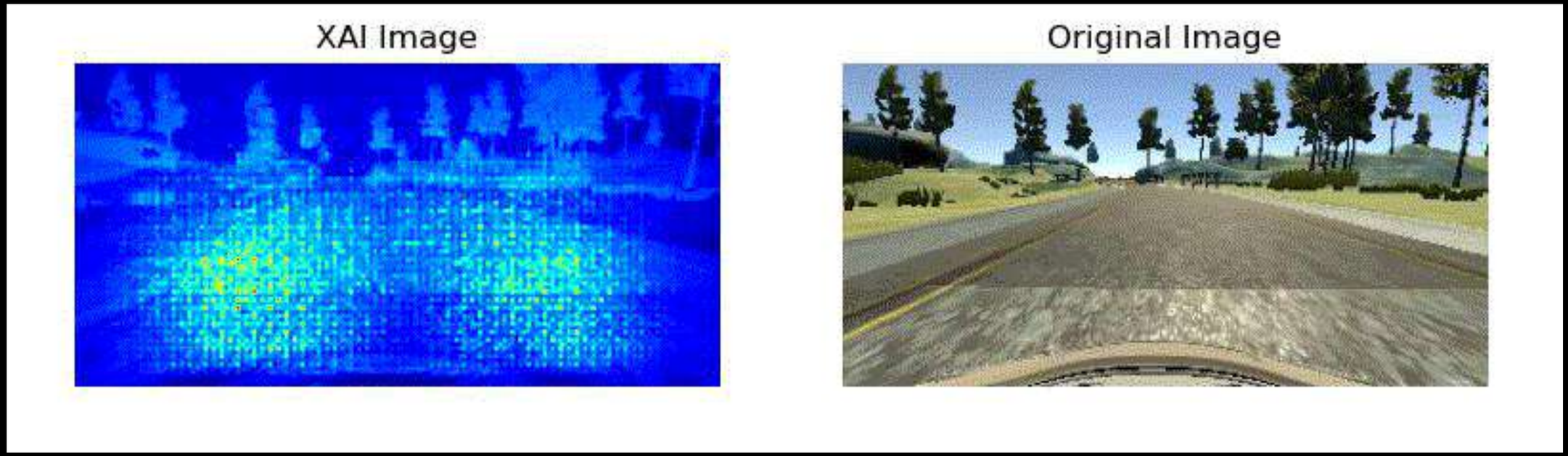
correct *misbehaviour predictions* 6 *to* 9 *seconds before*

White Box Unexpectedness Score



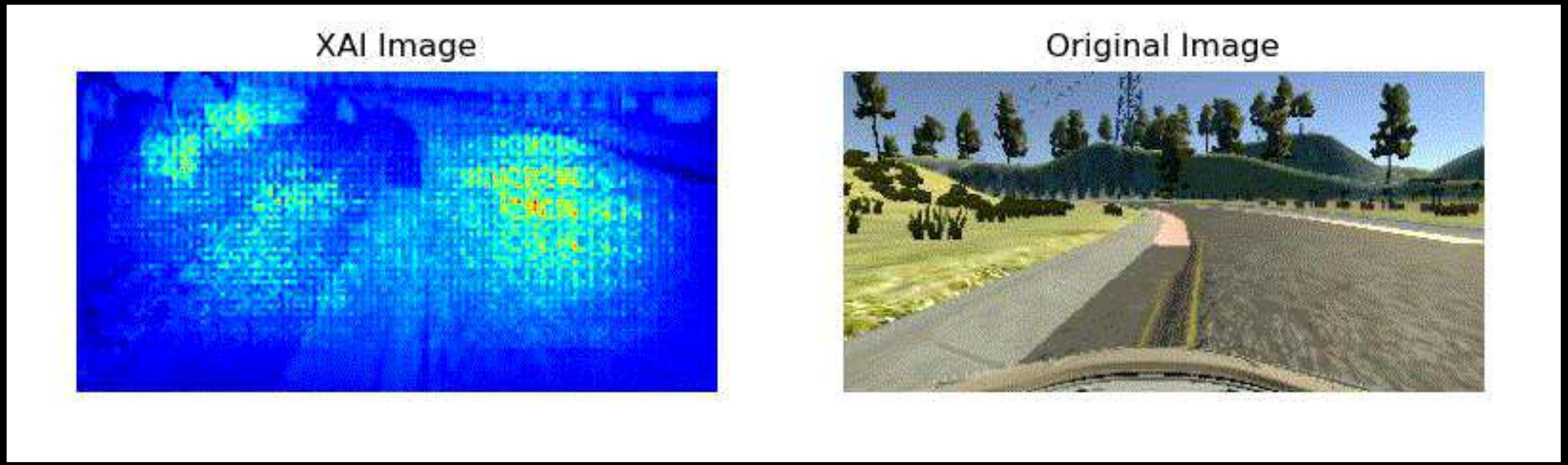
XAI for failure prediction

Nominal



XAI for failure prediction

Uncertain/Unexpected



Confidence Score Synthesis

$$\overline{\boldsymbol{h}} = \frac{1}{WHC} \sum_{i=1,j=1,c=1}^{W,H,C} h_{[i][j][c]}$$

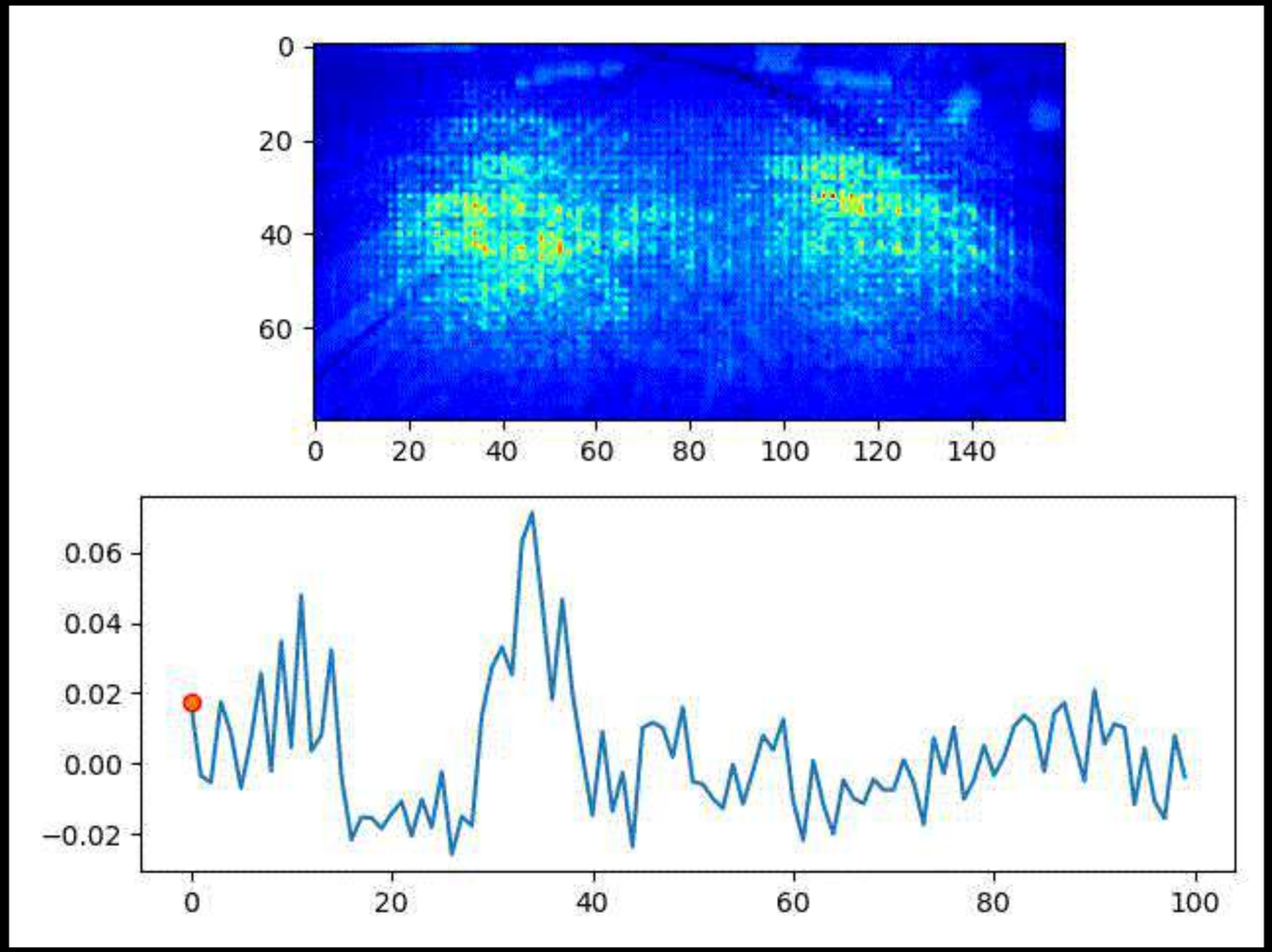
HA = Heatmap
Average Function

$$\overline{\nabla \boldsymbol{h}_t} = \frac{1}{WHC} \sum_{i=1,j=1,c=1}^{W,H,C} h_{t[i][j][c]} - h_{t-1[i][j][c]}$$

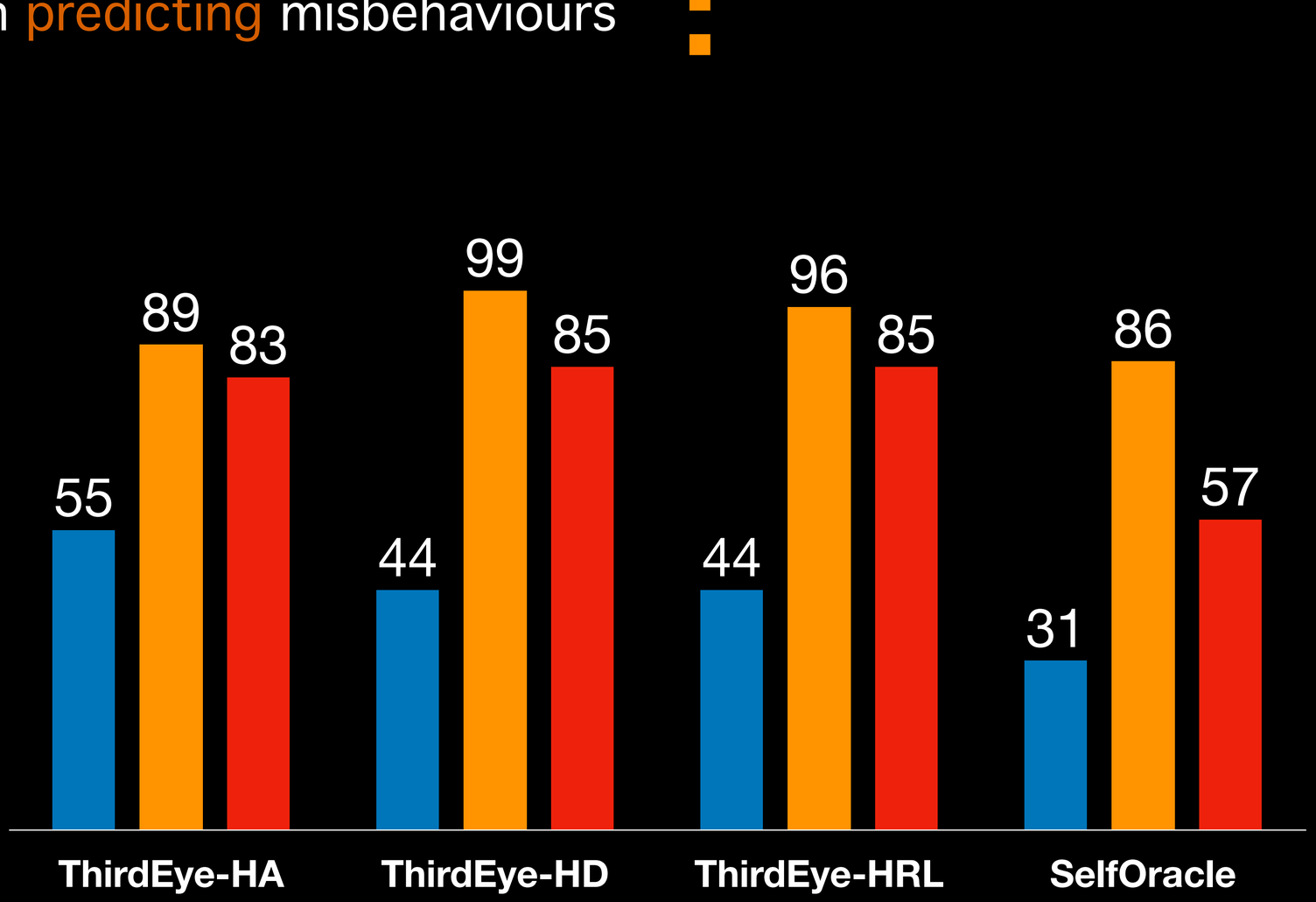
HD = Heatmap
Derivative Function

$$\boldsymbol{h_e} = \mathscr{L}(\boldsymbol{h}, dec(enc(\boldsymbol{h})))$$

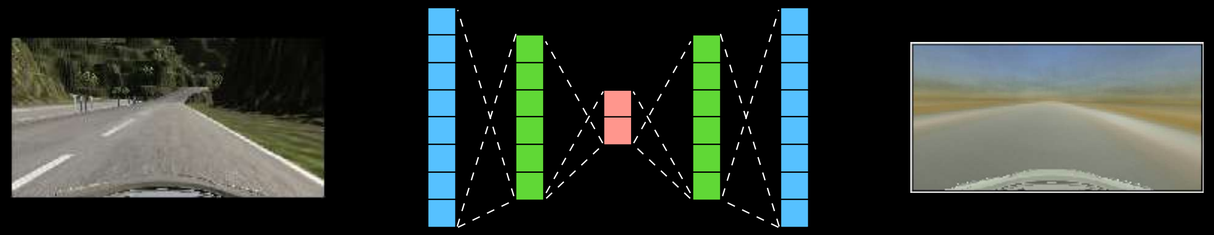
HRL = Heatmap
Reconstruction Loss


0
20
40
60
0
20
40
60
80
100
120
140
0.06
0.04
0.02
0.00
−0.02
0
20
40
60
80
100

Is our approach effective in predicting misbehaviours ?



SelfOracle (ICSE 2020)

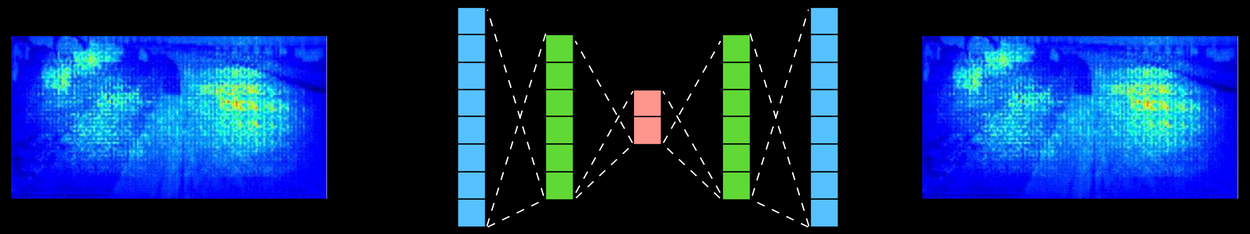


Autoencoder-based

Black-box

Computationally Fast

ThirdEye (ASE 2022)

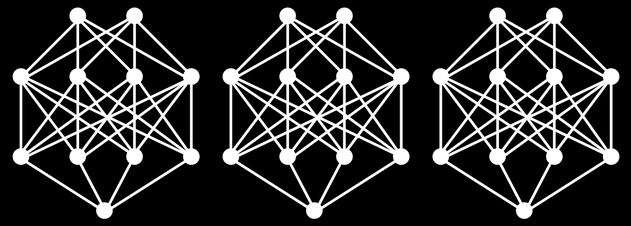


Heatmap-based

White-box

Computationally Expensive

Deep Ensembles



Uncertainty score

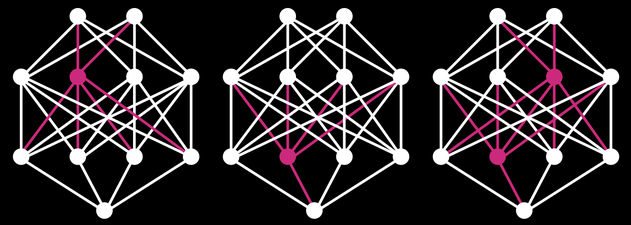
🤔

Theoretically grounded

Widely applicable

Computationally Expensive

Monte-Carlo Dropout



Uncertainty score

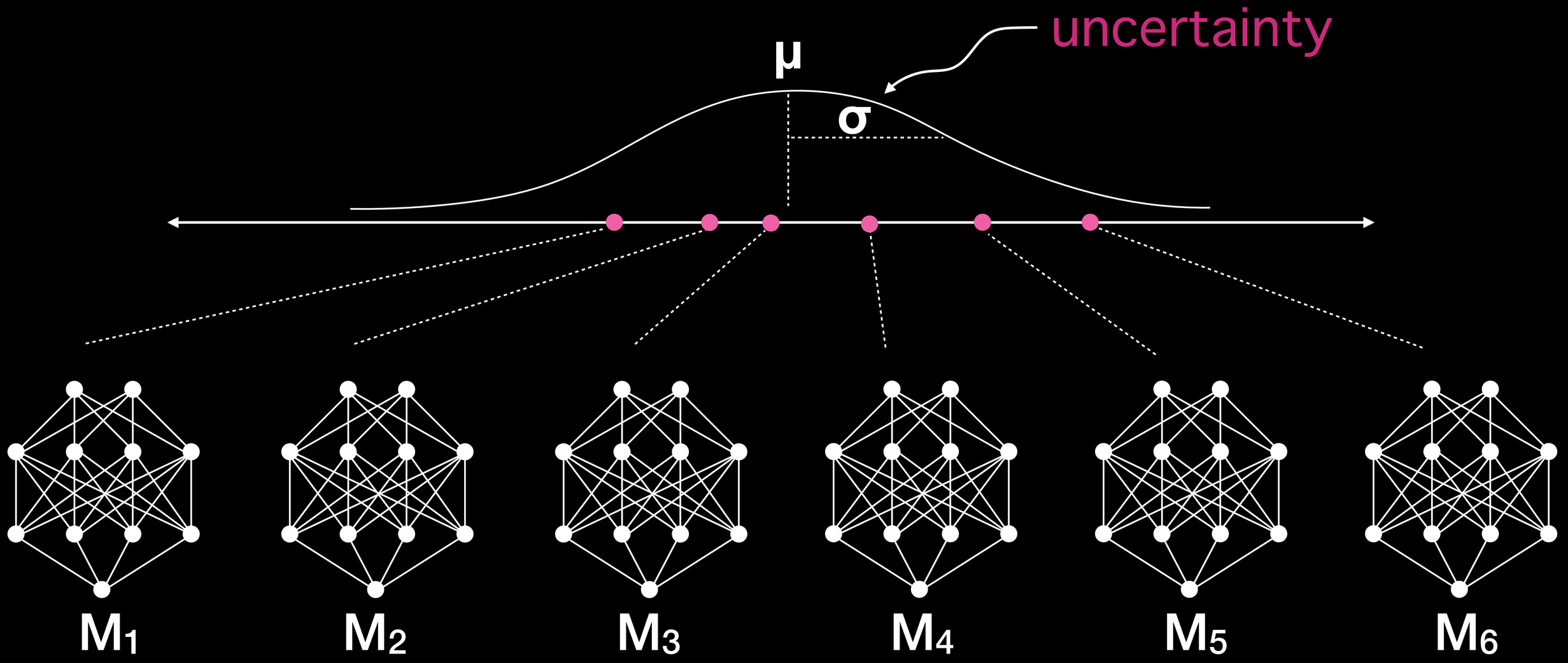
🤔

Approximation of BNNs

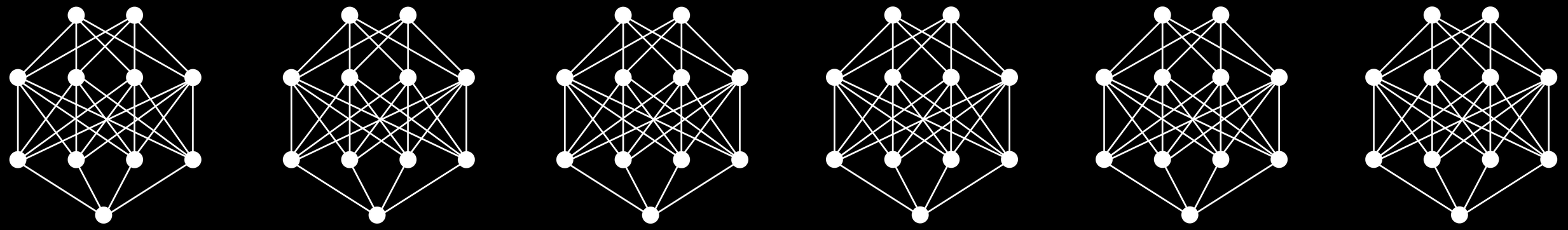
Intrusive Approach

Computationally Feasible

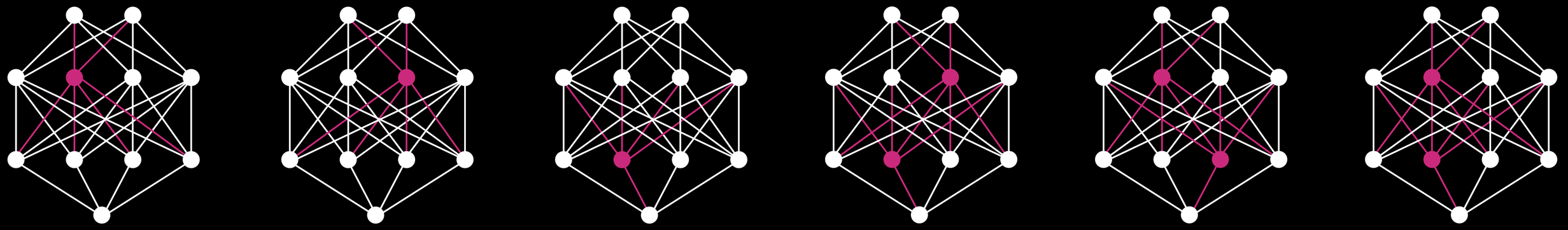
Deep Ensembles



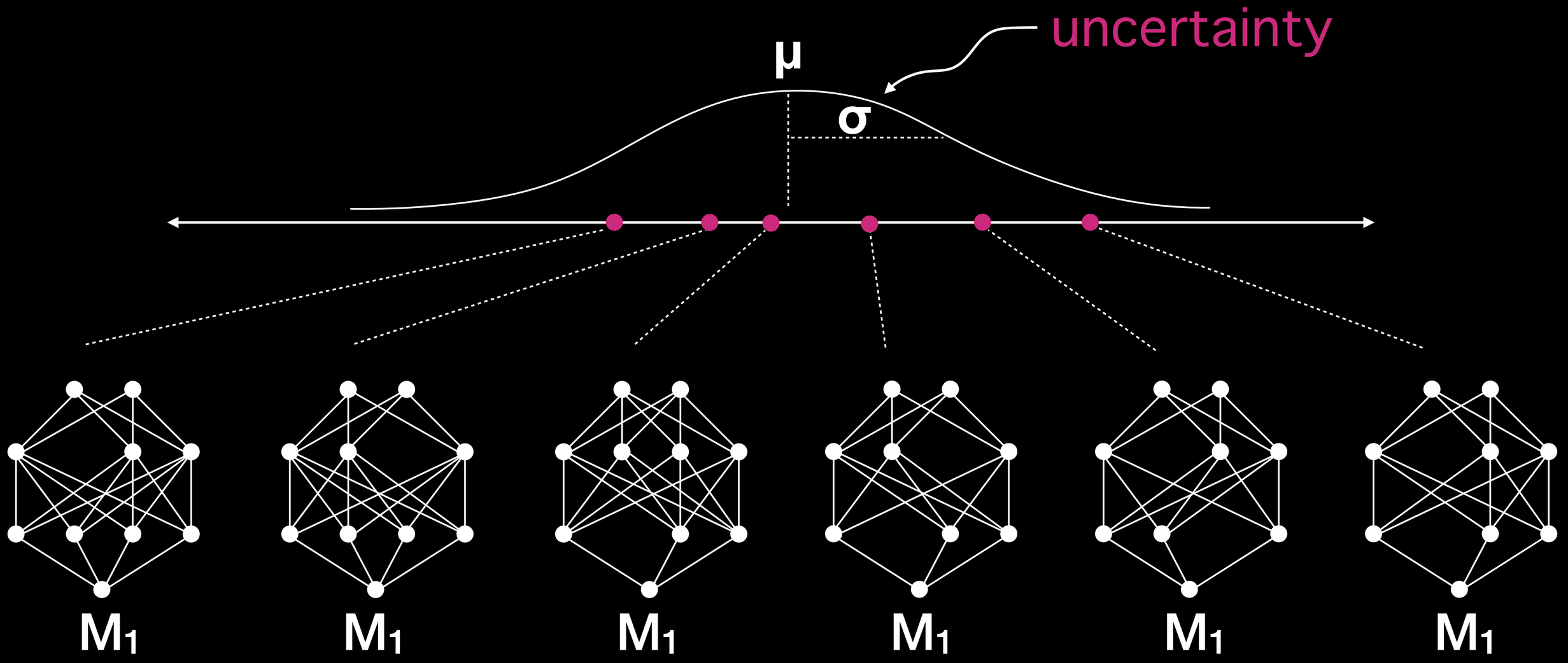
Monte-Carlo Dropout



Monte-Carlo Dropout



Monte-Carlo Dropout

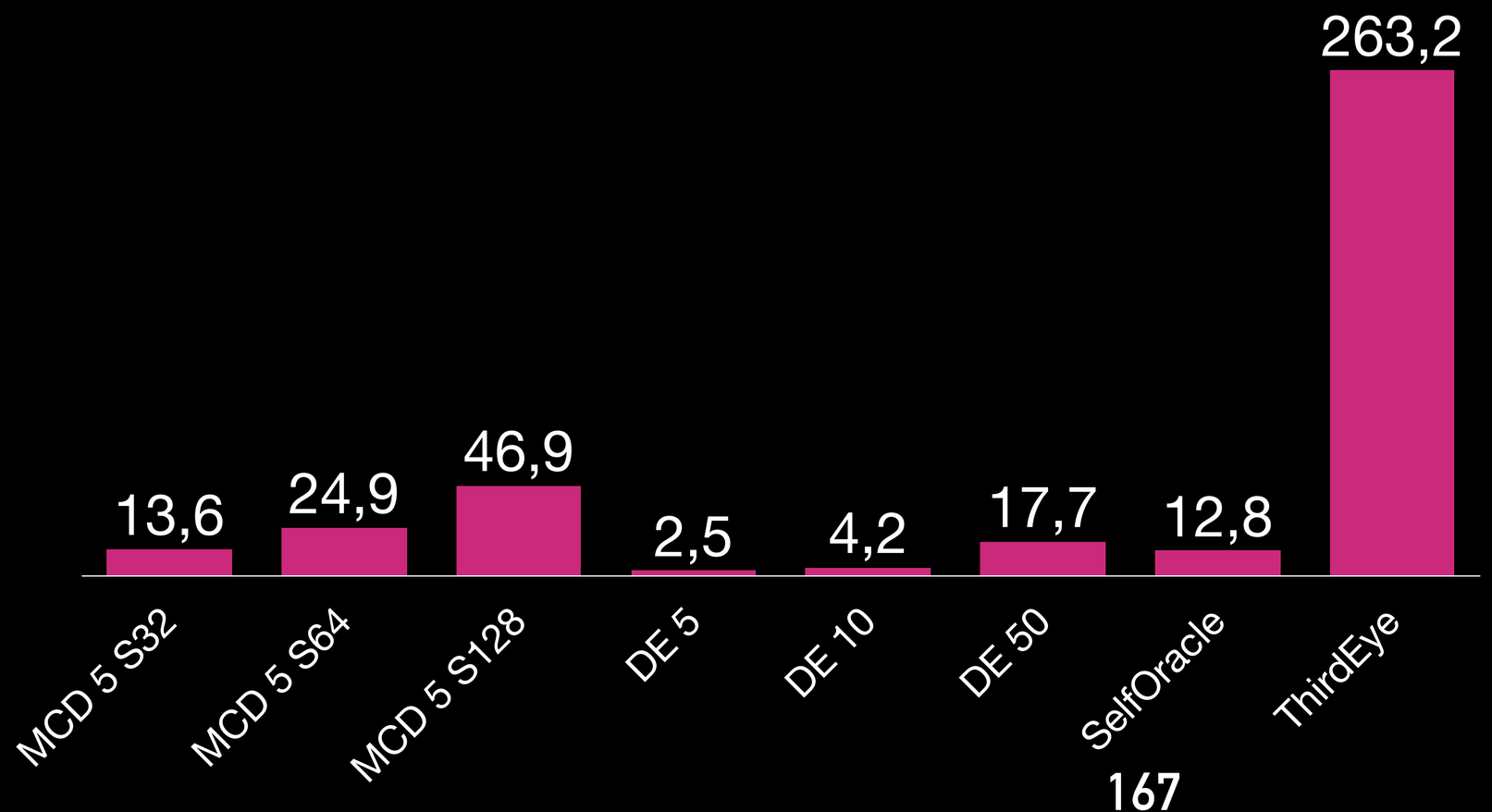


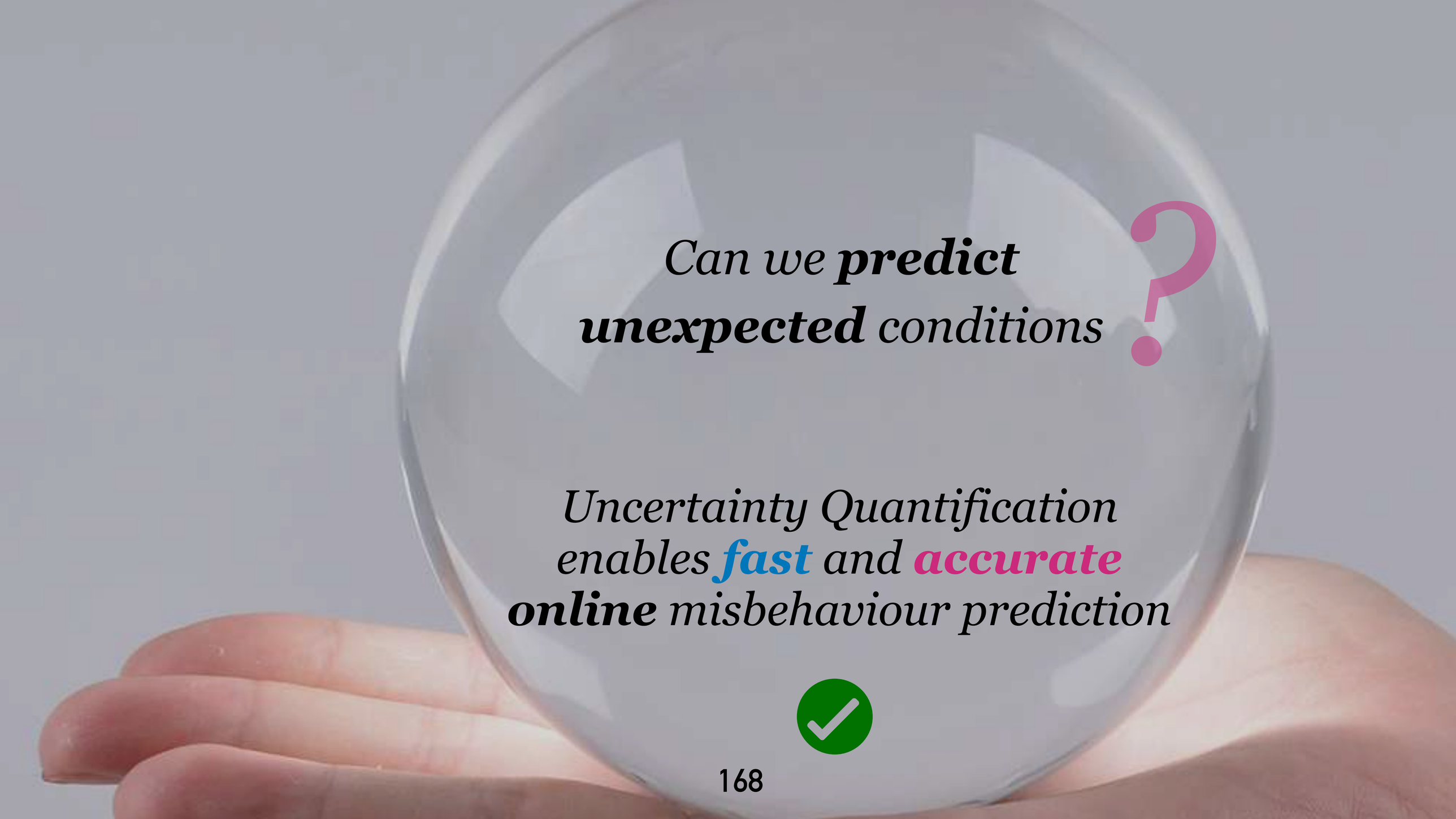
Is our approach effective in predicting misbehaviours ?

MCD 5 S32:
dropout rate = 5%
S32 = 32 samples

DE 5:
five models

Is our approach efficient in predicting misbehaviours ?




*Can we **predict** **unexpected** conditions*
?
*Uncertainty Quantification enables **fast** and **accurate** **online** misbehaviour prediction*

How to design them in production?

How to design them in production?

Steps ahead

Testing and Evaluation of Autonomous Driving Systems

From Simulated to Real-world Test Environments

Andrea Stocco

SIESTA summer school

02.09.2026

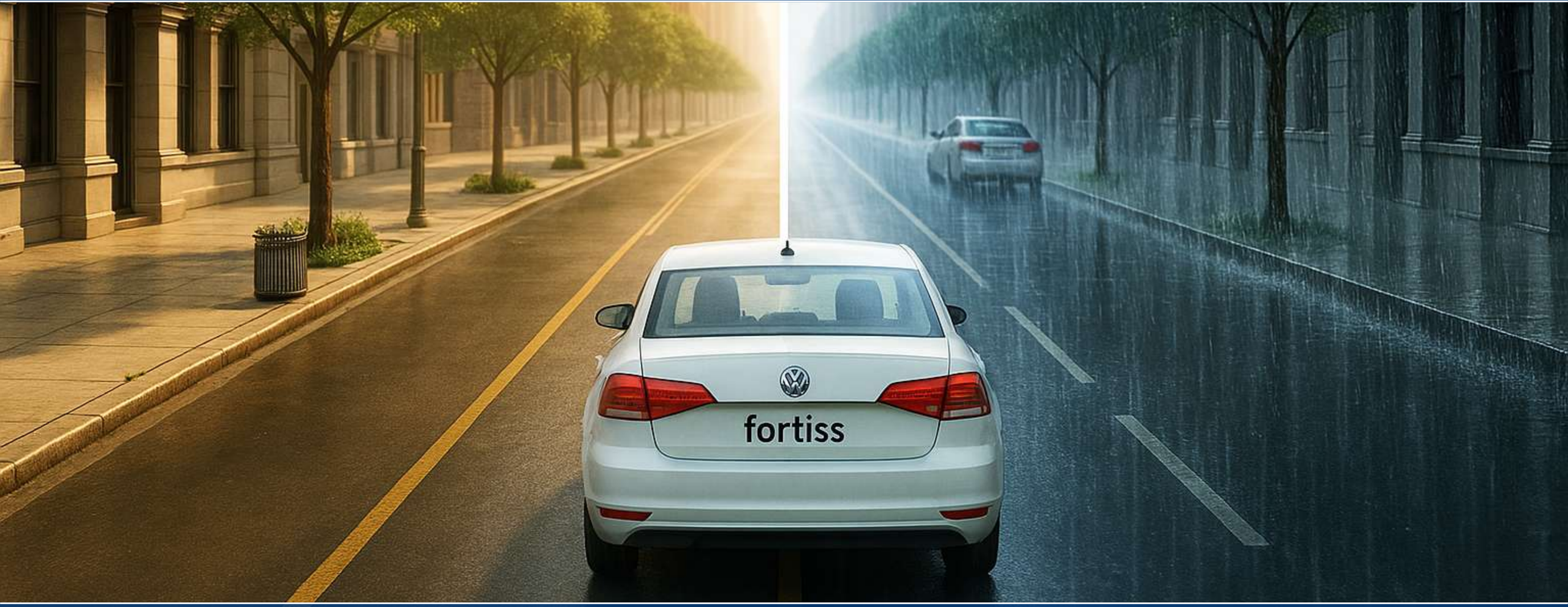


Image generated with DALL·E 3