Developing Multi-Fidelity Test Plans for Evolving and Heterogeneous AI-Enabled Systems

DATAWorks 2026

Tuesday, April 21, 2026

Zichong Yang
Ziran Wang
Jitesh H. Panchal (presenter)

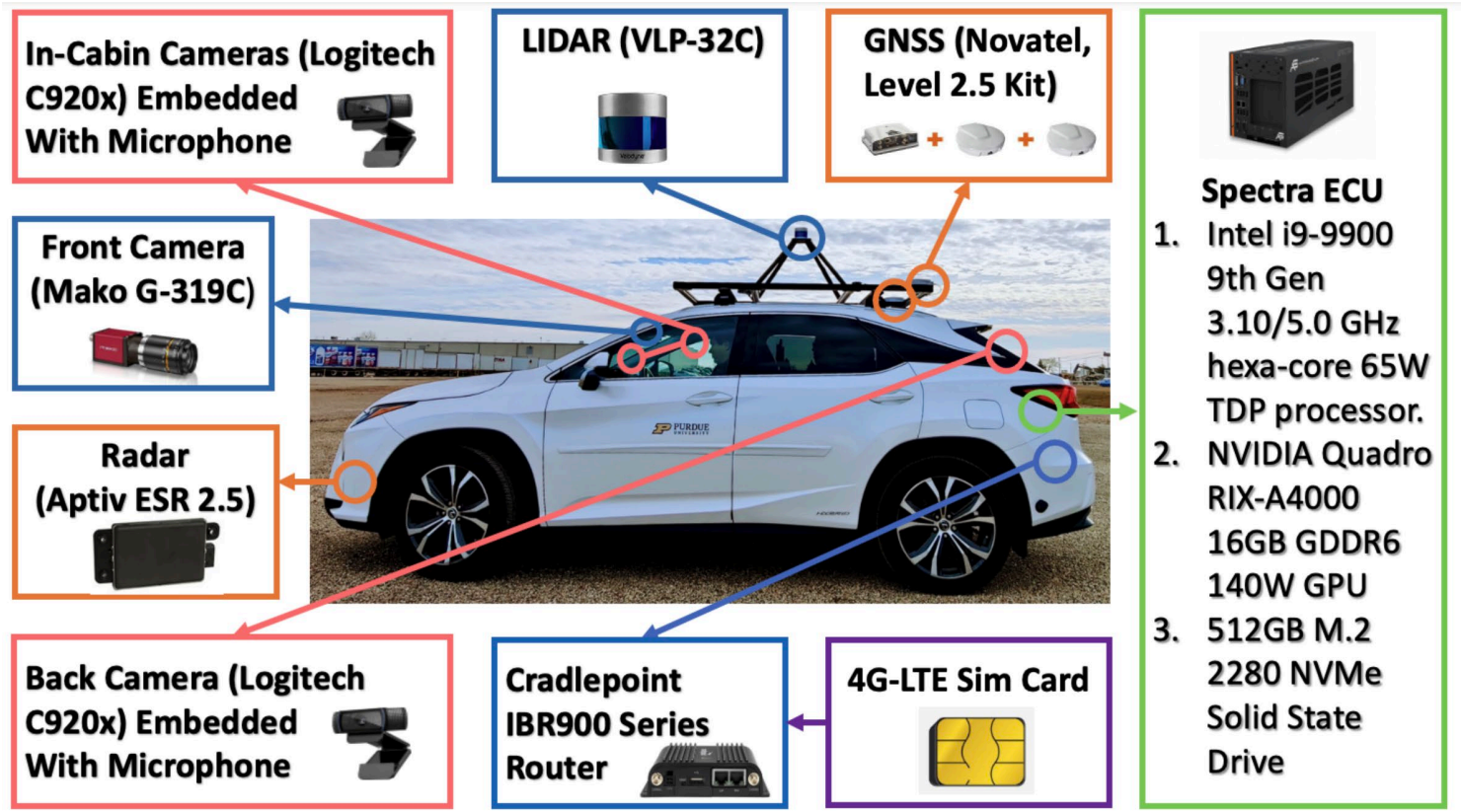
Purdue University, West Lafayette, Indiana

Motivational Context – Autonomous Vehicles

Autonomous vehicles serve as a representative autonomous AI-enabled system, integrating sensors, computing, and actuators.

Hardware Composition

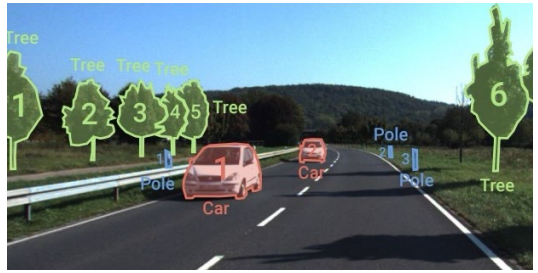
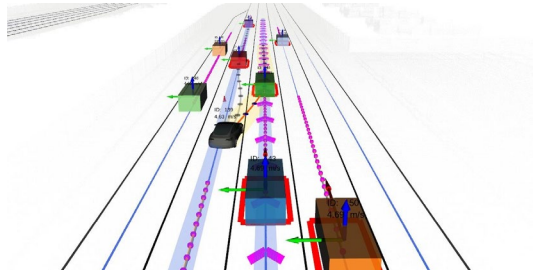
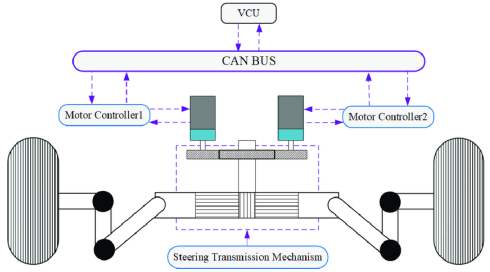
- Sensors
- Communication
- Computing
- Actuators



Courtesy: Prof. Ziran Wang (Purdue)

Autonomous Driving Software

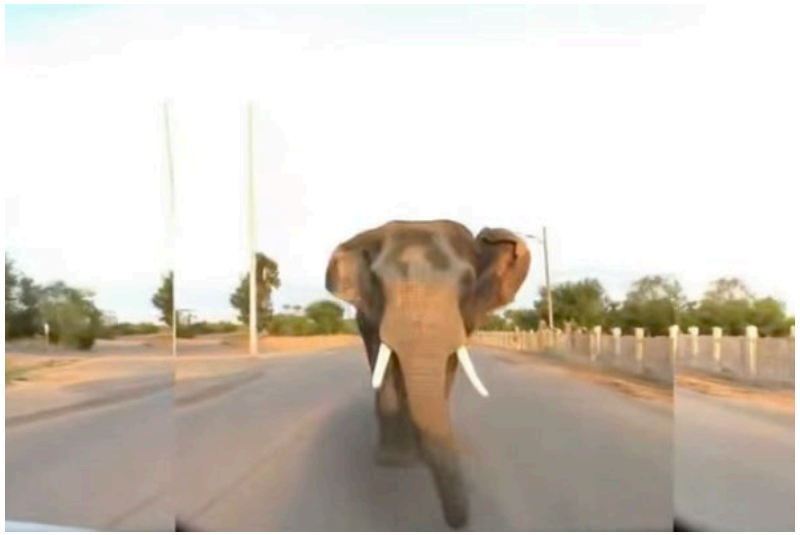
- Classical autonomous driving systems utilize modularized AI sub-systems.
- Each sub-system requires their own data for training and testing.

Perception		**Segmentation**	**Object Detection**	**Object Tracking**	**Localization**
Planning		**Route Planning**	**Behavior Planning**	**Motion Planning**	**Prediction**
Control		**Steering Control**	**Torque Control**	**Emission Control**	**Energy Management**

Challenge 1: Complex, High-Dimensional Testing Space

False negative cases

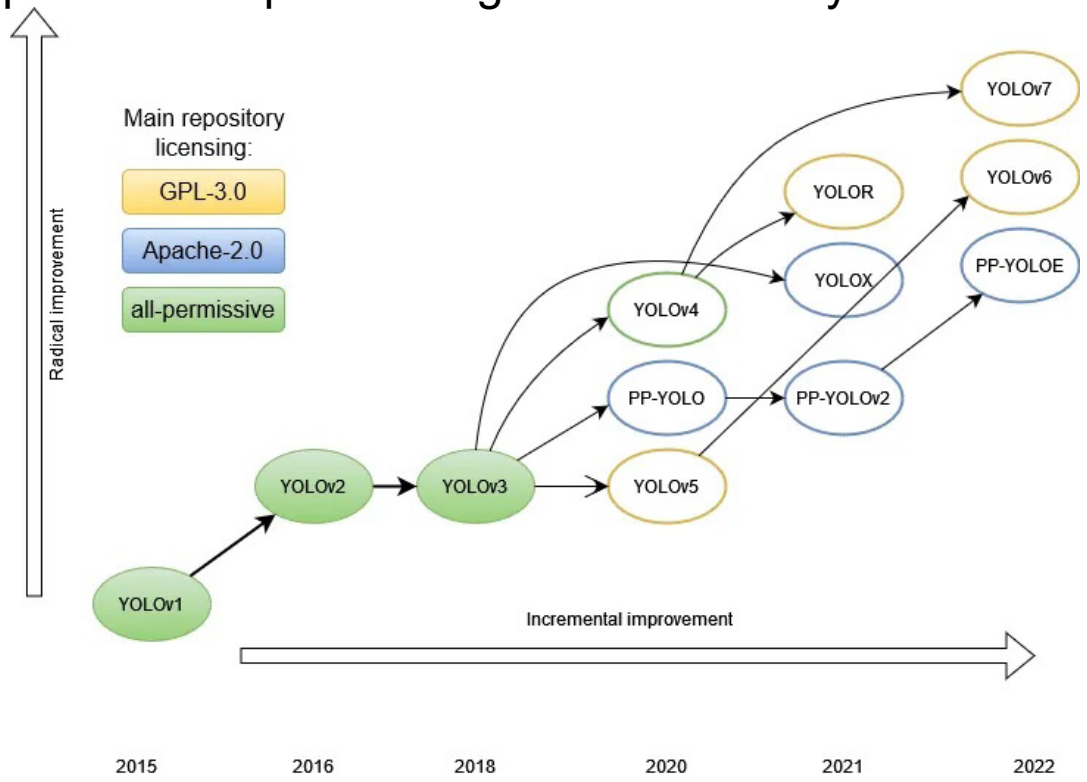
Rare scenarios



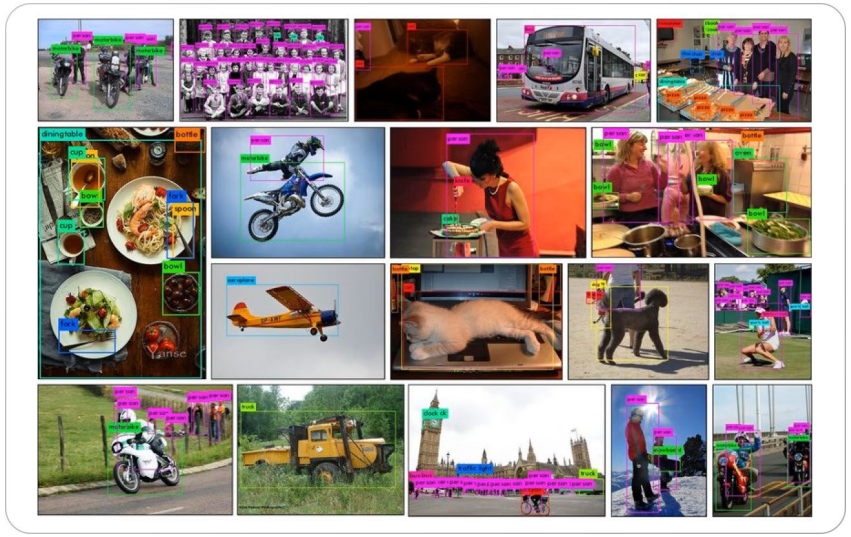
Out of distribution (OOD) cases

Challenge 2: Rapidly Evolving Models & Training Data

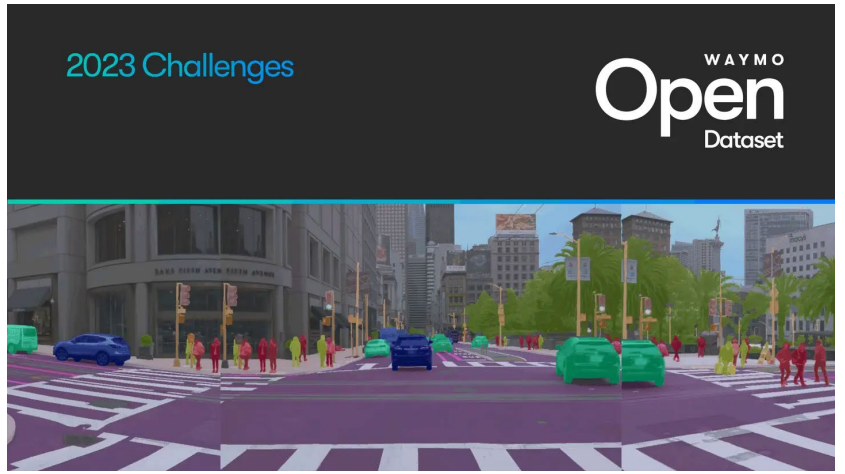
Newer variants with more advanced network structure and training approaches improve the general efficiency and accuracy.



https://medium.com/deelvin-machine-learning/the-evolution-of-the-yolo-neural-networks-family-from-v1-to-v7-48dd98702a3d



Generic Training Datasets

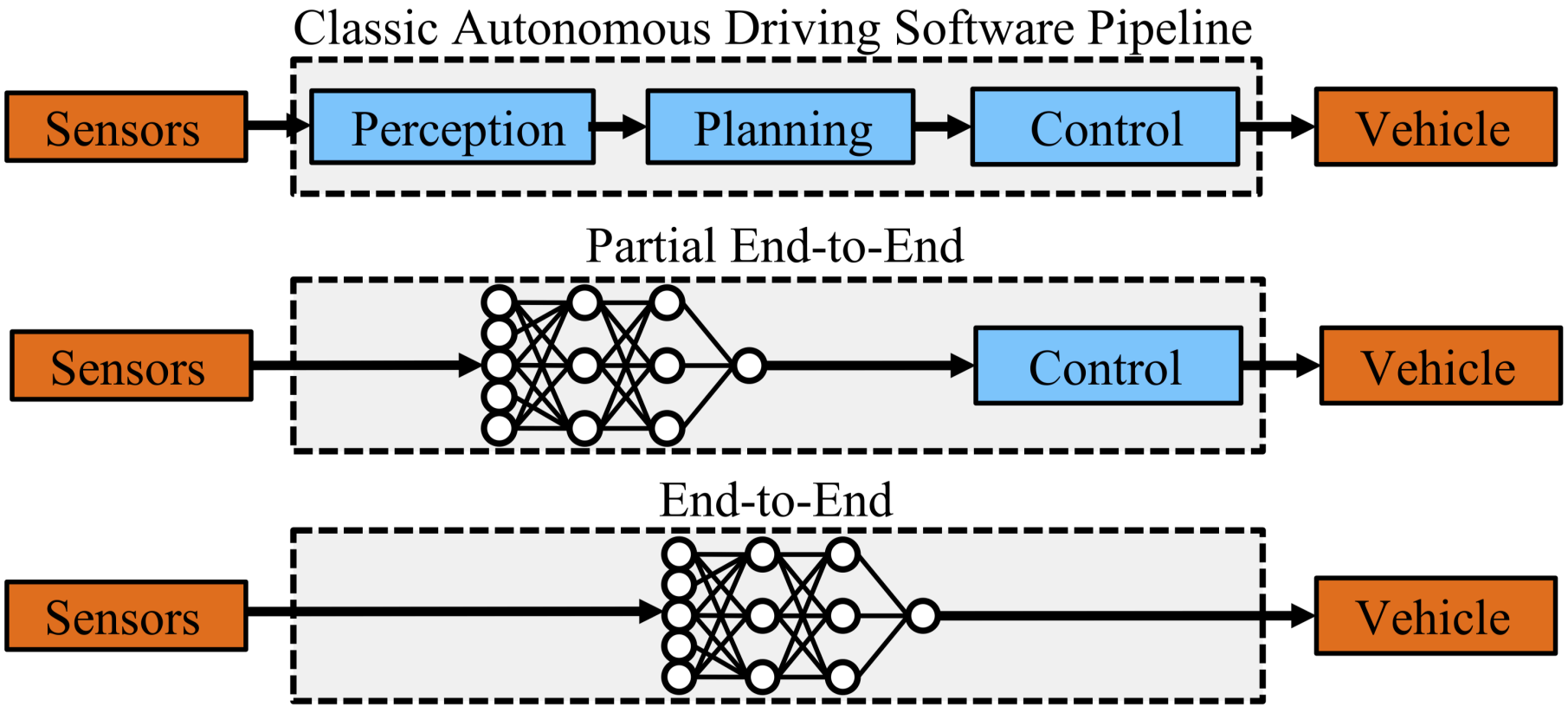


Application-Specific Fine-Tuning

Fine-tuning existing models increases the performance of the system on specific tasks.

Evolution to an End-to-End Approach

- End-to-End autonomous driving systems use only one AI system.
- Single dataset type to train the system, reducing intermediate annotation.



Challenge 3: High Cost of Real-World Testing

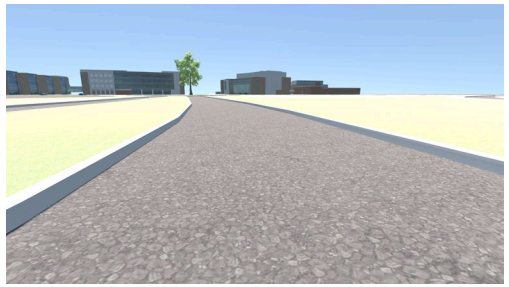
- **Fixed costs:** One-time expenses required to establish the testing infrastructure.
- **Evolving costs:** Recurring variable expenses incurred during the testing process.

	Real-world Testing
Equipment	Physical Vehicle + Modifications (More than $200,000)
Operational Environment	Closed and Open Road
Human Labor	Inspector All the Time
Weather effects	Snow, clouds, sunny, day, night
Time	Align with Physical Time Span

It is infeasible to adequately test AI in operation.

Opportunity: Multiple Fidelities of Testing Environments

Increasing Fidelity



General purpose 3D rendering software

Specialized driving T&E software (CARLA)

Generative AI enabled scenario generation

Real-world testing

Scene Creation	Manual construction	Parameter adjustments	Natural Language	Physical environment
Scene Diversity	Low	Moderate	High	Cost related
Computing	Low	High	Very High	Low
Visual Realism	Basic 3D graphics	Domain-specific realism	Close to reality	Actual reality

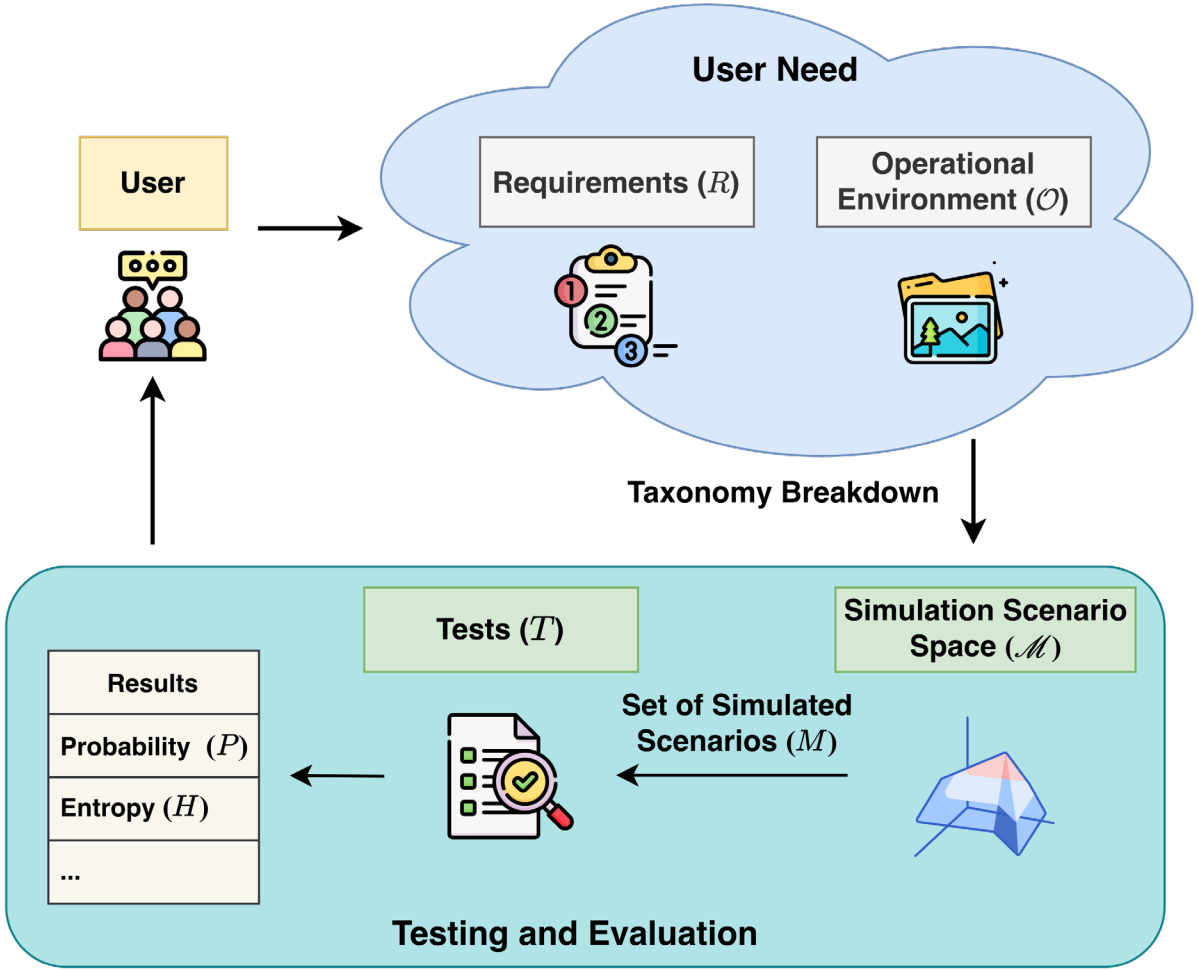
Proposed Multi-Fidelity T&E Approach

Acquisition Side

- Define Requirements
- Specify Operational Environment

Testing and Evaluation Side

- Simulation Space Definition through Taxonomy Breakdown
- Sequential Test Selection through Bayesian Optimization
- Confidence on Testing Results update through Bayes' Rule and Entropy Theory



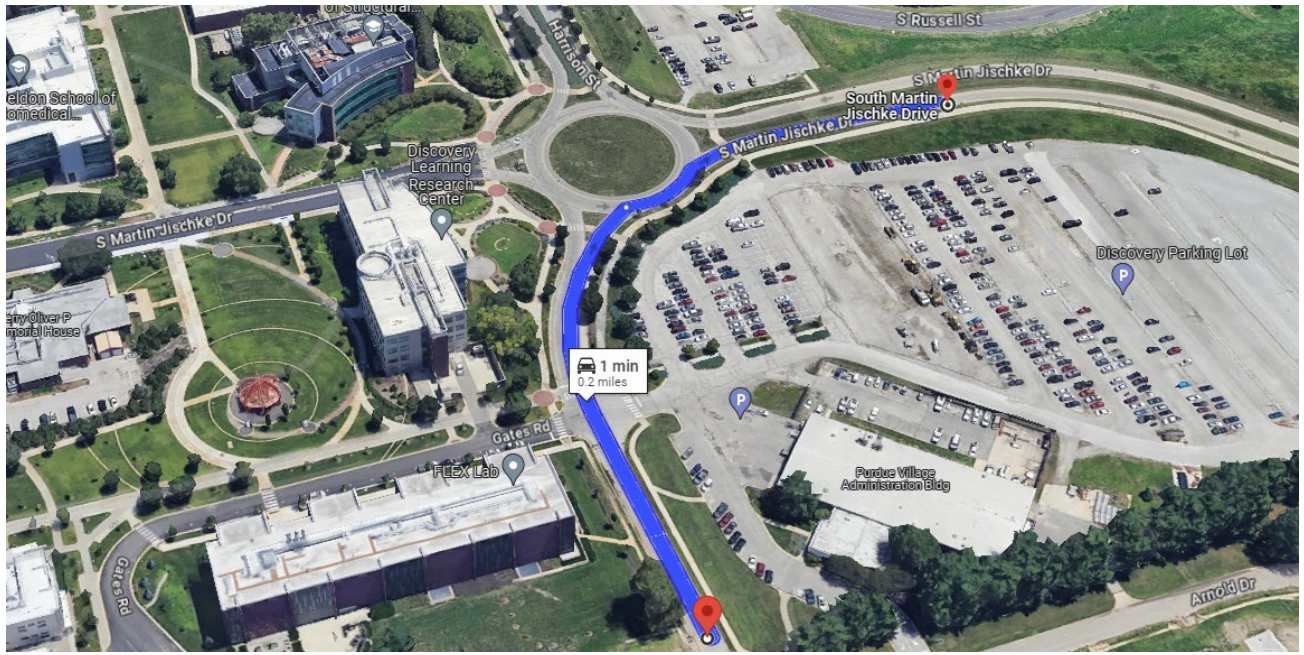
Example: Testing of the Vision System

- **AI model:** detection system on an autonomous vehicle
- **Input:** Video file from camera
- **Output:** Object location in each frame

https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.alten.com%2Fnext-generation-camera-based-adas-development

Requirements

- **System tested:** YOLOv8 (object detection)
- **Setting:** Autonomous Vehicle driving through a roundabout
- **Simulation:** Unity simulation engine and occlusion post-processing

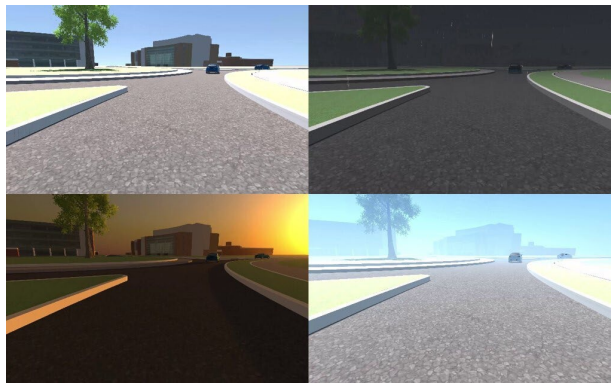


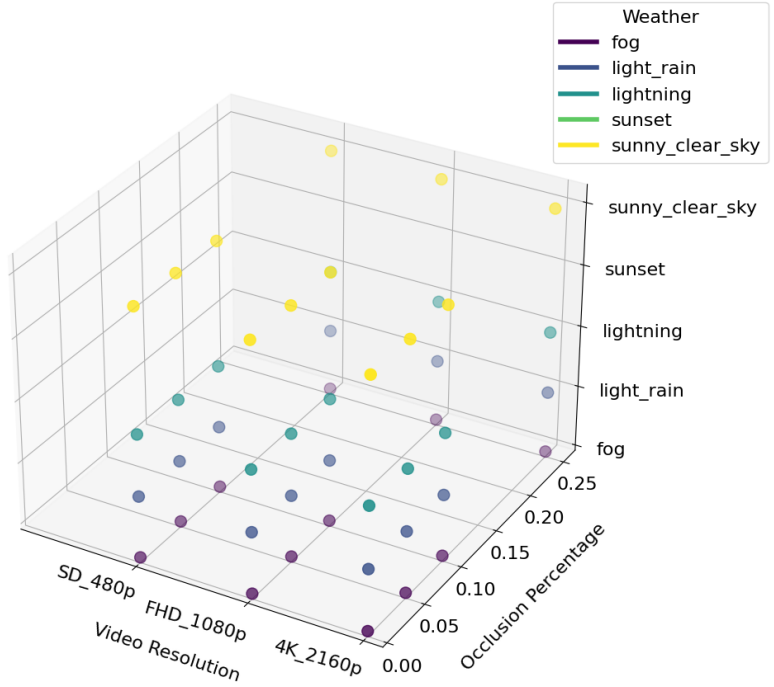
R	Requirements
r_1	System must operate on a 1080p HD video format
r_2	System must operate in direct overhead sunlight
r_3	System must operate in raining weather condition
r_4	System must operate under 25% random occlusion of vision field
r_5	System must operate with a 30 fps operational speed
r_6	System must have average confidence over 80% in vehicle prediction
r_i	…

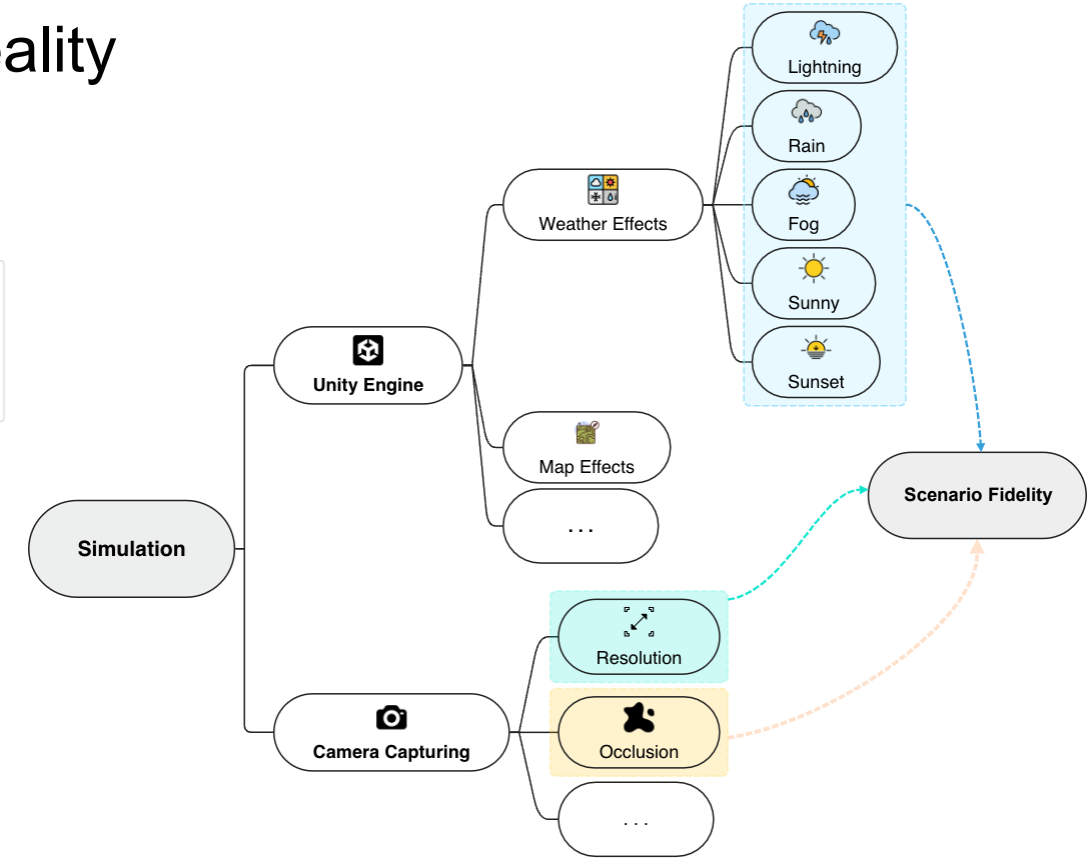
Part 1: Simple Scenario Space

Constructing the Simulation Scenario Space

- ***Single Scenario Instance (m_j):*** A generated scenario
- ***Scenario Fidelity:*** accuracy in representing reality
- ***Cost:*** $c = c_p \times c_l \times c_o$

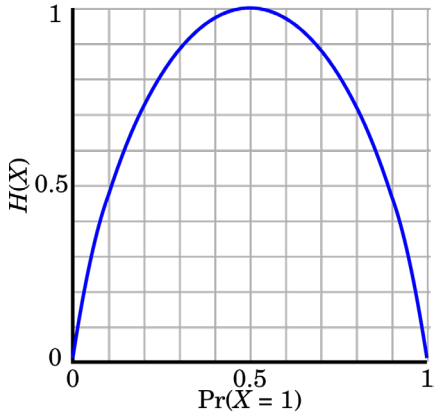






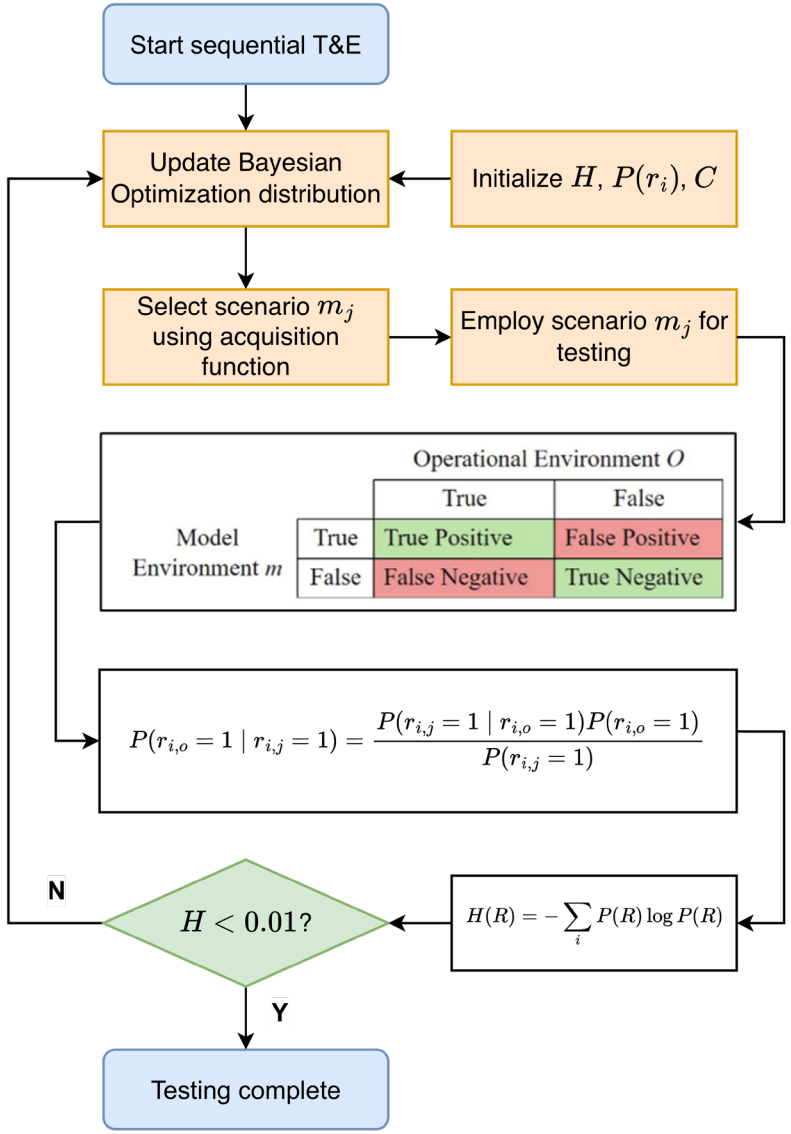
Sequential Test Design

- **Random variables:** $r_{i,j}$, $r_{i,o}$ (probability of requirement r_i being satisfied)
- **Objectives of the T&E process:**
  - minimize uncertainty about whether the system satisfies the requirements in the operational environment
  - Minimize the cost of T&E
- **Model selection:** Acquisition Function
- **Belief update:** Bayes' Rule



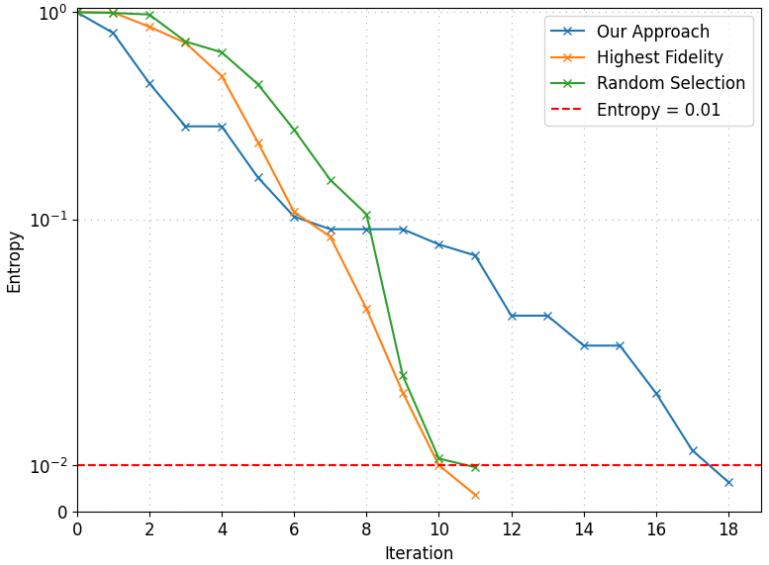
Uncertainty measure: Shannon Entropy

$$H(r_i) = -P(r_i)log_2P(r_i) - (1 - P(r_i))log_2(1 - P(r_i))$$

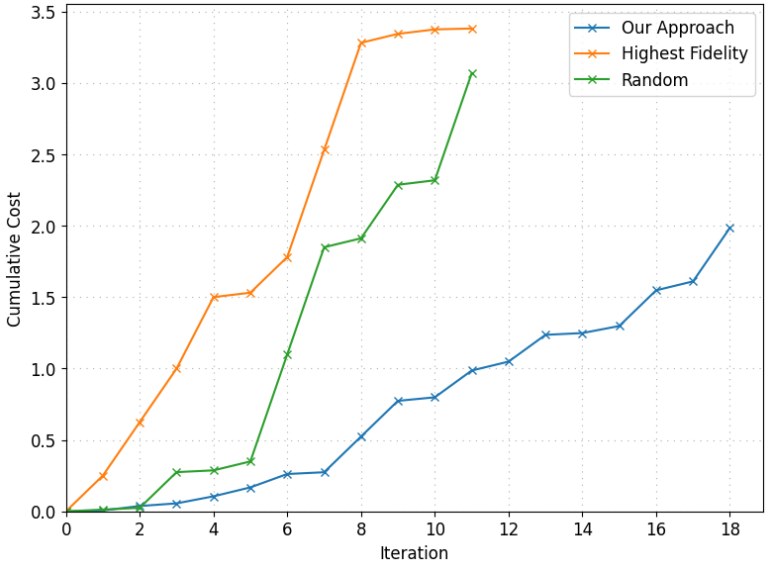


Results

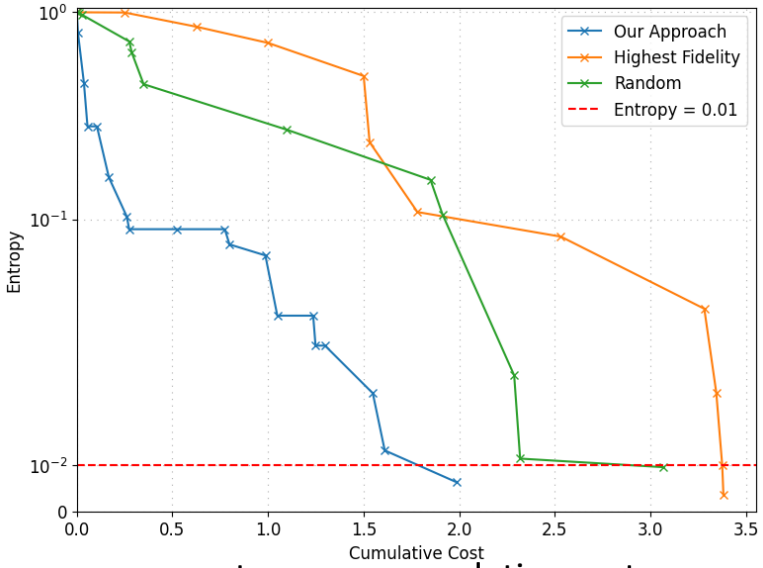
- Stop condition: entropy $H < 0.01$.
- Entropy and accumulated costs are recorded.
- Our approach: **41% reduced cost**, broader exploration of model space



entropy vs. number of tests



cumulative cost vs. number of tests



entropy vs. cumulative cost

Conclusion: The proposed method can ***find the testing sequence with maximum utility*** by experimenting with testing the perception system of an autonomous vehicle.

Part 2: Risk-based Analysis with Evolving AI model

Risk-Based Testing

- Risk = probability of failure x severity (impact) of consequence
- Model selection is driven by the objective to maximize the knowledge of high-risk failure modes.

RISK ASSESSMENT MATRIX				
SEVERITY / PROBABILITY	Catastrophic (1)	Critical (2)	Marginal (3)	Negligible (4)
Frequent (A)	High	High	Serious	Medium
Probable (B)	High	High	Serious	Medium
Occasional (C)	High	Serious	Medium	Low
Remote (D)	Serious	Medium	Medium	Low
Improbable (E)	Medium	Medium	Medium	Low
Eliminated (F)	Eliminated			

Prioritization of safety-critical requirements.

Amland, Ståle, and Hulda Garborgsv. "Risk based testing and metrics." 5th International Conference EuroSTAR. 1999.
Amland, Ståle. "Risk-based testing:: Risk analysis fundamentals and metrics for software testing including a financial application case study." Journal of Systems and Software. 2000.

Requirements, Failure Modes and Risks

Requirements $R = \{r_1, r_2, \dots, r_i\}$ ← Manual definition from UL4600* standard.

Failure Modes $F_i = \{f_{i,1}, \dots, f_{i,j}\}$ ← Probability that the failure mode will happen $P_f(\text{prediction}, \text{target})$

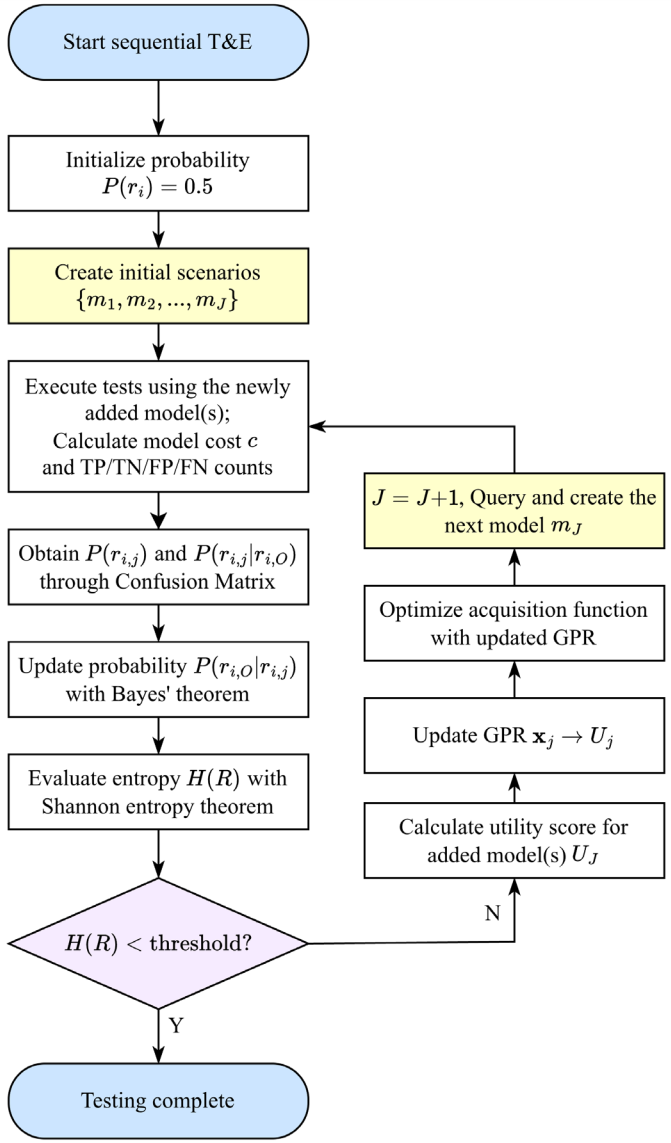
Risk $Re_f = C_f P_f$ ← C_f describes the level of impact

Requirements	Failure Modes	Impact / Consequence
High accuracy in detecting pedestrians	False Positive	Low (blocking traffic)
	False Negative	High (lethal)
	Misclassification	Medium (unstable prediction)
Low latency in detecting pedestrians	Medium Latency	Low (larger deceleration)
	Critical Latency	Medium (emergency break)
	High Latency	High (lethal)
…		

*https://users.ece.cmu.edu/~koopman/ul4600/index.html

Updated Testing on Single Version AI

- We expect to get from test a minimal confidence p that the probability a failure mode will not happen is q.
- Then, we have $N \geq \left\lceil z_p^2 \frac{\hat{\theta}(1-\hat{\theta})}{(\hat{\theta}-q)^2} \right\rceil$ through Binomial Proportion Confidence Interval which determines the minimal number of test cases required.
- The actual probability $\hat{\theta}$ updates after every model is simulated and tested, which updates N.



Example of Termination Criterion

- For example: we want to be at least p=99% sure that there will be no false positives of pedestrian prediction in q=86% of the time
- We first sample 5 simulation instances through Latin Hypercube Sampling
- After running the initial model, we have

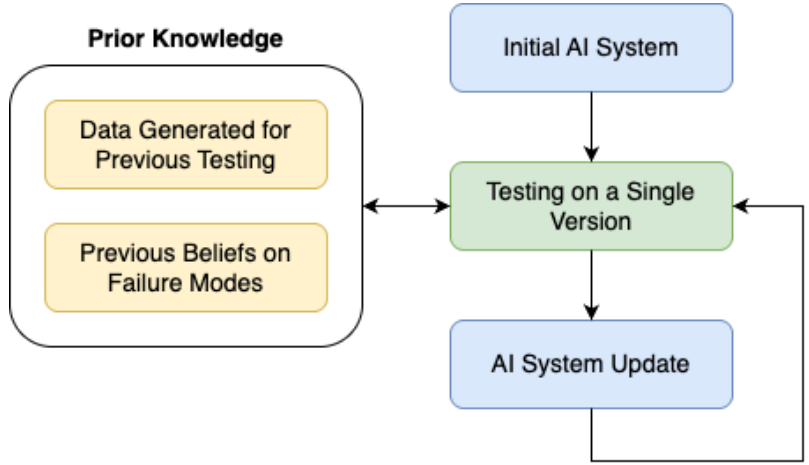
$$N \geq \left\lceil 1.645^2 \frac{\hat{\theta}(1-\hat{\theta})}{(\hat{\theta}-0.86)^2} \right\rceil = 46 \text{ models}$$

- But with iteration running, $\hat{\theta}$ increased. After 18 iterations, $N \geq 17$, equal to 17 models required. Thus, we can stop testing.

Step (k)	Quality	Resolution	Weather	$P(f_{i,j}, m_k)$	N_k
1	0.0	0.0	0.0	0.818	1.8
2	100.0	1080.0	1.0	0.861	15.4
3	50.0	540.0	2.0	0.864	328.7
4	70.0	720.0	3.0	0.868	221.6
5	90.0	960.0	4.0	0.868	46.5
6	0.1	887.1	1.9	0.853	54.3
7	91.8	960.2	3.9	0.868	29.8
8	99.7	343.3	3.0	0.866	22.4
9	0.5	394.8	4.4	0.856	23.9
10	1.8	646.3	4.0	0.856	24.9
11	99.2	627.8	4.8	0.870	19.3
12	99.2	456.5	1.8	0.863	17.8
13	99.9	804.4	4.7	0.869	15.2
14	1.1	1067.3	2.7	0.857	15.8
15	99.8	233.3	1.1	0.853	17.5
16	99.9	706.4	1.4	0.860	17.3
17	0.9	763.9	4.3	0.855	18.3
18	99.4	885.4	1.3	0.864	17.4
19	99.9	567.5	1.8	0.865	16.3
20	0.6	271.8	3.5	0.852	17.7

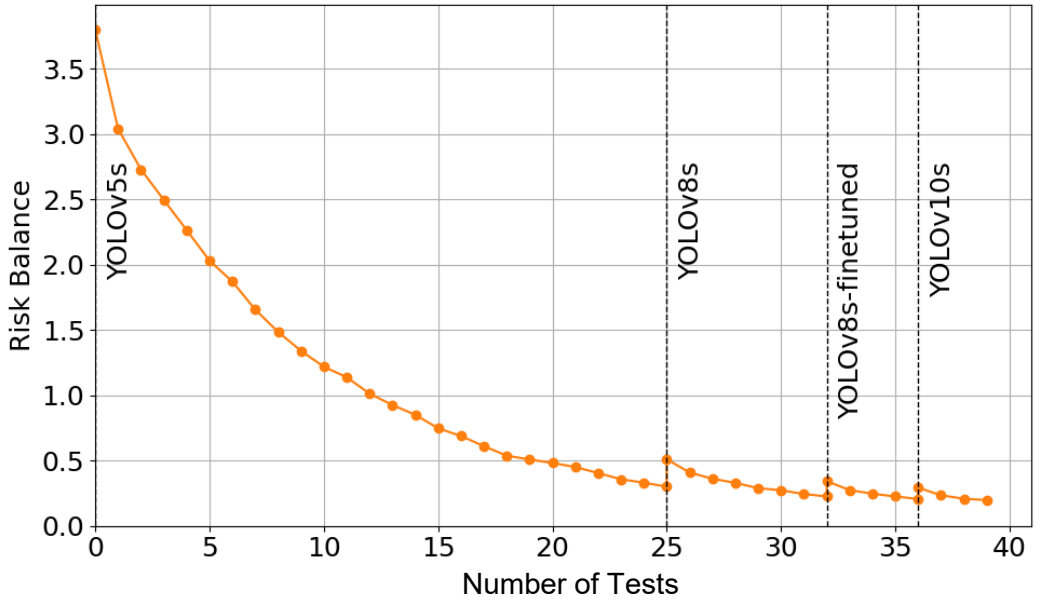
Testing After Version Update

- After testing on each version, we obtained information as prior knowledge for future tests.
- Previous data generated for testing, including images and ground truth, can be used to directly test the updated AI system.
- Previous beliefs on failure modes can help us prioritize tests with higher risks.
- Retaining previous information can reduce the overall cost of testing.



Results

- Applied 4 YOLO versions for testing: 5s, 8s, traffic fine-tuned 8s, 10s.
- 5 requirements with 13 related failure modes designed in reference to UL 4600.
- For each generation of the AI system, using more scenario instances reduces the overall risk.
- A spike following a version update indicates the presence of new, unknown information about the system.
- Our approach significantly reduces the number of tests required through utilizing prior information.



	Number of Tests (with prior information)	Number of Tests (without prior information)
YOLOv5s	25	25
YOLOv8s	7	27
Fine-tuned YOLOv8s	5	23
YOLOv10s	4	21

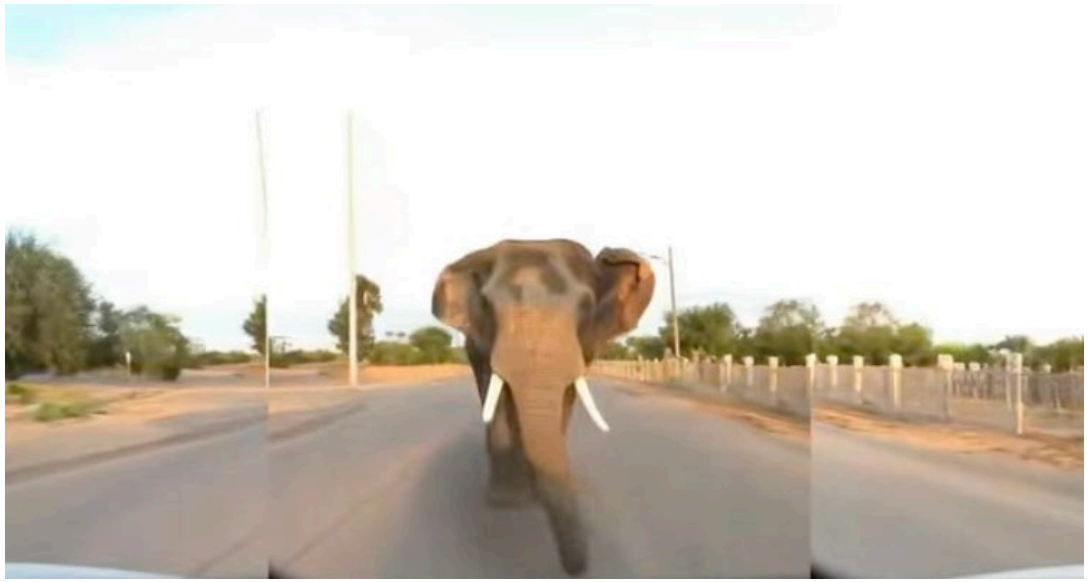
Part 3: Latent Space of Testing Scenarios

World Foundation Models for AV Testing

- High fidelity data generation tool: But what scenarios to generate?
- Prior methods require manual definition of failure modes.
- We want to automatically find out the seen failure modes, potential failures unseen in the training data.

Rare but seen: Pedestrians blocking traffic (captured from Nvidia's Physical AV dataset)

https://huggingface.co/datasets/nvidia/PhysicalAI-Autonomous-Vehicles



Rare and unseen: Encounter with an elephant (generated by Waymo's world foundation model)

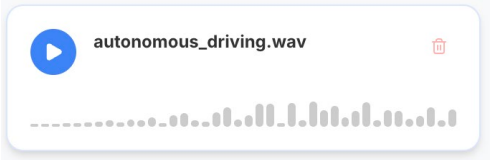
https://waymo.com/blog/2026/02/the-waymo-world-model-a-new-frontier-for-autonomous-driving-simulation/

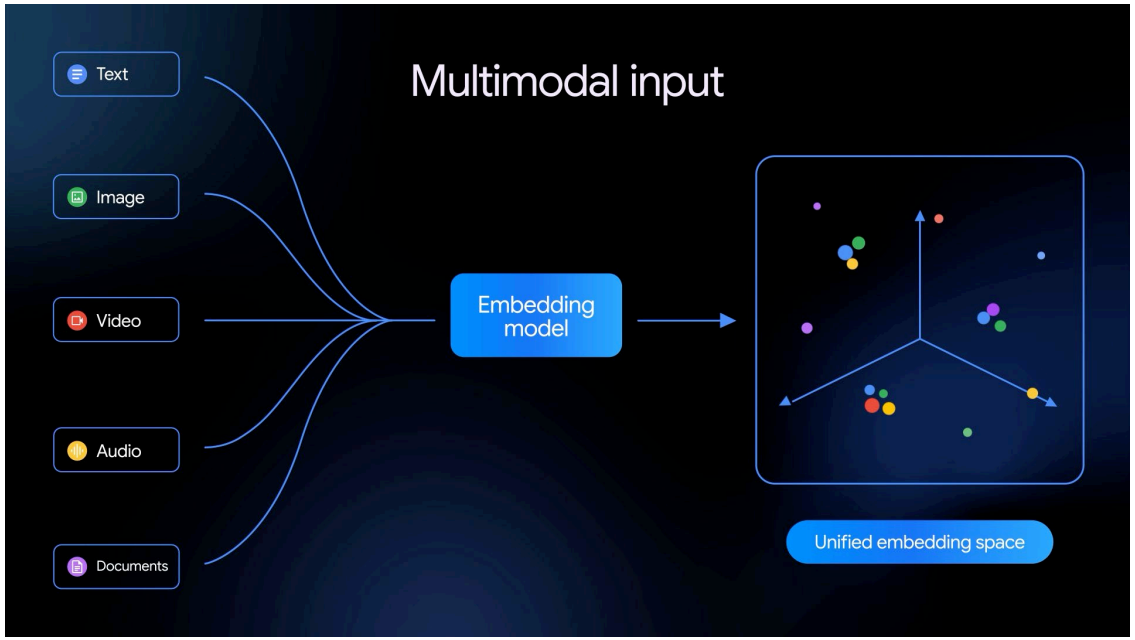
Embedding Space – A Learned Scenario Space

Instead of manual definition of available scenarios, we let the embedding model to learn a scenario space.

- **Multi-Modal Translation:** Converts diverse data into a unified vector space.
- **Contextual Organization:** Groups disparate data points by conceptual similarity.
- **Interoperable Search:** Allows querying across different media types.

"Autonomous Driving"

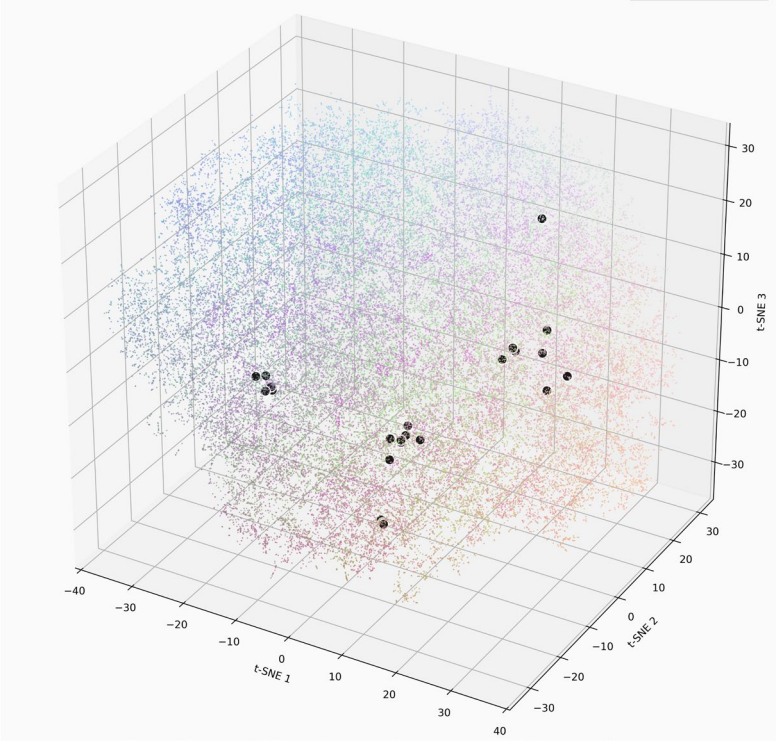




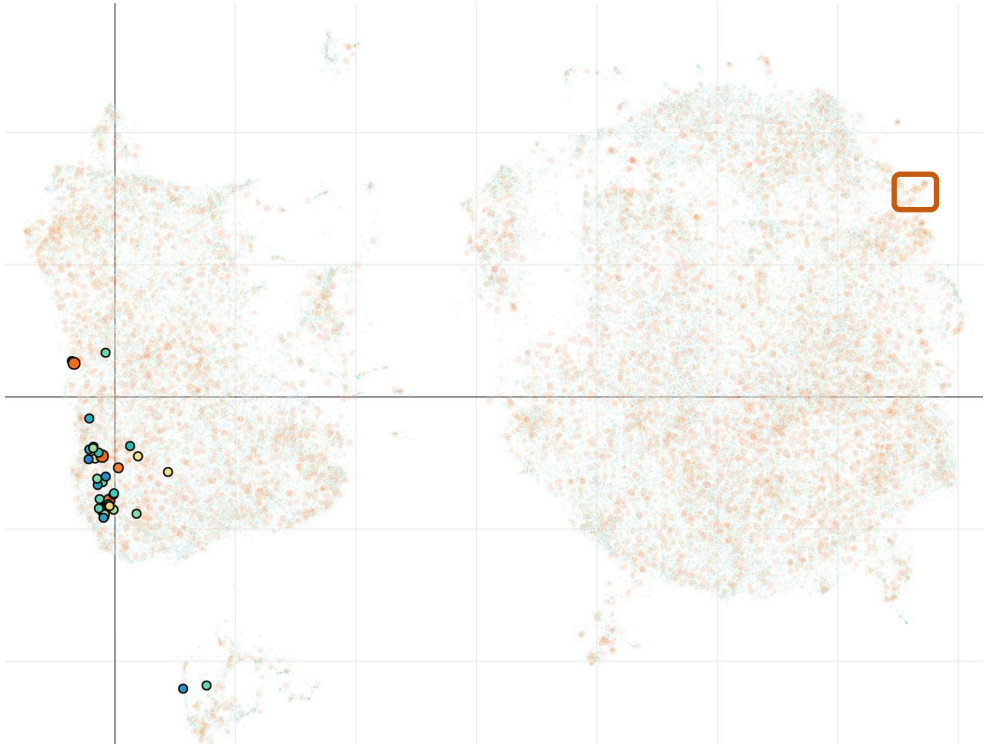
https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-embedding-2/

Scenario Space Encoding

- Through NVIDIA Cosmos Embed*, we map over 300,000 scenes into a scenario space.
- By looking into the scenario space, we find low score or low-density clusters.



Visualizing the first 3 of 768 dimensions of the new scenario space. ● is the original dataset generated through simulation.



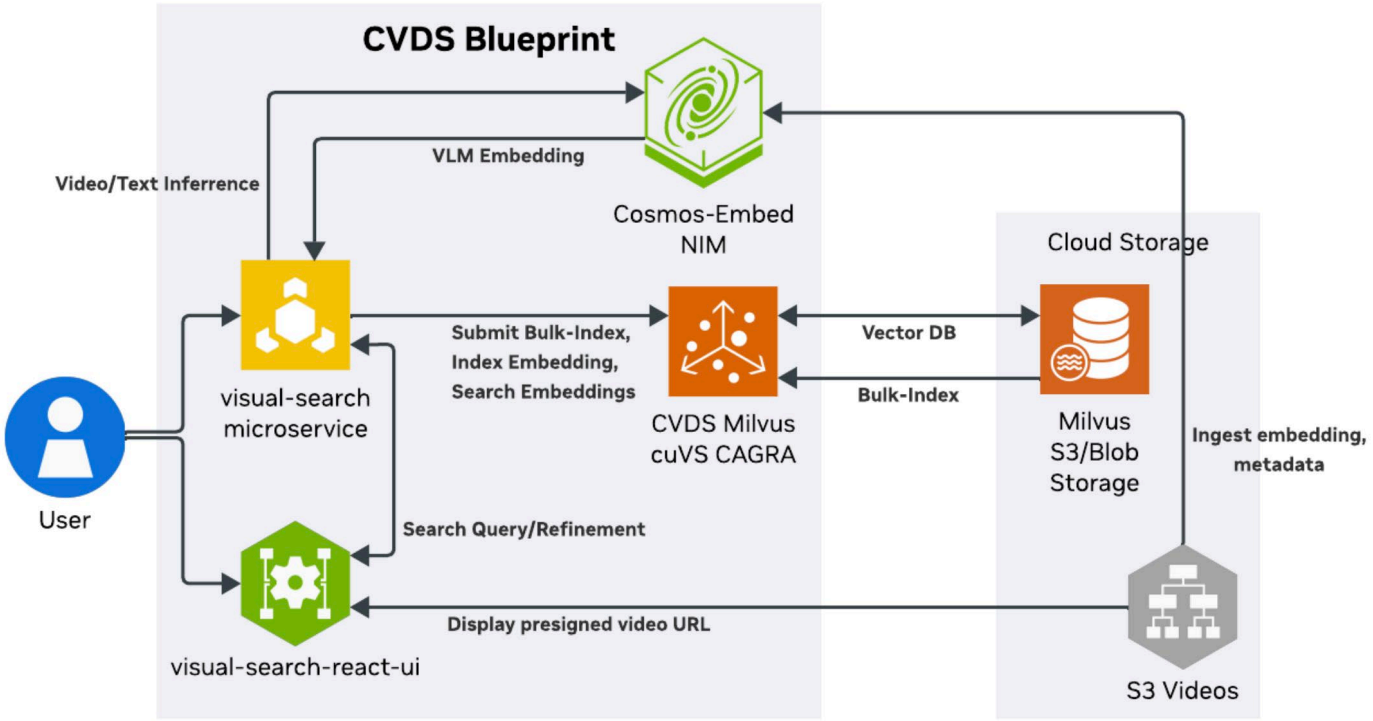
Highlighted points: Top 40 results for “A car pulling out of a driveway directly into the ego vehicle's path at night”

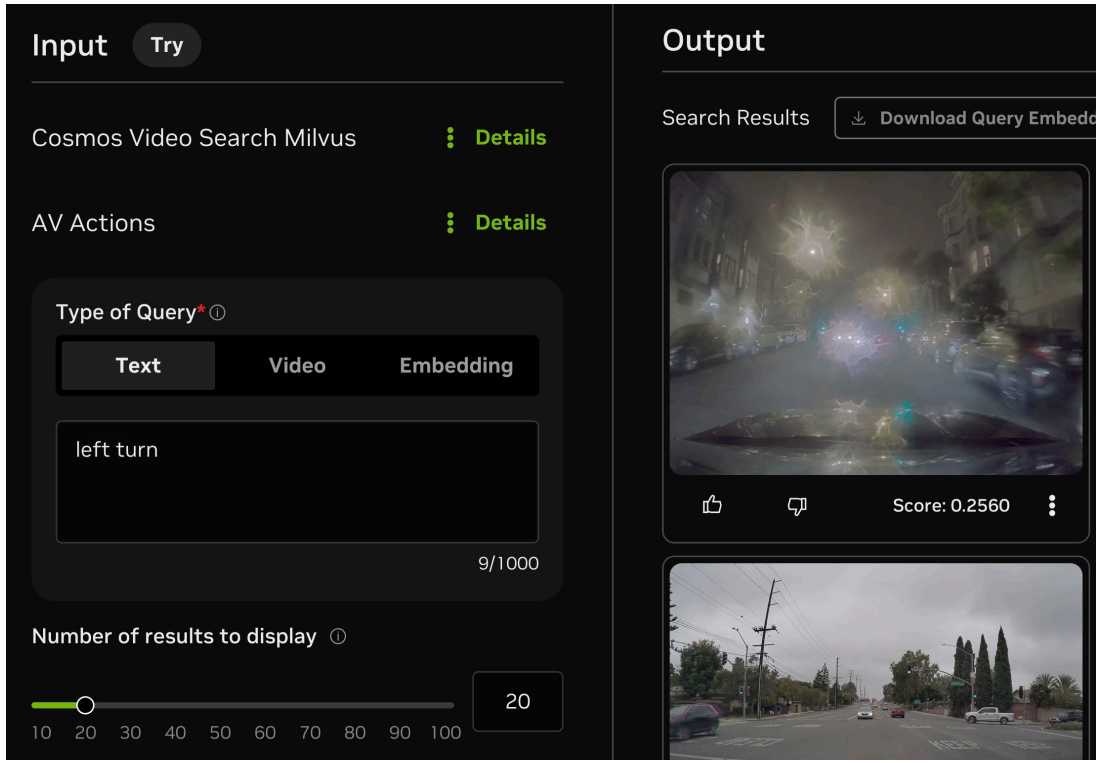
▢: Possible low score cluster

*https://huggingface.co/nvidia/Cosmos-Embed1-448p

Searching the Embedding Space

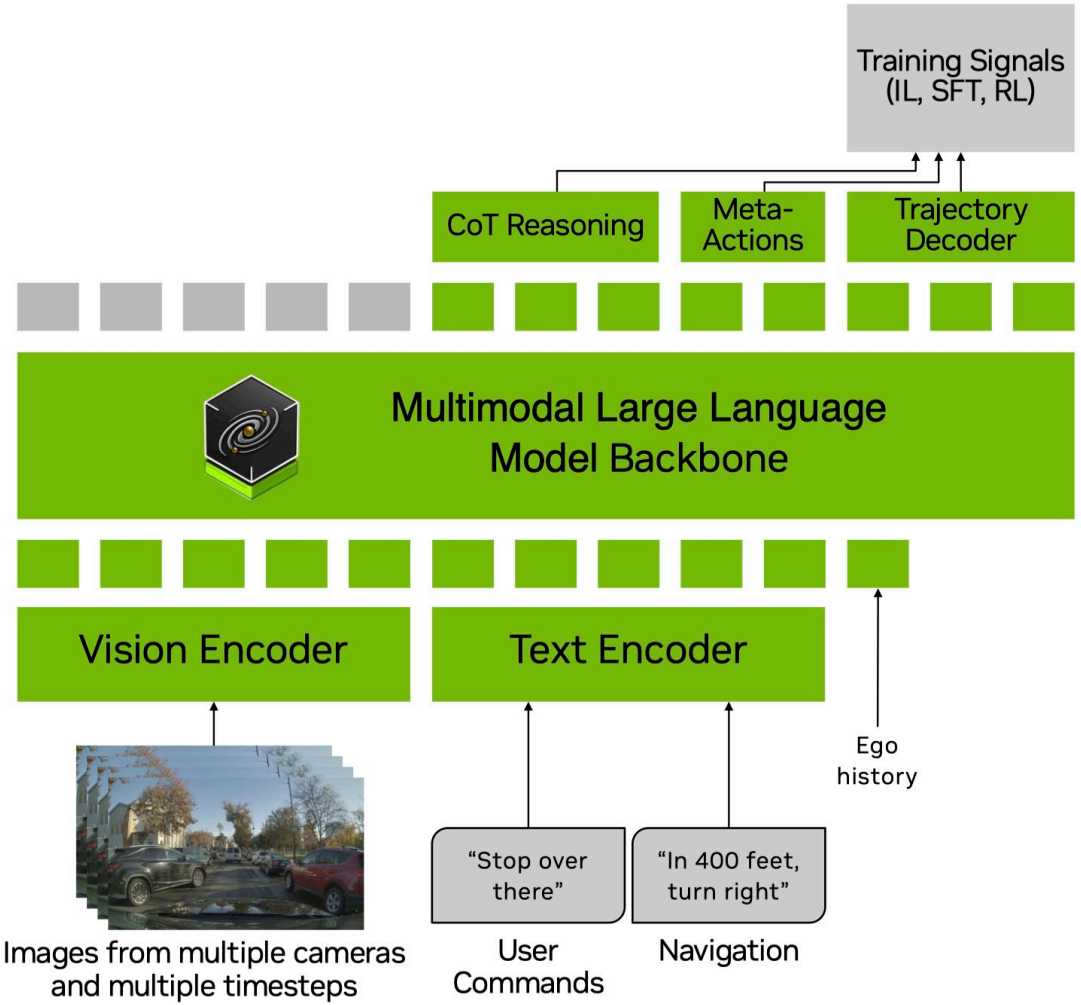
- Cosmos Dataset Search: Nvidia's cloud-based solution.
- Utilize cosmos search to create embedding.

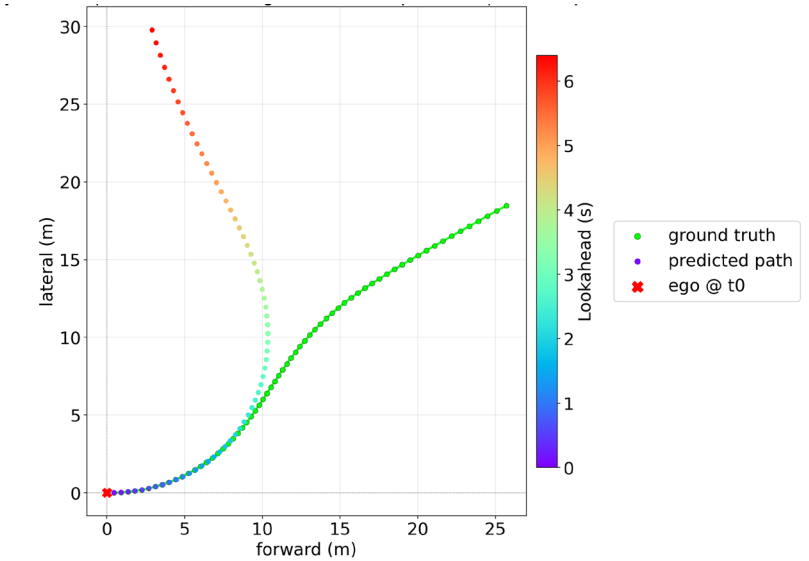




https://build.nvidia.com/nvidia/cosmos-dataset-search

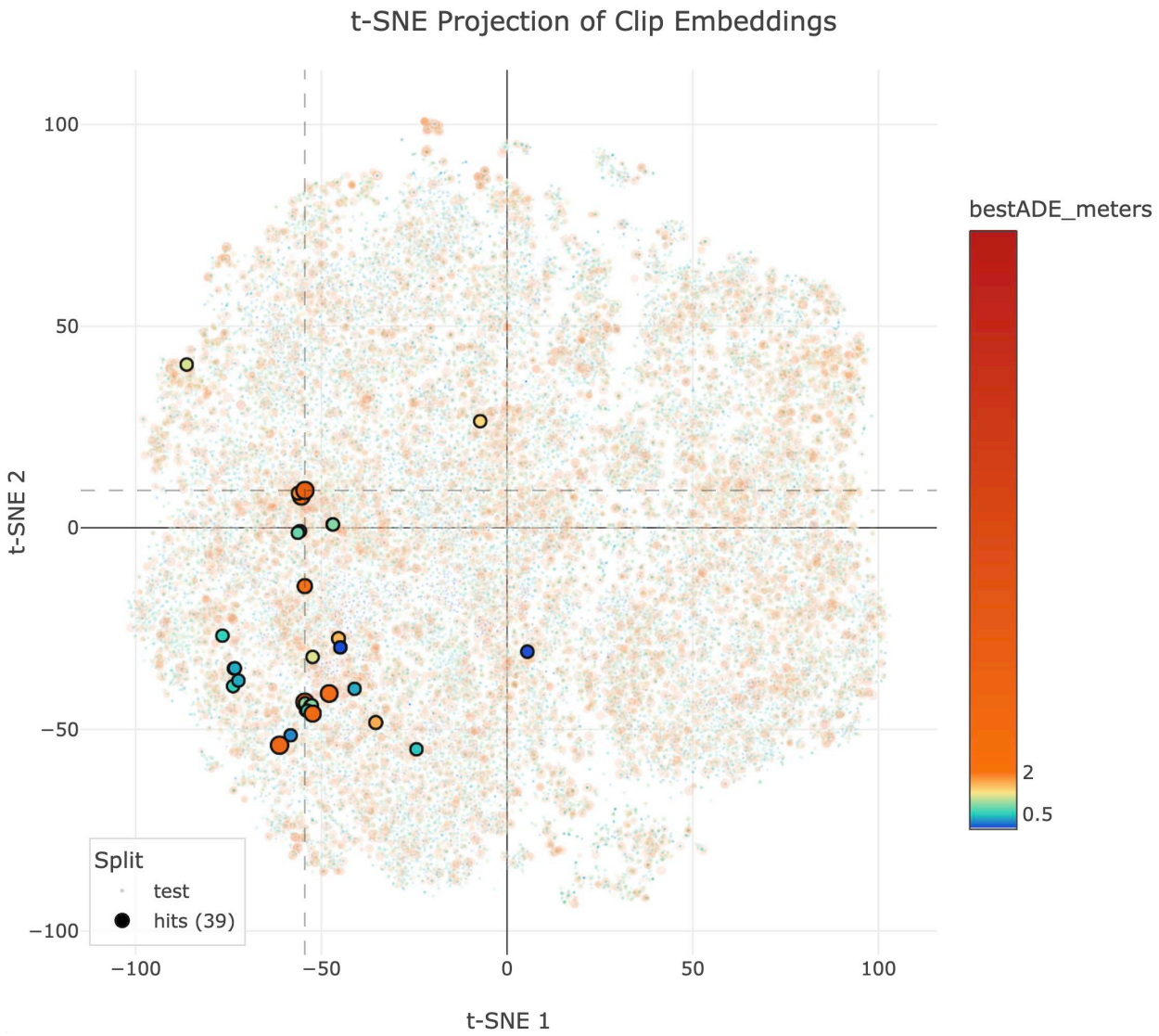
Open Loop Testing using NVIDIA Alpamayo

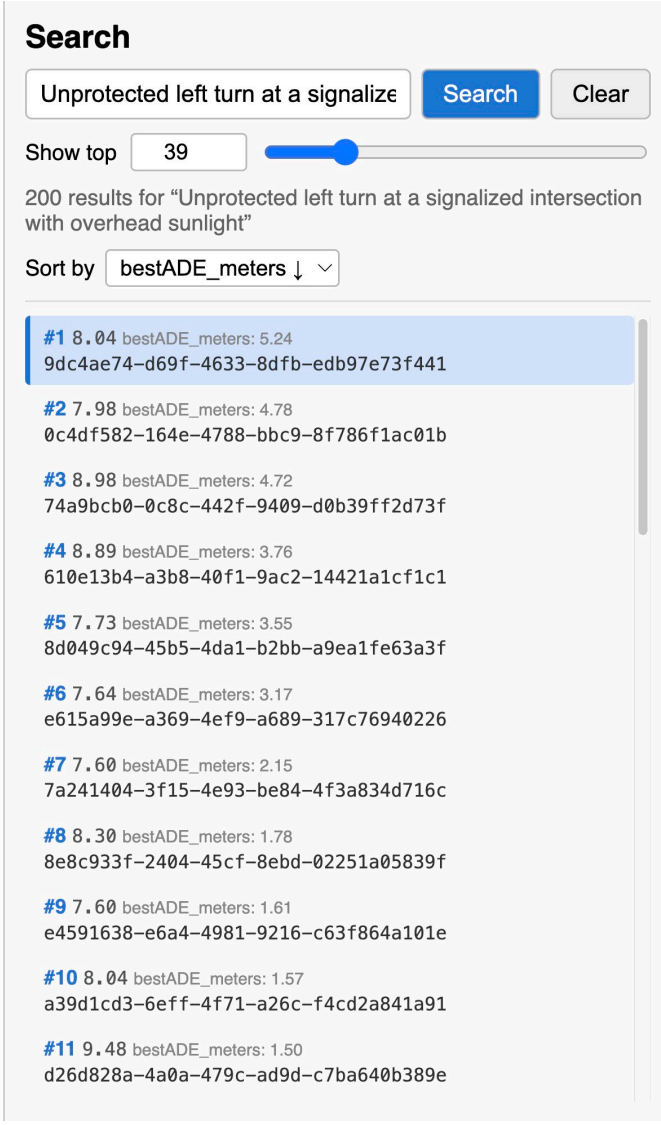




https://www.nvidia.com/en-us/solutions/autonomous-vehicles/alpamayo/

Testing Scores (Open Loop)





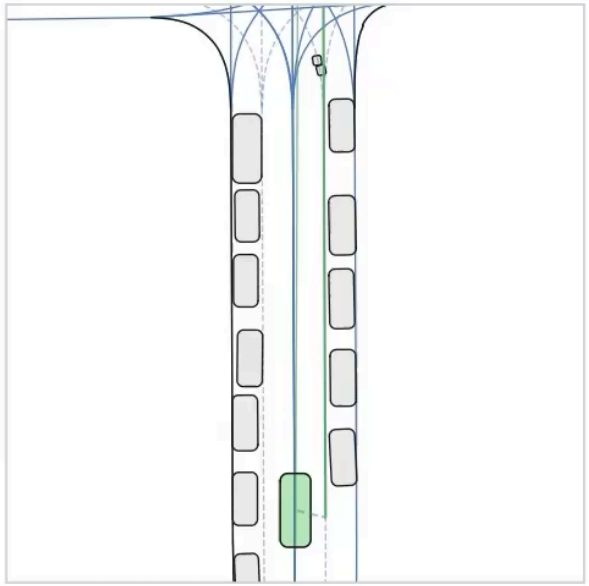
Search String:
“Unprotected left turn at a signalized intersection with overhead sunlight.”

ADE: Average displacement error (Lower is better)

New Scenario Generation

- Upon finding a potential risk (new unseen scenario or low-score existing scenario), we can feed the representation in the form of an embedding into the world foundation model to generate new synthetic data.
  https://www.nvidia.com/en-us/ai/cosmos/
- The synthetic data can be used later for training and close-loop testing.

Synthetic video of “Driving down a narrow suburban road on a cloudy day” generated through NVIDIA’s world foundation model.



Run: LR-08de2b96-87c9-444a-85ab-6fa7f54c583b
Clip: clipgt-clipgt-048b974e-1546-488a-b8f9-d32bff77f5aa
Batch: 0:76c908fa-24a5-11f1-82ab-4d6b050ee9c2

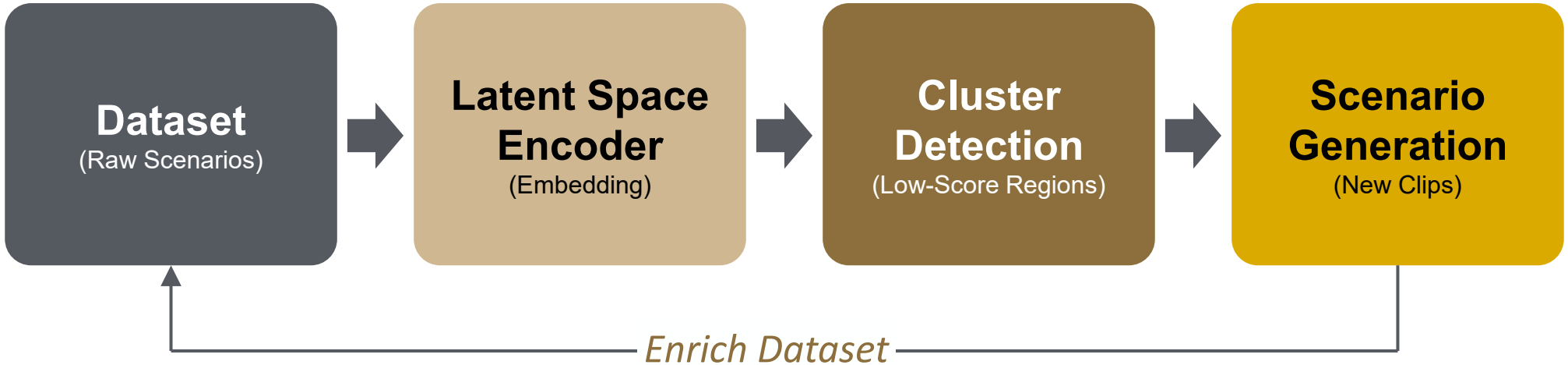
	Agg	Per-Ts
offroad_or_collision_at_fault	1.00 (max)	N/A
collision_any	0.00 (max)	0.00
collision_at_fault	0.00 (max)	N/A
collision_front	0.00 (max)	0.00
collision_lateral	0.00 (max)	0.00
collision_rear	0.00 (max)	0.00
offroad	1.00 (max)	1.00
dist_to_gt_trajectory	0.00 (max)	0.00
dist_to_gt_location	0.00 (max)	0.00
progress	0.00 (last)	0.00
progress_rel	1.00 (min)	1.00
safety_monitor_triggered	0.00 (max)	0.00

Time: 9787425000

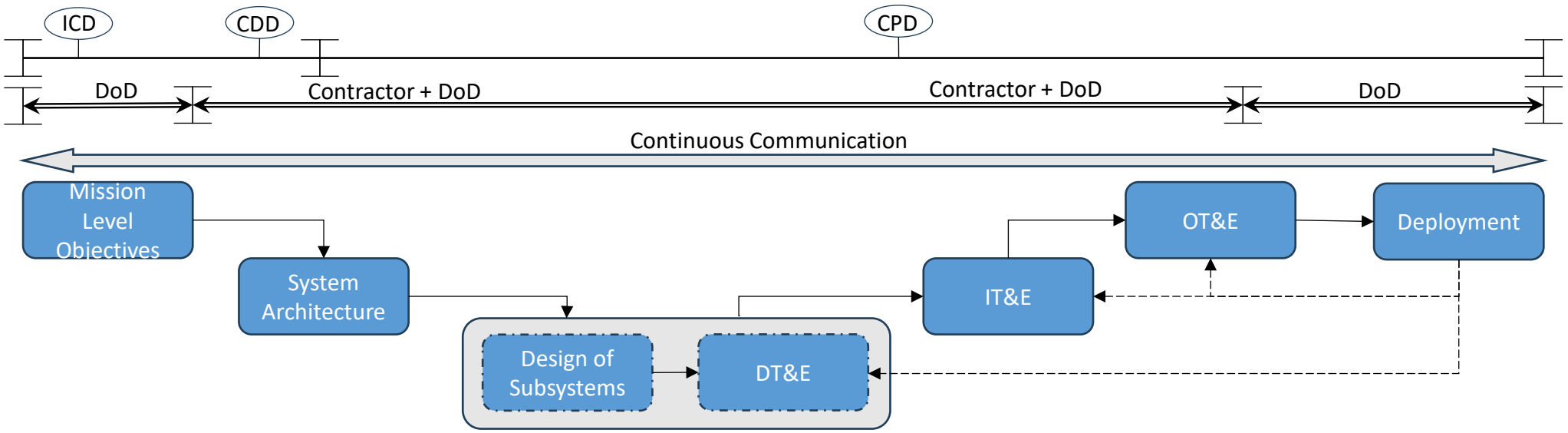
Closed-loop evaluation of the generated scene in AlpaSim (Metrics such as safety, comfort, and drivable area compliance can be measured).

Automatic Dataset Scaling

- All dataset points converted into a unified latent space representation. For example, an interaction and environment aware scenario embedding space.
- Automatically find within the space clusters with low score or no data points.
- Feed the latent space representation along with custom constraints to generate corresponding clips.



Implications for the Acquisition Process



- **Efficiency of T&E:** Reducing the cost and time for T&E using evidence generated throughout the systems engineering process.
- **Access Rights and IP:** Cost-benefit analysis of access to algorithms, training data, and test data.
- **Common Infrastructure:** World models, embedding models, and test environments.

Thank you!

Acknowledgment: This material is based upon work supported, in part, by the U.S. Department of Defense, Director, Operational Test and Evaluation (DOT&E) through the Office of the Assistant Secretary of Defense for Research and Engineering (ASD(R&E)) under Contract HQ003419D0003. Any views, opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Department of Defense (specifically DOT&E and ASD(R&E)).