Trajectory Forecasting

Ingrid Navarro
ingridn@andrew.cmu.edu

Agenda

- What is Trajectory Forecasting?
- Is Trajectory Forecasting Solved?
- Trajectory Forecasting in my PhD

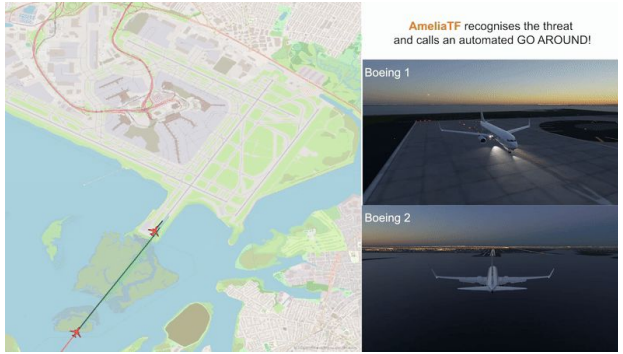
What is Trajectory Forecasting?

Why do we care about trajectory forecasting in robotics?

Trajectory forecasting because helps us understand how humans interact in shared spaces, anticipate their future actions, and use that knowledge to guide safe and effective decision-making and decision-support systems.

Source: KXAN

Source: giphy.com



Source: Amelia

Trajectory Forecasting

Goal: predict the future states for an agent or set of agents given a history of states and context surrounding an agent,

How do we define these states? What length of history is enough?

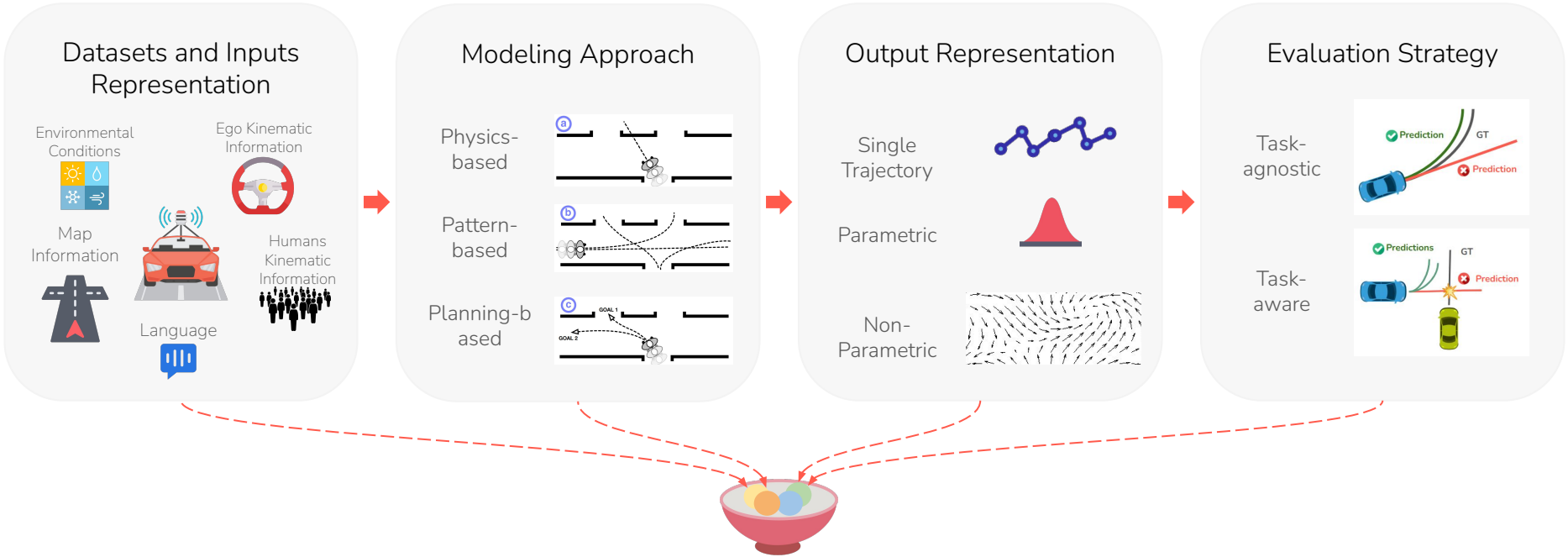
$$\hat{\mathbf{x}}_{\mathbf{F}} \sim p_\theta\left(\mathbf{x}_F \middle| \mathbf{x}_H, \mathbf{c}_H\right)$$

How do we represent the predictions? How far into the future should we predict?

How do we design a model of human behavior?

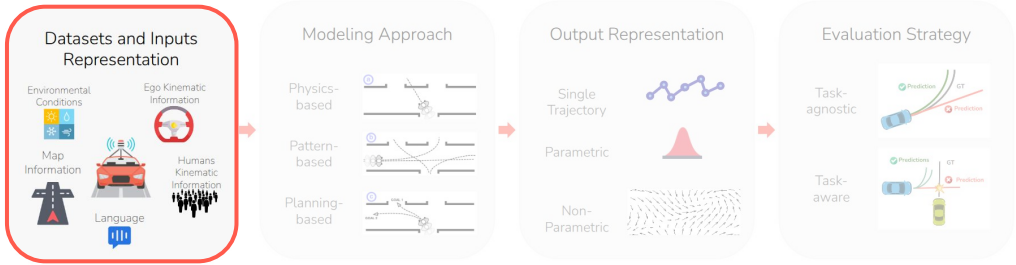
What context information is useful for predicting an agent's future?

Key Ingredients



Rudenko, A., Palmieri, L., Herman, M., Kitani, K. M., Gavrila, D. M., & Arras, K. O. (2020). *Human motion trajectory prediction: A survey*. The International Journal of Robotics Research, 39(8), 895-935.

Dataset and Inputs Representation



Common domains and use cases

Human Motion

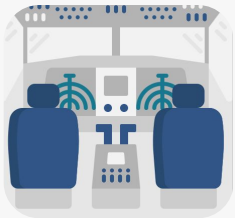
Use cases: service robots, surveillance, sport analysis.

Driving

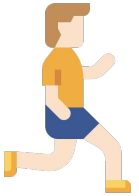
Use cases: personal self-driving, transportation of goods, traffic management.

Aviation

Use cases: traffic management, decision-support systems for task load reduction.



Human Motion

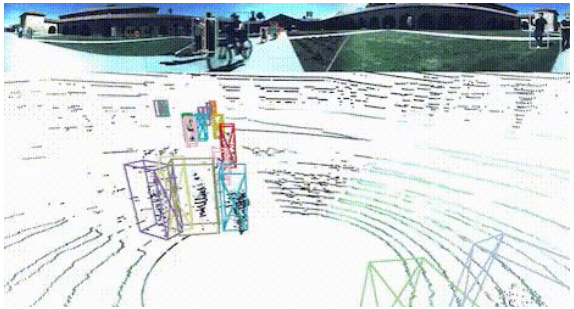


ETH[1] / UCY[2]

Subdomain: Surveillance

Captures human behavior in public spaces, like hotels, universities and stores.

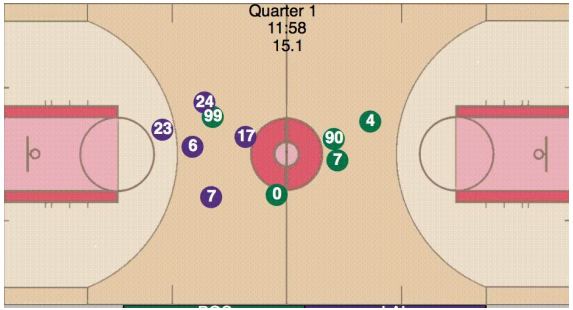
JRDB [3]



Subdomain: Service Robots (Social Navigation)

Captures human behavior in public spaces from a first-person-view.

SportVU Data [4]



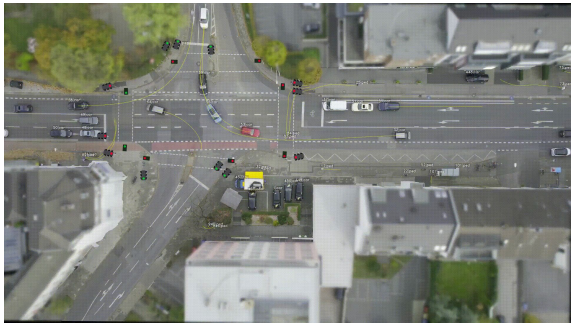
Subdomain: Sports

Utilizes a motion capture system that tracks players, balls, referees.

[1] Lerner, A., Chrysanthou, Y., & Lischinski, D. (2007, September). *Crowds by example*. In *Computer graphics forum* (Vol. 26, No. 3, pp. 655-664). Oxford, UK: Blackwell Publishing Ltd.
[2] Pellegrini, S., Ess, A., Schindler, K., & Van Gool, L. (2009, September). *You'll never walk alone: Modeling social behavior for multi-target tracking.* In 2009 IEEE 12th international conference on computer vision (pp. 261-268). IEEE.
[3] Martin-Martin, R., Patel, M., Rezatofighi, H., Shenoi, A., Gwak, J., Frankel, E., ... & Savarese, S. (2021). *Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. IEEE transactions on pattern analysis and machine intelligence*, *45*(6), 6748-6765.
[4] *SportsVU Data*: https://github.com/linouk23/NBA-Player-Movements/tree/master/examples

Urban/Highway Driving

LevelXData (inD, exiD, hiD) [1]



Subdomain: Traffic Management

Captures urban settings from a top-down perspective, typically using RGB data. Focus is generally on developing intelligent mobility solutions.

WOMD [2]

Subdomain: Autonomous Driving

Captures urban/highway interactions from first-person view.

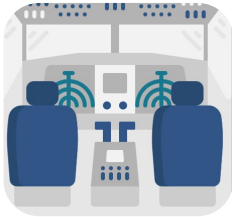
MAN Truck Scenes [3]

Subdomain: Autonomous Trucking

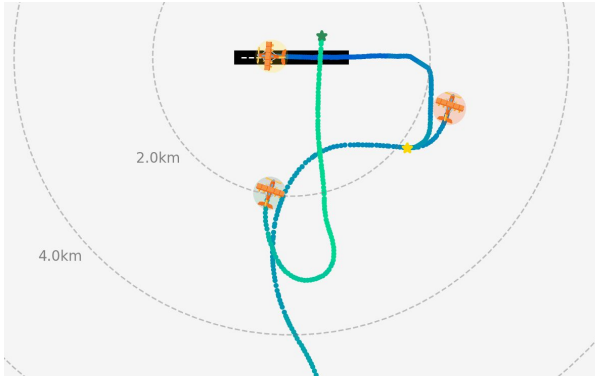
Focus on covering truck-specific surroundings (e.g., container terminals), larger local regions to understand the changes between chassis and cabin.

[1] *LevelXData*: https://levelxdata.com
[2] Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., ... & Anguelov, D. (2021). *Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset.* In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9710-9719).
[3] Fent, F., Kuttenreich, F., Ruch, F., Rizwin, F., Juergens, S., Lechermann, L., ... & Lienkamp, M. (2024). Man truckscenes: A multimodal dataset for autonomous trucking in diverse conditions. Advances in Neural Information Processing Systems, 37, 62062-62082.

Aviation



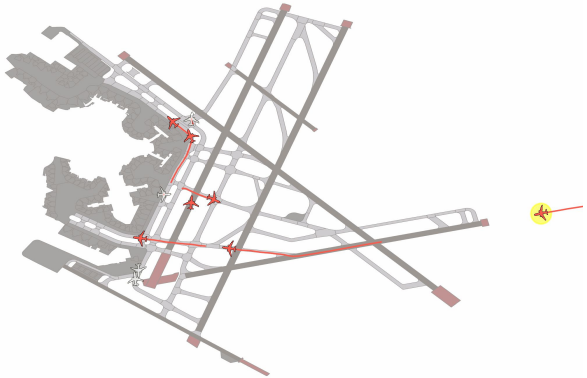
TrajAir [1]



Subdomain: Non-towered Airports

Captures aircraft interaction in a non-towered airport, i.e., airports where there is no centralized coordination via air traffic controllers. Interactions are centralized.

Amelia [2]

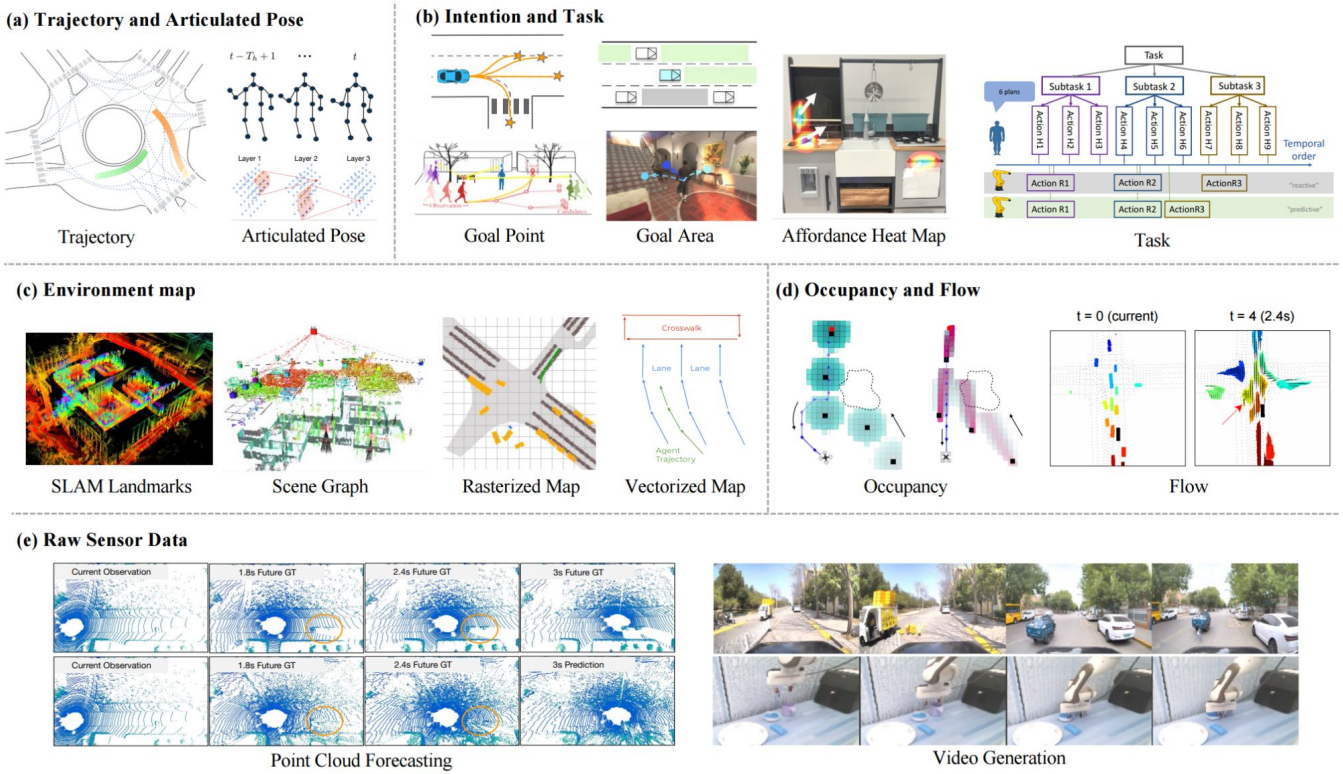


Subdomain: Airport Surface Movement Operations

Captures diverse airport operations in terminal airspace. Airports are primarily towered, meaning there is a centralized authority coordinating interactions.

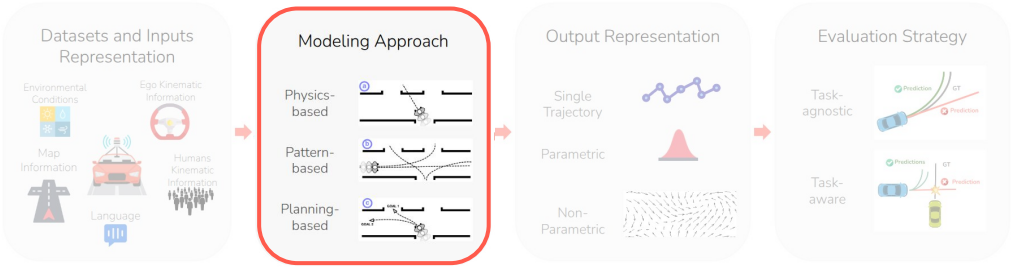
[1] Patrikar, J., Moon, B., Oh, J., & Scherer, S. (2022, May). **Predicting like a pilot: Dataset and method to predict socially-aware aircraft trajectories in non-towered terminal airspace**. In 2022 international conference on robotics and automation (icra) (pp. 2525-2531). IEEE.
[2] Navarro, I., Ortega, P., Patrikar, J., Wang, H., Ye, Z., Park, J. H., ... & Scherer, S. (2024). *AmeliaTF: A Large Model and Dataset for Airport Surface Movement Forecasting*. In AIAA AVIATION FORUM AND ASCEND 2024 (p. 4251).

Input Representations



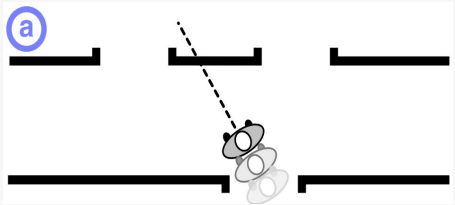
Wang, L., Lavoie, M. A., Papais, S., Nisar, B., Chen, Y., Ding, W., ... & Waslander, S. (2025). *Deployable and Generalizable Motion Prediction: Taxonomy, Open Challenges and Future Directions*. arXiv preprint arXiv:2505.09074.

Modeling Approaches



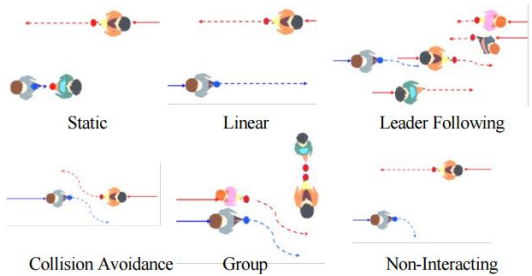
Physics-based Models

Motion is predicted by forward simulating a set of dynamics equations that follow a physics-inspired model.

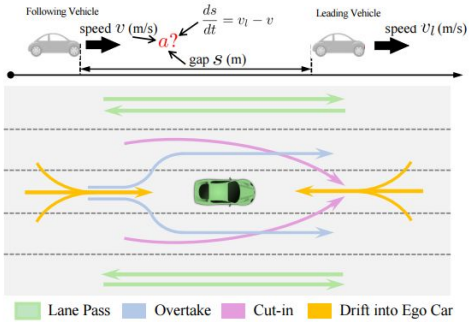


Generally simple and acceptable under mild conditions, such as short-horizon predictions.

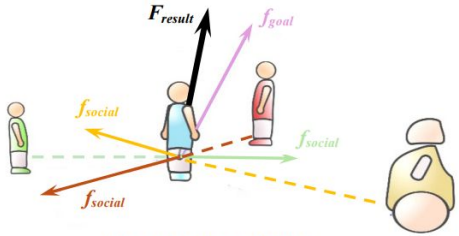
Difficult to effectively design hand-crafted models that account for all environmental and social complexities.



(a) Pedestrian Motion Patterns



(b) Vehicle Motion Patterns



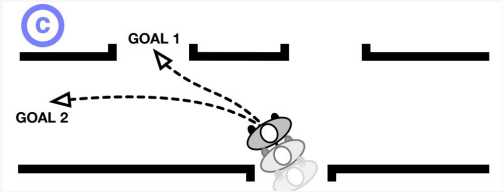
(c) Social Force Model

(d) Occupancy Probability

Wang, L., Lavoie, M. A., Papais, S., Nisar, B., Chen, Y., Ding, W., ... & Waslander, S. (2025). *Deployable and Generalizable Motion Prediction: Taxonomy, Open Challenges and Future Directions*. arXiv preprint arXiv:2505.09074.

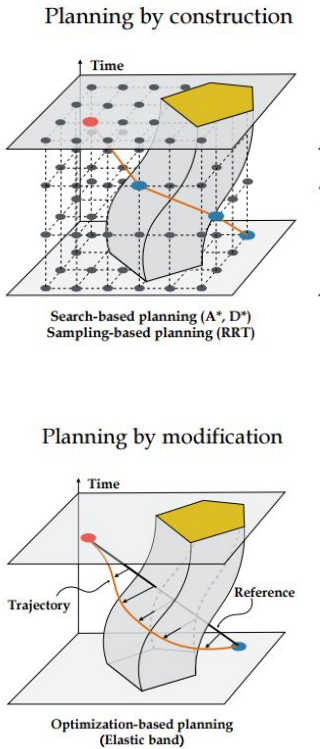
Planning-based Models

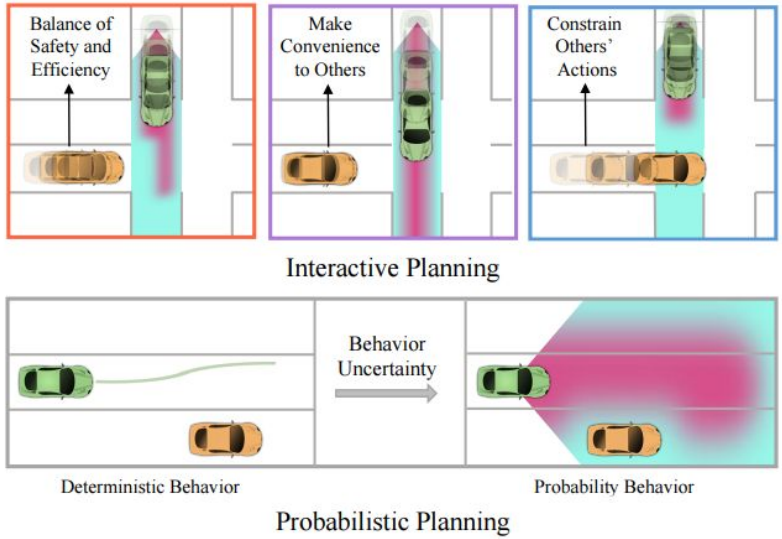
Motion is predicted by explicitly reasoning about the agent's long-term motion goals, and computing possible paths that attain those goals.



Generally good in structured settings.

Often assume agents are rational and make optimal motion decisions. May not generalize well under complex and less structured settings.

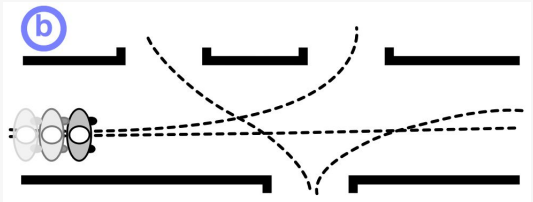




Wang, L., Lavoie, M. A., Papais, S., Nisar, B., Chen, Y., Ding, W., ... & Waslander, S. (2025). *Deployable and Generalizable Motion Prediction: Taxonomy, Open Challenges and Future Directions*. arXiv preprint arXiv:2505.09074.

Pattern-based Models

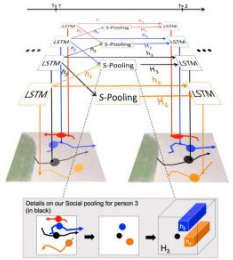
Approximate a dynamics function from data, discovering statistical behavioral patterns.



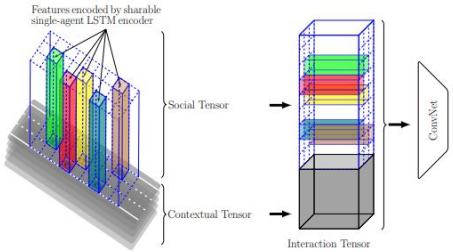
Can capture complex motion patterns directly from the data. More suitable for longer-horizon settings.

Generalizability to unseen settings is challenging.

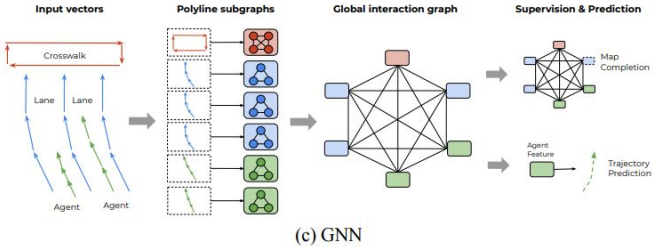
Main focus of this lecture!



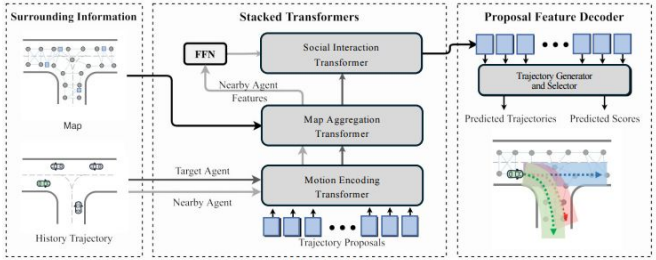
(a) RNN



(b) CNN



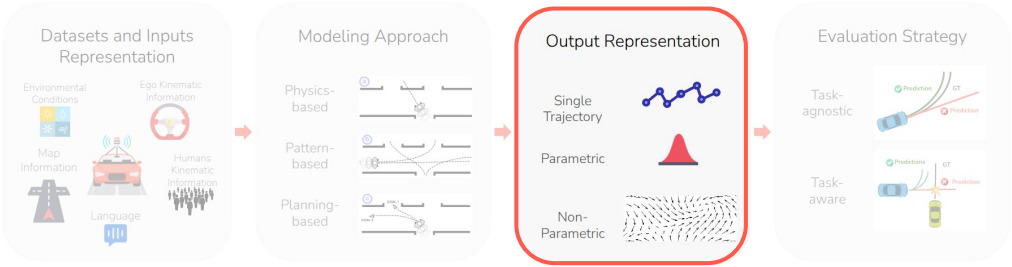
(c) GNN



(d) Attention

Wang, L., Lavoie, M. A., Papais, S., Nisar, B., Chen, Y., Ding, W., ... & Waslander, S. (2025). *Deployable and Generalizable Motion Prediction: Taxonomy, Open Challenges and Future Directions*. arXiv preprint arXiv:2505.09074.

Output Representations



Types of Predictions

Single Trajectory

Model predicts a single, deterministic trajectory.

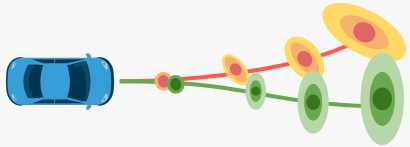
Simple to implement.

Cannot characterize multimodality and uncertainty.

Examples: RNNs

Parametric

Model outputs the parameters of a probability distribution, or mixture of distributions.



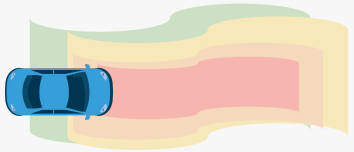
Captures prediction uncertainty, and mixtures can represent multiple possible futures.

Underlying distribution might not match the true distribution. Potential mixture collapse.

Examples: GMMs, C-VAEs

Non-Parametric

Model outputs trajectory samples without assuming a parametric distribution.



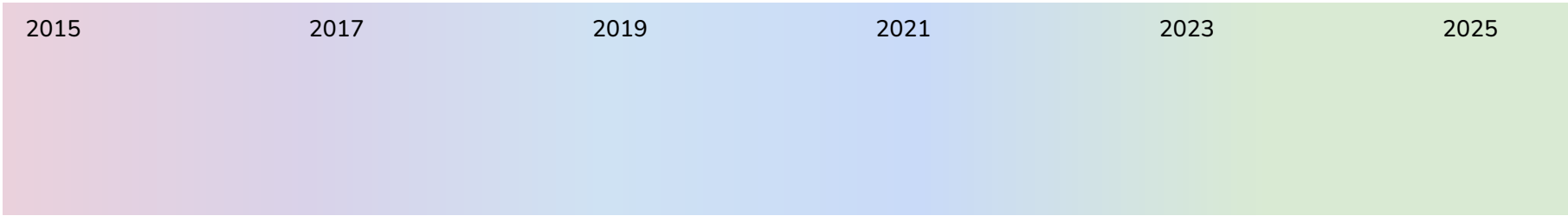
Flexible at capturing complex multimodal behaviors.

Often more difficult to evaluate and computationally expensive.

Examples: Normalizing Flows

Evolution of Pattern-based Approaches

Pattern-based Approaches Timeline



LSTMs + Social Pooling

2015 2017 2019 2021 2023 2025

Idea: Use recurrent sequence models (LSTMs) to model agent temporal dependencies and add a “social pooling” layer to let hidden states of nearby agents influence each other.

Representative Work: Social LSTM (Alahi et. al, CVPR 2016)

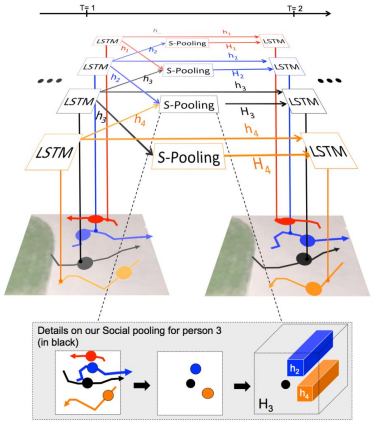
Influence: First type of strategy to replace hand-crafted rules with *learned* interactions. Established a problem formulation that was widely explored by many follow-up works.

Strengths: Handle variable length histories; relatively simple and intuitive; strong performance in short-horizon settings.

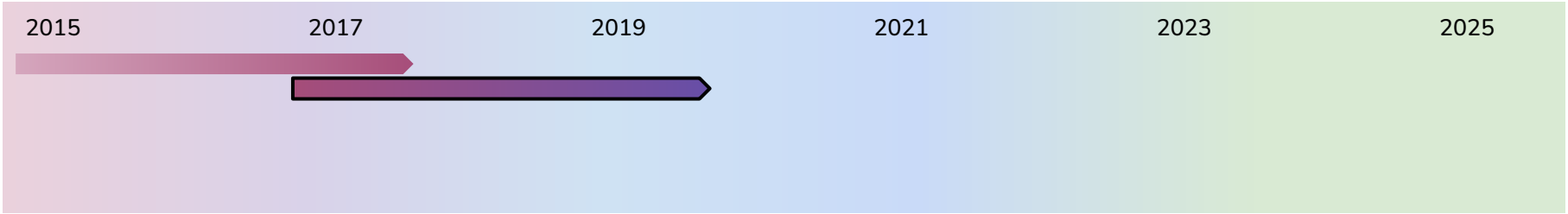
Limitations: Unable to represent multiple futures, relied on hand-designed pooling (e.g., k-closest agents).

Social LSTM:
Human Trajectory Prediction in Crowded Spaces

Alexandre Alahi*, Kratarth Goel*, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, Silvio Savarese
Stanford University



LSTMs + Social Pooling + GANs



Idea: Use GANs to produce multiple plausible futures but also train models adversarially to produce socially plausible trajectories.

Representative Work: Social GAN (Gupta et. al, CVPR 2018)

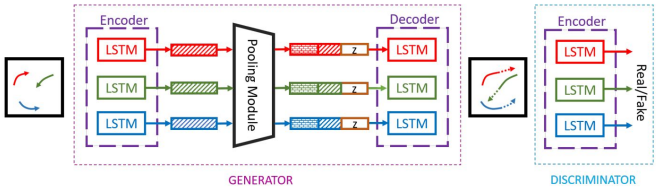
Influence: Addressed the problem of multi-modality

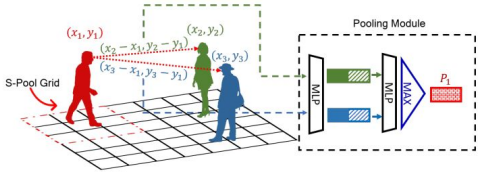
Strengths: Improved sample realism and diversity.

Limitations: Difficult to train, evaluation of diversity/quality/realism was hard; no likelihood associated to a sample.

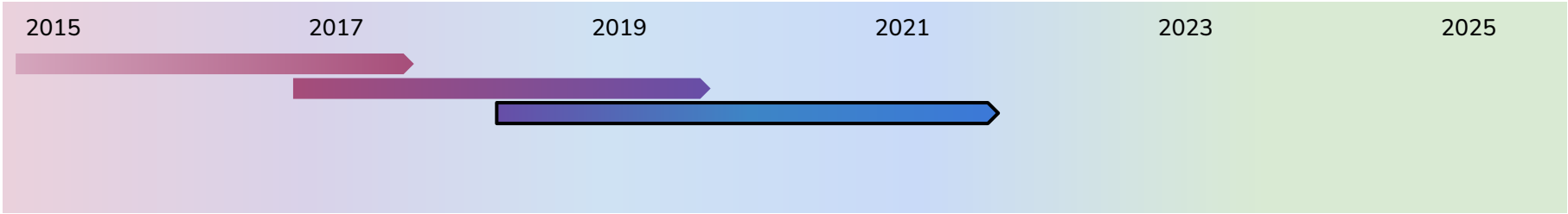
Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks

Agrim Gupta[1] Justin Johnson[1] Li Fei-Fei[1] Silvio Savarese[1] Alexandre Alahi[1,2]
Stanford University[1] École Polytechnique Fédérate de Lausanne[2]





Structured Stochastic Models



Idea: Represent scenes as spatio-temporal graphs (agents=nodes, interactions=edges), use probabilistic models to decode trajectory information (C-VAEs, VRNN, GMMs).

Representative Work: Trajectron (Ivanovic et. al, ICCV 2019), Multipath (Varadarajan et al, ICRA 2021).

Influence: Enabled conditioning and principled/structured representations.

Strengths: Probabilistic methods for representing multimodality and uncertainty, can incorporate context.

Limitations: Constructing graph-based representations still relied on heuristics. Mode collapse.

The Trajectron: Probabilistic Multi-Agent Trajectory Modeling With Dynamic Spatiotemporal Graphs

Boris Ivanovic Marco Pavone
Stanford University

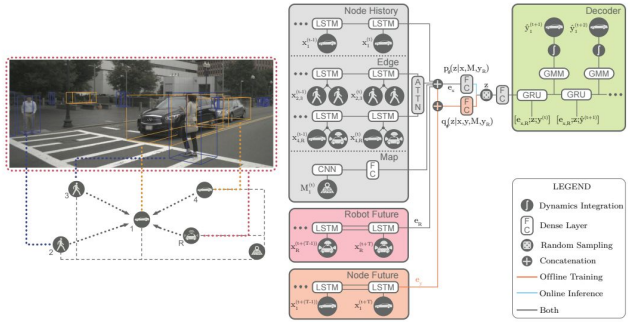
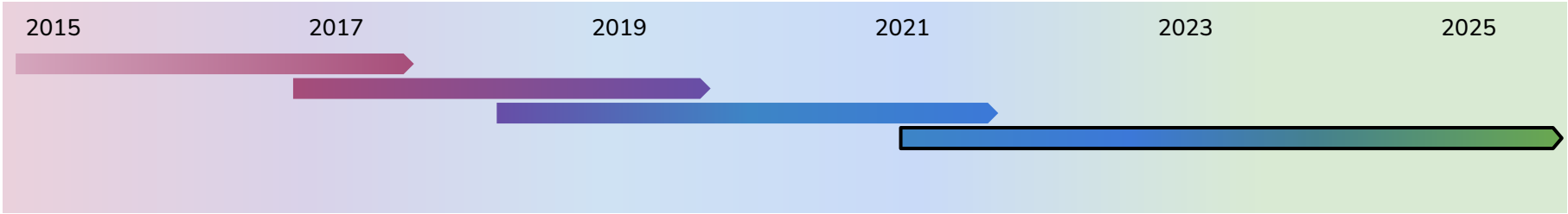


Figure is actually Trajectron++ (2021)

Transformer and Attention-based Models



Idea: Use transformers to jointly model social, temporal and other contextual relationships. Other attention-based models (e.g., GATs) were widely explored here.

Representative Work(s): Agent-Former (Yuan et al, ICCV 2021), Scene-Transformer (Ngiam et al, ICLR 2021), Motion-Transformer (Shi et al, NeurIPS 2022), Wayformer (Nayakanti et al, ICRA 2023), DAGNet (Monti et al, ICPR 2021) and many others.

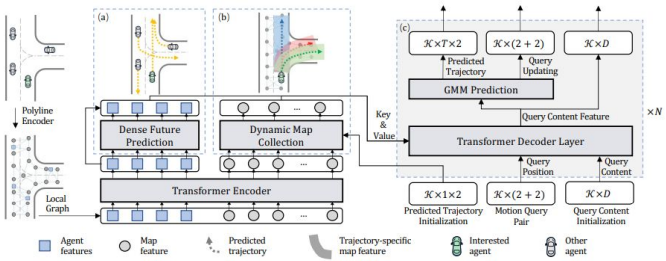
Influence: Attention enabled better, longer-range modeling; worked well with large-scale datasets; enabled much more design flexibility.

Strengths: Parallelizable, good at capturing complex dependencies, longer contexts, easily compatible with deterministic and probabilistic models.

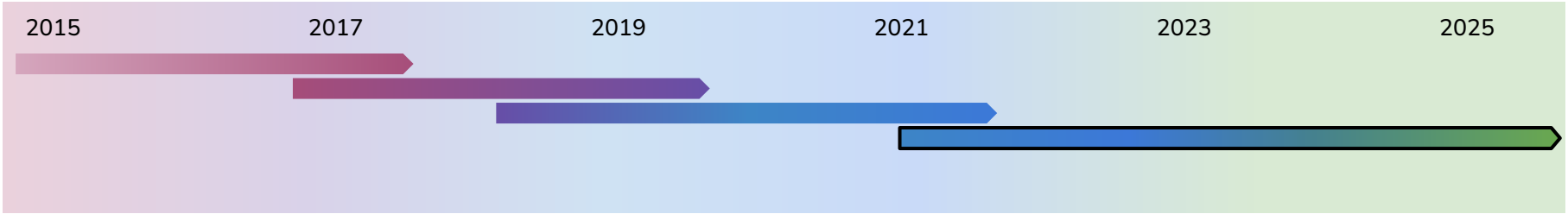
Limitations: Data hungry, computational expensive for larger scenes and context.

Motion Transformer with Global Intention Localization and Local Movement Refinement

Shaoshuai Shi, Li Jiang, Dengxin Dai, Bernt Schiele
Max Planck Institute for Informatics, Saarland Informatics Campus
{sshi, lijiang, ddai, schiele}@mpi-inf.mpg.de

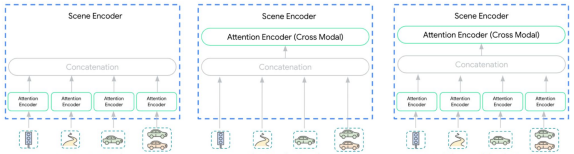


Transformer-based Models

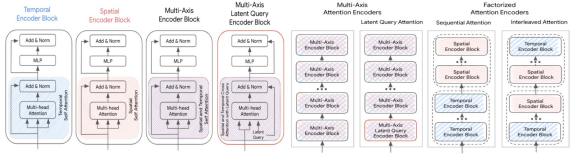


Wayformer: Motion Forecasting via Simple & Efficient Attention Networks

Nigamaa Nayakanti* nigamaa@waymo.com
Rami Al-Rfou* rmyeid@waymo.com
Aurick Zhou aurickz@waymo.com
Kratarth Goel kratarth@waymo.com
Khaled S. Refaat krefaat@waymo.com
Benjamin Sapp bensapp@waymo.com



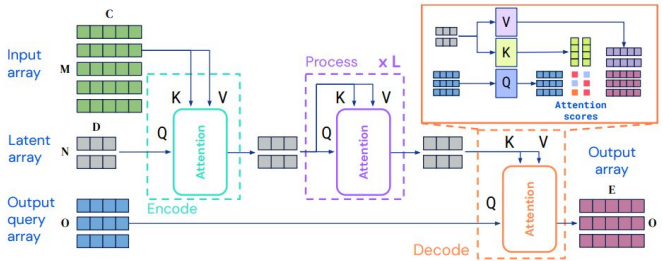
(a) Late Fusion (b) Early Fusion (c) Hierarchical Fusion



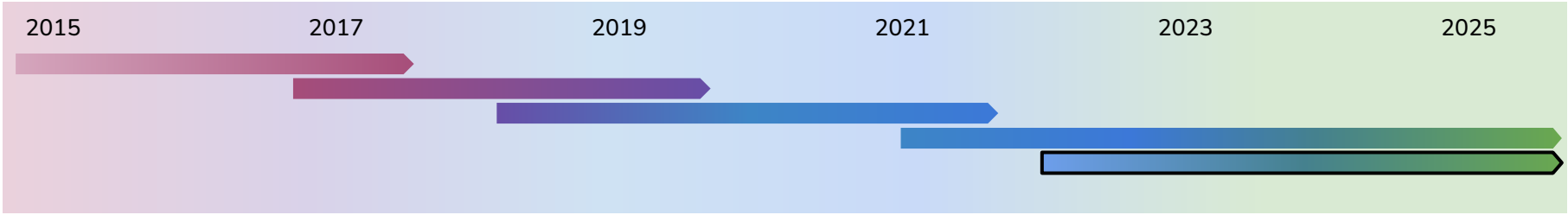
(a) Encoder Blocks (b) Encoders

PERCEIVER IO: A GENERAL ARCHITECTURE FOR STRUCTURED INPUTS & OUTPUTS

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu,
David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff,
Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, João Carreira
DeepMind



Diffusion-based Models



Idea: Learn joint distributions by gradually denoising samples.

Representative Work: Scene-Diffuser (Jiang et al, NeurIPS, 2024).

Influence: Enabled controllability, which propelled other subfields like scenario generation.

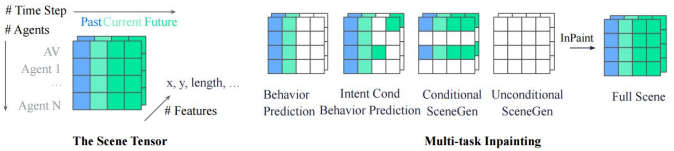
Strengths: Strong sample diversity and realism; flexibility / controllability.

Limitations: Denoising can have a high inference cost.

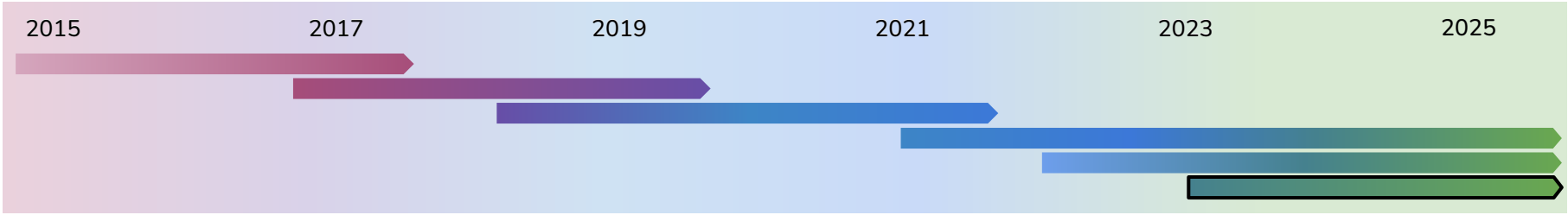
SceneDiffuser: Efficient and Controllable Driving Simulation Initialization and Rollout

Chiyu Max Jiang Yijing Bai* Andre Cornman* Christopher Davis* Xiukun Huang*
Hong Jeon* Sakshum Kulshrestha* John Lambert* Shuangyu Li* Xuanyu Zhou*
Carlos Fuertes Chang Yuan Mingxing Tan Yin Zhou Dragomir Anguelov

Waymo LLC



Trajectory Forecasting + Large Models



Idea: Use LLMs/VLMs to inject high-level reasoning priors into trajectory forecasting models.

Representative Work: TBD. See figure on the left for a taxonomy.

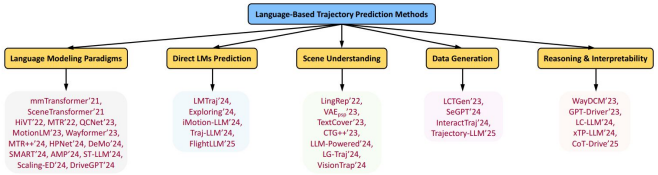
Influence: Providing predictors with strong commonsense and open-world reasoning.

Strengths: Strong zero-shot, few-shot reasoning; flexible conditioning.

Limitations: Computationally expensive; prone to hallucinations; often not great at spatial reasoning.

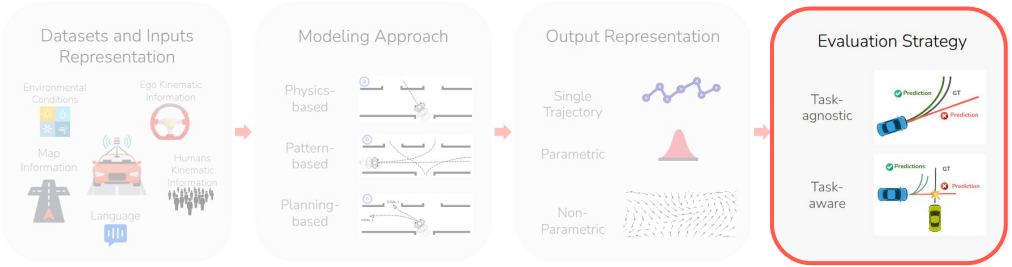
Trajectory Prediction Meets Large Language Models: A Survey

Yi Xu Ruining Yang Yitian Zhang Yizhou Wang
Jianglin Lu Mingyuan Zhang Lili Su Yun Fu
Department of Electrical and Computer Engineering, Northeastern University



How do we know we have a ***good*** model?

Evaluation Strategies

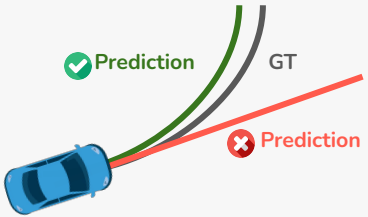


Task-Agnostic Evaluation

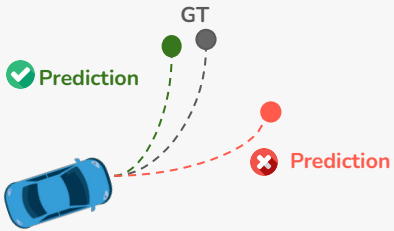
Metrics that focus on quantifying the similarity between the predicted and ground truth motion, but also capture the multi-modality of human behavior.

Accuracy / Realism

Minimum Average Displacement Error (mADE)

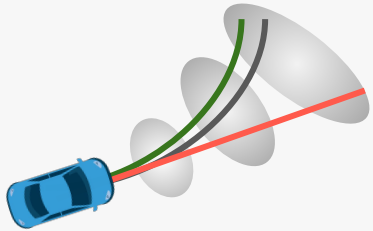


Minimum Final Displacement Error (mFDE)



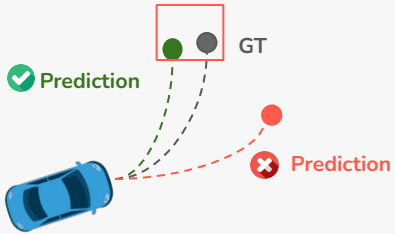
Uncertainty

Kernel Density Estimate (KDE)-based NLL



Coverage

Mean Average Precision (mAP)

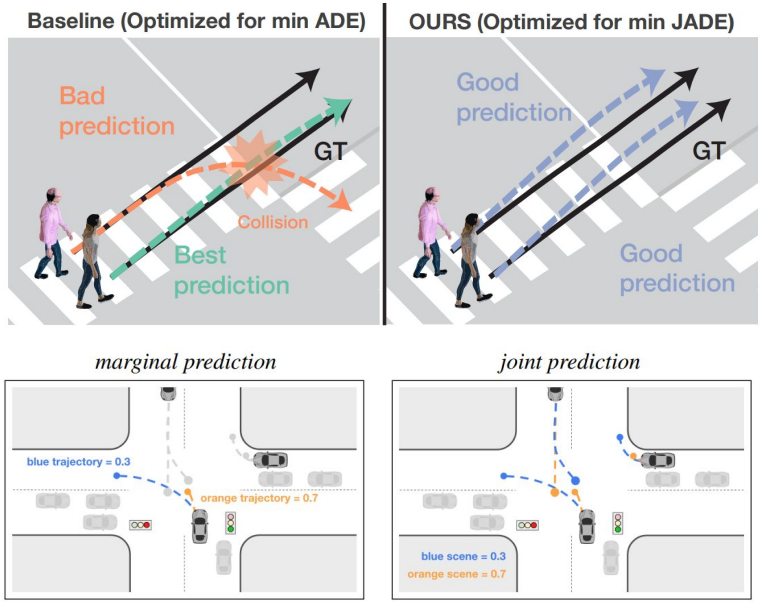


Wang, L., Lavoie, M. A., Papais, S., Nisar, B., Chen, Y., Ding, W., ... & Waslander, S. (2025). *Deployable and Generalizable Motion Prediction: Taxonomy, Open Challenges and Future Directions*. arXiv preprint arXiv:2505.09074.

Task-Agnostic Evaluation

However, single-agent metrics do not capture joint performance, which can lead to unnatural and inconsistent predicted social behavior, e.g.;

- Colliding trajectories
- Diverging trajectories for social groups



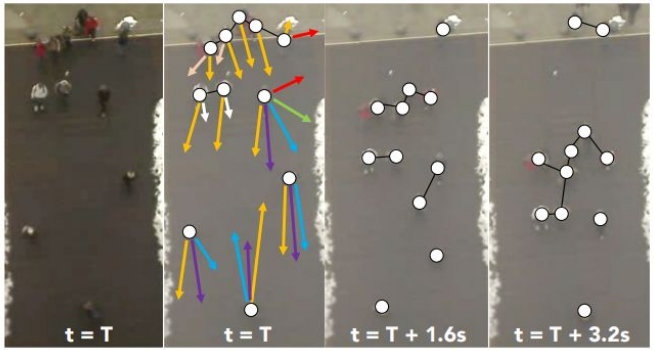
Sources: top figure [1], bottom figure [2]

[1] Weng, E., Hoshino, H., Ramanan, D., & Kitani, K. (2023). *Joint metrics matter: A better standard for trajectory forecasting.* In Proceedings of the IEEE/CVF International Conference on Computer Vision

[2] Ngiam, J., Caine, B., Vasudevan, V., Zhang, Z., Chiang, H. T. L., Ling, J., ... & Shlens, J. (2021). *Scene transformer: A unified architecture for predicting multiple agent trajectories.* arXiv preprint arXiv:2106.08417.

Benchmarks: ETH / UCY (Human Crowds)

Method	$\text{ADE}_{20}/\text{FDE}_{20} \downarrow$ (m), $K = 20$ Samples					
	ETH	Hotel	Univ	Zara1	Zara2	Average
SGAN [15]	0.81/1.52	0.72/1.61	0.60/1.26	0.34/0.69	0.42/0.84	0.58/1.18
SoPhie [44]	0.70/1.43	0.76/1.67	0.54/1.24	0.30/0.63	0.38/0.78	0.54/1.15
Transformer-TF [12]	0.61/1.12	0.18/0.30	0.35/0.65	0.22/0.38	0.17/0.32	0.31/0.55
STAR [55]	0.36/0.65	0.17/0.36	0.31/0.62	0.26/0.55	0.22/0.46	0.26/0.53
PECNet [34]	0.54/0.87	0.18/0.24	0.35/0.60	0.22/0.39	0.17/0.30	0.29/0.48
Trajectron++ [45]	**0.39**/0.83	**0.12/0.21**	**0.20/0.44**	**0.15**/0.33	**0.11**/0.25	**0.19**/0.41
Ours (AgentFormer)	0.45/**0.75**	0.14/0.22	0.25/0.45	0.18/**0.30**	0.14/**0.24**	0.23/**0.39**



Yuan, Y., Weng, X., Ou, Y., & Kitani, K. M. (2021). *Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting.* In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9813-9823).

Benchmarks: WOMD, nuScenes, Argoverse (Urban Driving)

Method Name	Object type	Evaluation time	Soft mAP v2	mAP v2	minADE	minFDE	Miss rate	Overlap Rate	Uses public model pretraining?
	All	Avg	Show results with						
MTR v3	All	Avg	0.4967	0.4859	0.5554	1.1062	0.1098	0.1279	
IMPACT	All	Avg	0.4801	0.4598	0.5563	1.1295	0.1087	0.1258	
RMP_Ensemble	All	Avg	0.4737	0.4531	0.5564	1.1188	0.1084	0.1259	
ModeSeq	All	Avg	0.4737	0.4665	0.5680	1.1766	0.1204	0.1275	
HeLoFusion	All	Avg	0.4732	0.4624	0.5690	1.1596	0.1183	0.1243	
TrajFlow	All	Avg	0.4710	0.4604	0.5714	1.1667	0.1162	0.1272	
BeTop	All	Avg	0.4698	0.4587	0.5716	1.1668	0.1183	0.1272	
BehaveOcc	All	Avg	0.4678	0.4566	0.5723	1.1668	0.1176	0.1278	
RMP-YOLO	All	Avg	0.4673	0.4523	0.5737	1.1697	0.1160	0.1266	
QMTR	All	Avg	0.4649	0.4445	0.5702	1.1627	0.1177	0.1269	
UniMotion	All	Avg	0.4648	0.4540	0.5693	1.1574	0.1156	0.1266	

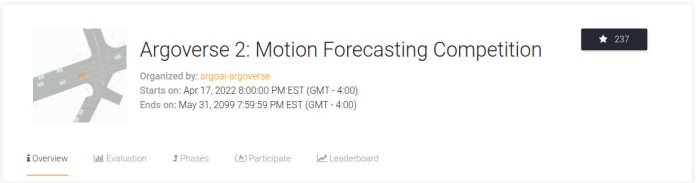
nuScenes prediction task

Leaderboard

Search:

Export as JSON

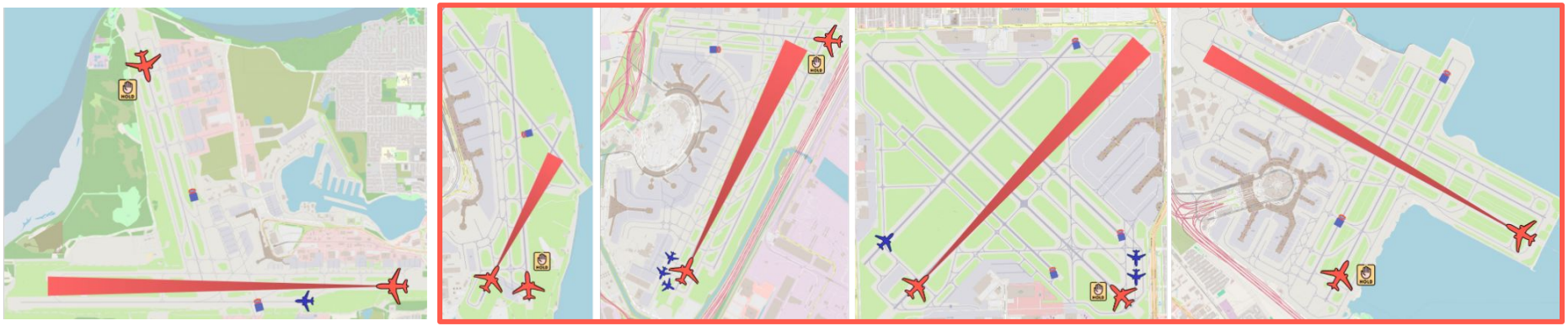
Date	Name	MinADE_5	MinADE_10	MissRateTopK_2_5	MissRateTopK_2_10	MinFDE_1	OffRoadRate
2022-05-25	DGCN_ST_LANE	1.092	0.974	46.99%	42.22%	3.624	0.124
2024-04-12	CASPFormer	1.148	1.148	47.83%	47.83%	6.702	0.011
2022-05-01	DSS	1.192	0.887	49.23%	30.99%	6.640	0.018
2022-09-06	Jack	1.211	0.894	56.65%	33.38%	7.502	0.019
2022-05-19	LaPred++	1.235	1.054	52.60%	46.19%	7.580	0.060
2022-03-07	PGP	1.267	0.945	51.90%	34.34%	7.171	0.028
2022-01-24	CASPNet_v2	1.275	1.010	53.15%	31.81%	7.023	0.011
2021-11-17	THOMAS	1.325	1.041	55.10%	42.40%	6.712	0.031
2022-01-23	JAL-MTP	1.356	0.988	54.88%	35.75%	7.870	0.025
2021-09-15	Autobot	1.368	1.026	62.23%	43.87%	8.190	0.025



waymo.com/open/challenges/2024/motion-prediction/
https://www.nuscenes.org/prediction?externalData=all&mapData=all&modalities=Any
https://eval.ai/web/challenges/challenge-page/1719/overview

Benchmarks: Domain Shift Generalization (Aviation)

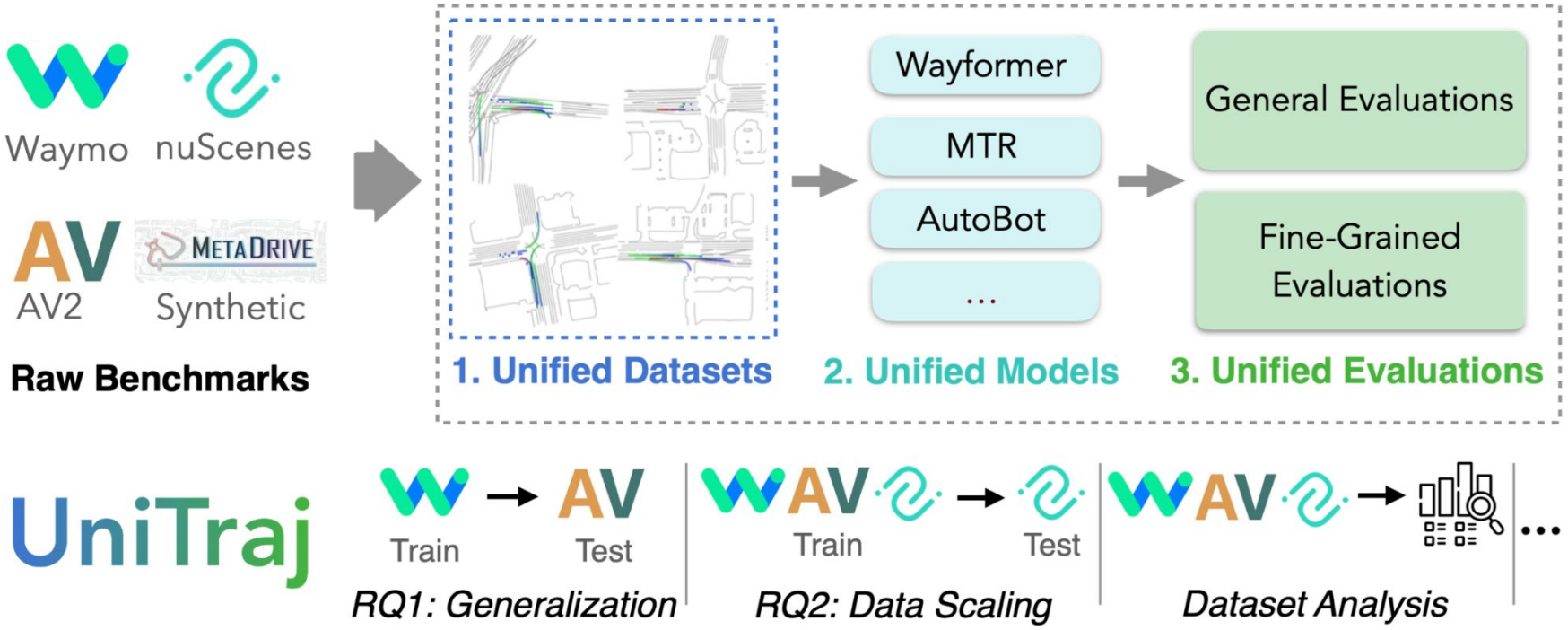
How can we design and evaluate models that generalize across a wide variety of contexts?



Amelia

Navarro, I., Ortega, P., Patrikar, J., Wang, H., Ye, Z., Park, J. H., ... & Scherer, S. (2024). *AmeliaTF: A Large Model and Dataset for Airport Surface Movement Forecasting.* In AIAA AVIATION FORUM AND ASCEND 2024

Benchmarks: UniTraj



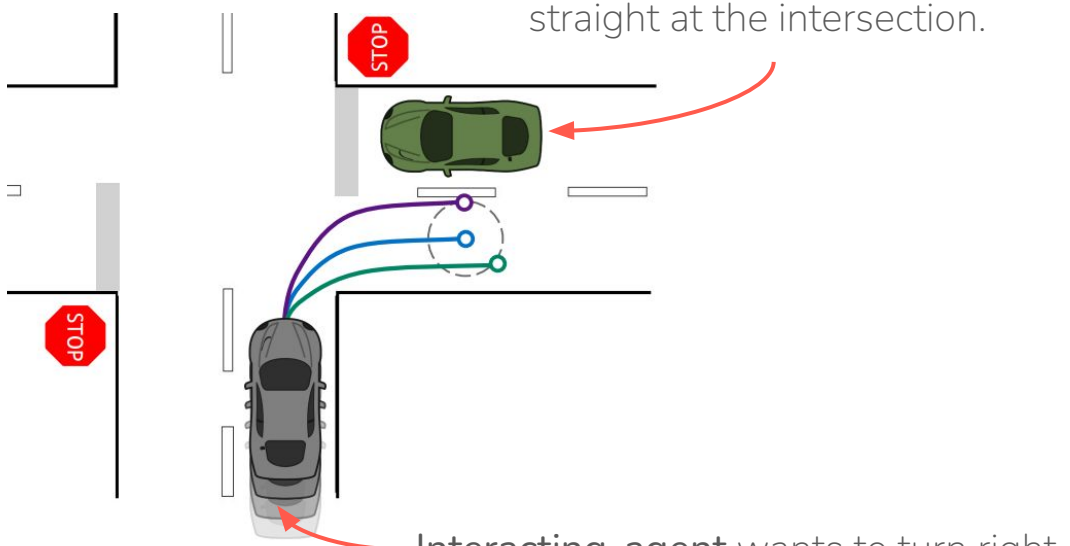
Feng, L., Bahari, M., Amor, K. M. B., Zablocki, É., Cord, M., & Alahi, A. (2024, September). ***Unitraj: A unified framework for scalable vehicle trajectory prediction***. In *European Conference on Computer Vision* (pp. 106-123). Cham: Springer Nature Switzerland.

Problems with Task-Agnostic Evaluation

The blue trajectory is the ground truth

The green and purple trajectories are the ego-agent's predictions.

Imagine that both were predicted to be equally likely and yield the same ADE value.



How do these predictions impact the ego-policy?

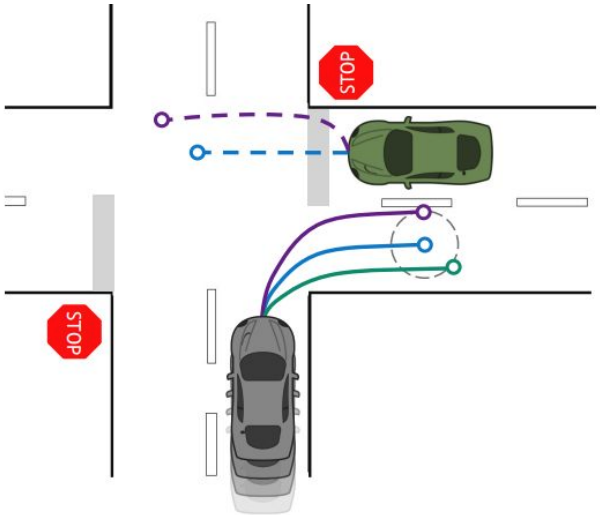
Ivanovic, B., & Pavone, M. (2021). *Rethinking trajectory forecasting evaluation.* arXiv preprint arXiv:2107.10297.

Problems with Task-Agnostic Evaluation

The blue trajectory is the ground truth

The green and purple trajectories are the ego-agent's predictions.

Imagine that both were predicted to be equally likely ~~and yield the same ADE value~~, but the purple one has lower ADE.



The purple prediction is used to inform the ego-agent.

The ego realizes it is a potentially dangerous trajectory.

Thus it takes an evasive maneuver, which incurs a higher planning cost than would've otherwise the ground truth plan.

Problems with Task-Agnostic Evaluation

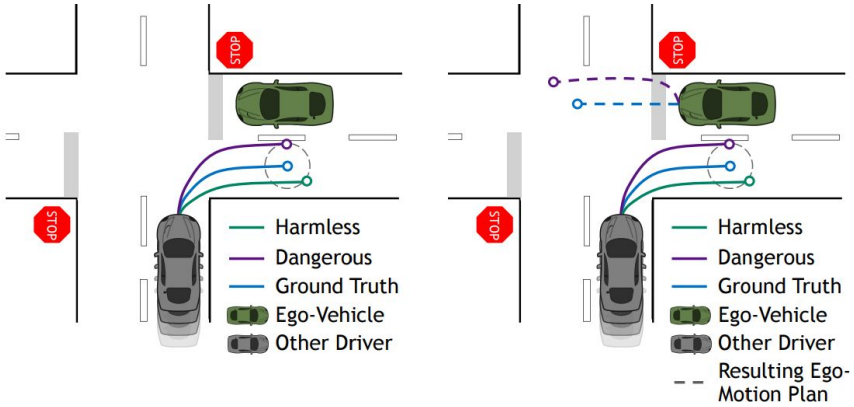
Task-agnostic metrics are disconnected from the downstream task and from real-world evaluation and deployment.

Ivanovic, B., & Pavone, M. (2021). *Rethinking trajectory forecasting evaluation.* arXiv preprint arXiv:2107.10297.

Task-Aware Evaluation

How do we design a cost or planning-informed metric?

- Focus is on evaluating and addressing:
  - Prediction under perception uncertainty [1].
  - Implications of prediction failures on downstream tasks [2, 3, 4].



[1] Stoler, B., Jana, M., Hwang, S., & Oh, J. (2023, October). *T2FPV: Dataset and method for correcting first-person view errors in pedestrian trajectory prediction.* In 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 4037-4044). IEEE.

[2] Ivanovic, B., & Pavone, M. (2021). *Rethinking trajectory forecasting evaluation.* arXiv preprint arXiv:2107.10297.

[3] Farid, A., Veer, S., Ivanovic, B., Leung, K., & Pavone, M. (2023, March). *Task-relevant failure detection for trajectory predictors in autonomous vehicles.* In Conference on Robot Learning (pp. 1959-1969). PMLR.

[4] Nakamura, K., Tian, T., & Bajcsy, A. (2025, January). *Not All Errors Are Made Equal: A Regret Metric for Detecting System-level Trajectory Prediction Failures.* In Conference on Robot Learning (pp. 4051-4065). PMLR.

Is Trajectory Forecasting Solved?

The State-of-the-Art

Motion Transformer (MTR), 1st Place Waymo Open Motion Prediction Challenge, 2022.

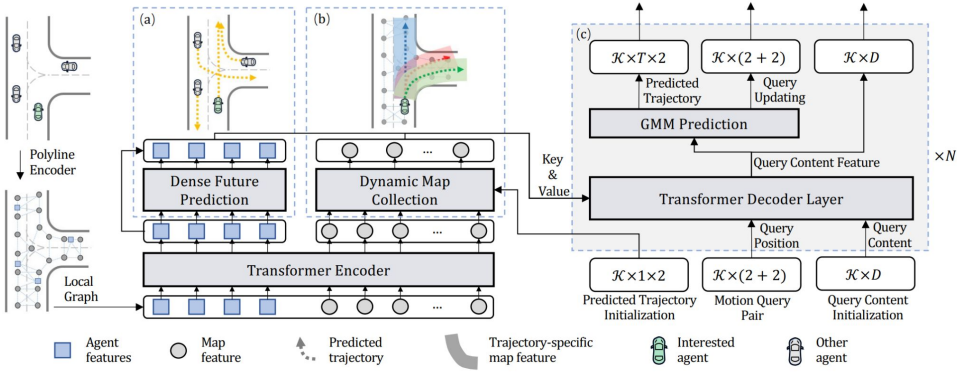
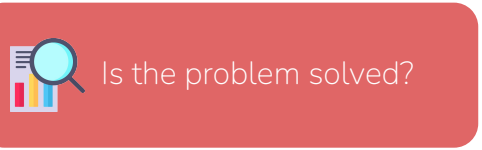


Table 1: Performance comparison of marginal motion prediction on the validation and test set of Waymo Open Motion Dataset. †: The results are shown in *italic* for reference since their performance is achieved with model ensemble techniques. We only evaluate our default setting MTR on the test set by submitting to official test server due to the limitation of submission times of WOMD.

	Method	Reference	minADE ↓	minFDE ↓	Miss Rate ↓	**mAP ↑**
Test	MotionCNN [26]	CVPRw 2021	0.7400	1.4936	0.2091	0.2136
	ReCoAt [66]	CVPRw 2021	0.7703	1.6668	0.2437	0.2711
	DenseTNT [21]	ICCV 2021	1.0387	1.5514	0.1573	0.3281
	SceneTransformer [37]	ICLR 2022	0.6117	**1.2116**	0.1564	0.2788
	MTR (Ours)	-	**0.6050**	1.2207	**0.1351**	**0.4129**
	†MultiPath++ [49]	ICRA 2022	*0.5557*	*1.1577*	*0.1340*	*0.4092*
	†MTR-Advanced-ens (Ours)	-	*0.5640*	*1.1344*	*0.1160*	*0.4492*

Shi, S., Jiang, L., Dai, D., & Schiele, B. (2022). *Motion transformer with global intention localization and local movement refinement.* Advances in Neural Information Processing Systems, 35, 6531-6543.

The State-of-the-Art



Waymo Open Motion Prediction Challenge leaderboard, 2024.

Method Name	Object type	Evaluation time	Soft mAP v2 ↓	mAP v2	minADE	minFDE	Miss rate	Overlap Rate
	All	Avg	Show results with					
MTR v3	All	Avg	0.4967	0.4859	0.5554	1.1062	0.1098	0.1279
IMPACT	All	Avg	0.4801	0.4598	0.5563	1.1295	0.1087	0.1258
RMP_Ensemble	All	Avg	0.4737	0.4531	0.5564	1.1188	0.1084	0.1259
ModeSeq	All	Avg	0.4737	0.4665	0.5680	1.1766	0.1204	0.1275
HeLoFusion	All	Avg	0.4732	0.4624	0.5690	1.1596	0.1183	0.1243
TrajFlow	All	Avg	0.4710	0.4604	0.5714	1.1667	0.1162	0.1272
BeTop	All	Avg	0.4698	0.4587	0.5716	1.1668	0.1183	0.1272
BehaveOcc	All	Avg	0.4678	0.4566	0.5723	1.1668	0.1176	0.1278
RMP-YOLO	All	Avg	0.4673	0.4523	0.5737	1.1697	0.1160	0.1266
QMTR	All	Avg	0.4649	0.4445	0.5702	1.1627	0.1177	0.1269
UniMotion	All	Avg	0.4648	0.4540	0.5693	1.1574	0.1156	0.1266
QMTR-V2	All	Avg	0.4646	0.4441	0.5700	1.1621	0.1174	0.1270

Variants of MTR

Evaluation metrics have not improved significantly since

https://waymo.com/open/challenges/2024/motion-prediction

Benchmark-Reality Gap

The Goal:

The Reality:

Human-involved rear-end accidents

1.9 per million miles traveled

vs.

ADS-involved rear-end accidents

9.1 per million miles traveled

Huang, Chunxi, and Xiao Wen. ***Characteristics of Rear-End Collisions: A Comparison between ADS-Involved Crashes and ADAS-Involved Crashes 2.***

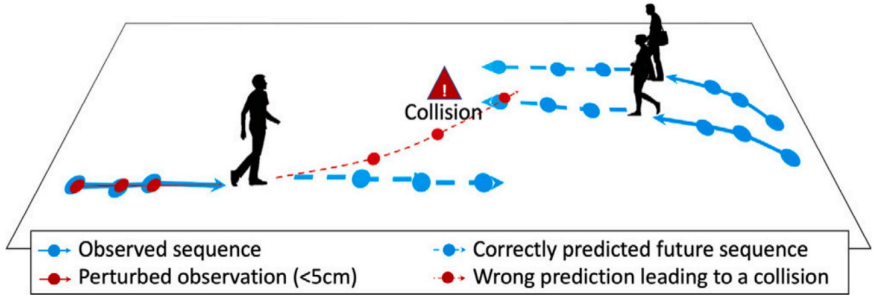
What's behind the generalization gap?

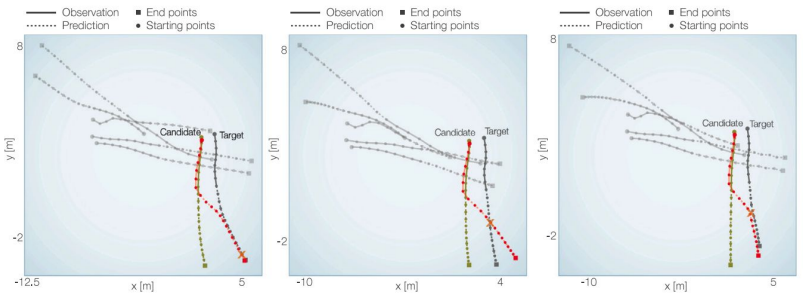
Existing challenges in the field:

Perception-level	Perception errors and perturbations to the input data
Data-level	Non-uniform data coverage
Control-level	How do we use forecasts downstream?

Brittleness to Perception Errors

Carefully-crafted perturbations can cause significant prediction failures that lead to unrealistic and/or unsafe behavior.

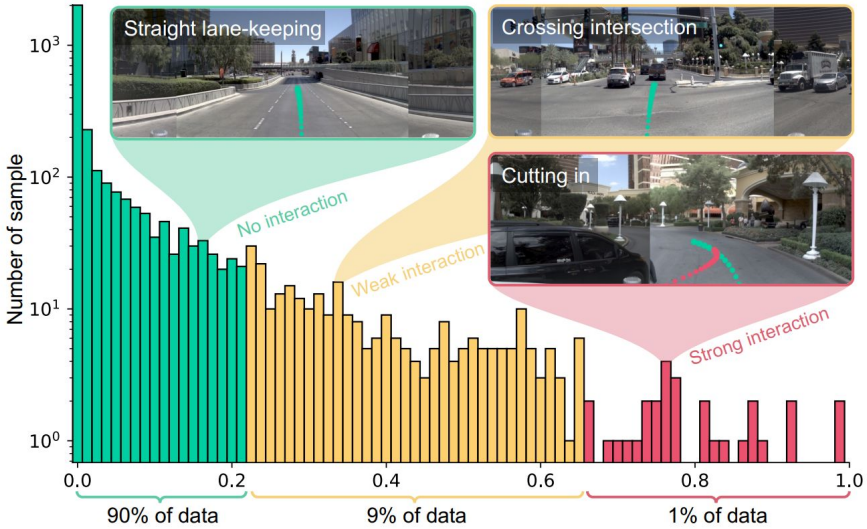




Saadatnejad, S., Bahari, M., Khorsandi, P., Saneian, M., Moosavi-Dezfooli, S. M., & Alahi, A. (2022). ***Are socially-aware trajectory prediction models really socially-aware?*** Transportation research part C: emerging technologies, 141, 103705.

Lack of Meaningful Interactivity

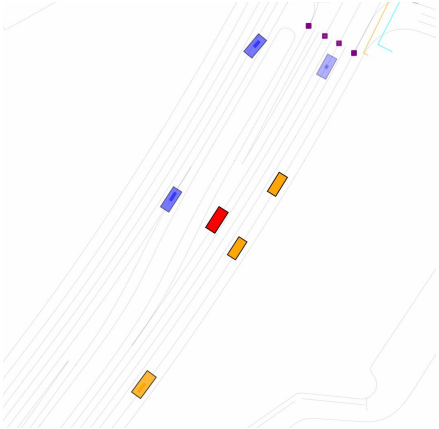
The majority of recorded driving data consists of uneventful driving with limited interactions [1].

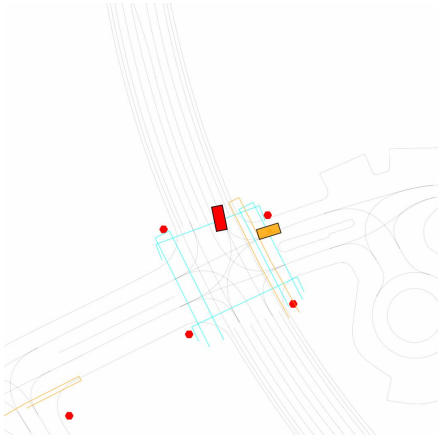


[1] Ding, W., Veer, S., Leung, K., Cao, Y., & Pavone, M. (2025). ***Surprise potential as a measure of interactivity in driving scenarios.*** arXiv preprint arXiv:2502.05677.

Lack of Meaningful Interactivity

The majority of recorded driving data consists of uneventful driving with limited interactions [1].





[1] Ding, W., Veer, S., Leung, K., Cao, Y., & Pavone, M. (2025). ***Surprise potential as a measure of interactivity in driving scenarios.*** arXiv preprint arXiv:2502.05677.

The Curse of Rarity

Complex and safety-critical scenarios in real-world datasets are rare, which makes leveraging real-world datasets to train and validate robust policies challenging.

[1] Stoler*, B., Navarro*, I., Jana, M., Hwang, S., Francis, J., & Oh, J. (2024, June). *Safeshift: Safety-informed distribution shifts for robust trajectory prediction in autonomous driving.* In 2024 IEEE Intelligent Vehicles Symposium (IV) (pp. 1179-1186). IEEE.
[2] Stoler, B., Navarro, I., Francis, J., & Oh, J. (2025). *SEAL: Towards safe autonomous driving via skill-enabled adversary learning for closed-loop scenario generation.* IEEE Robotics and Automation Letters, (99), 1-8.

Operational Restrictions

Ensuring uniform data coverage is difficult and expensive to attaining, which makes generalizability in unseen environment challenging.

Operational Restrictions:

Geofences

Weather

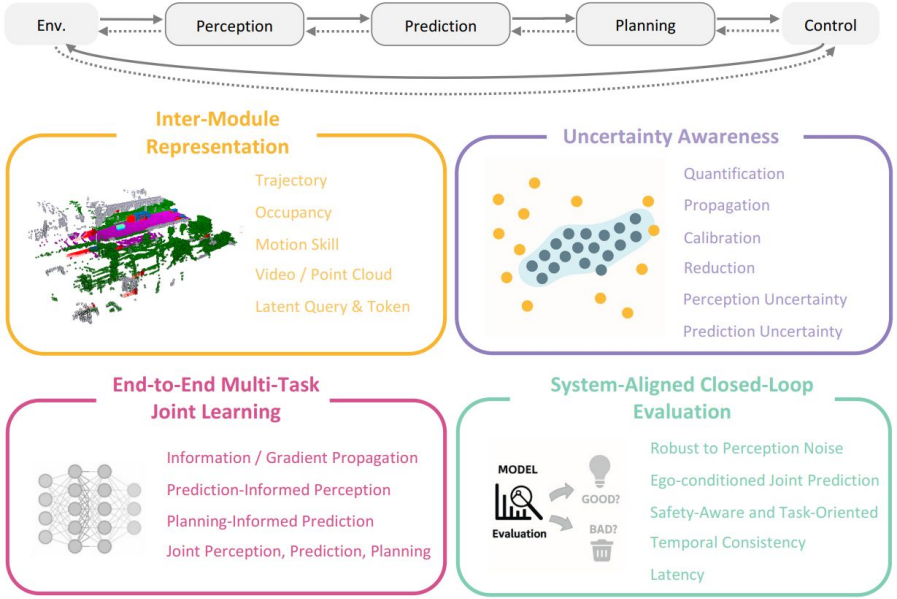
Time of Day

Road Type

Navarro, I., Ortega, P., Patrikar, J., Wang, H., Ye, Z., Park, J. H., ... & Scherer, S. (2024). *AmeliaTF: A Large Model and Dataset for Airport Surface Movement Forecasting.* In AIAA AVIATION FORUM AND ASCEND 2024

Deployability

How do we deploy trajectory forecasting models and guarantee generalizability and robustness?



Wang, L., Lavoie, M. A., Papais, S., Nisar, B., Chen, Y., Ding, W., ... & Waslander, S. (2025). *Deployable and Generalizable Motion Prediction: Taxonomy, Open Challenges and Future Directions*. arXiv preprint arXiv:2505.09074.

Trajectory Forecasting in my PhD

Some of Our Recent Works

Relevance to Trajectory Forecasting

SafeShift 

Safety-informed Distribution Shifts for Robust Trajectory Prediction in Autonomous Driving

IEEE Intelligent Vehicles Symposium, 2024

Scenario Mining; Robustness Benchmark

SEAL 

Skills-**E**nabled **A**dversary **L**earning for Closed-Loop Scenario Generation

IEEE Robotics and Automation Letters, 2025

Closed-Loop Evaluation; Robustness Benchmark

Amelia 

A Large Model and Dataset for Airport Surface Movement Forecasting

AIAA Aviation and Ascend Forum, 2024

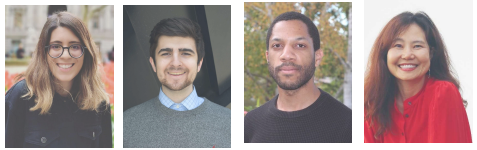
New Domain and Dataset; Domain Shift Benchmark

SoRTS 

Social **R**obot **T**ree **S**earch for Long-Horizon Navigation in Shared Airspace

IEEE Robotics and Automation Letters, 2024

New Domain; Model Deployability



SafeShift

Safety-informed Distribution Shift for Robust Trajectory Prediction in Autonomous Driving

IEEE Intelligent Vehicles Symposium, 2024

https://navars.xyz/safeshift

In the context of Trajectory Forecasting:

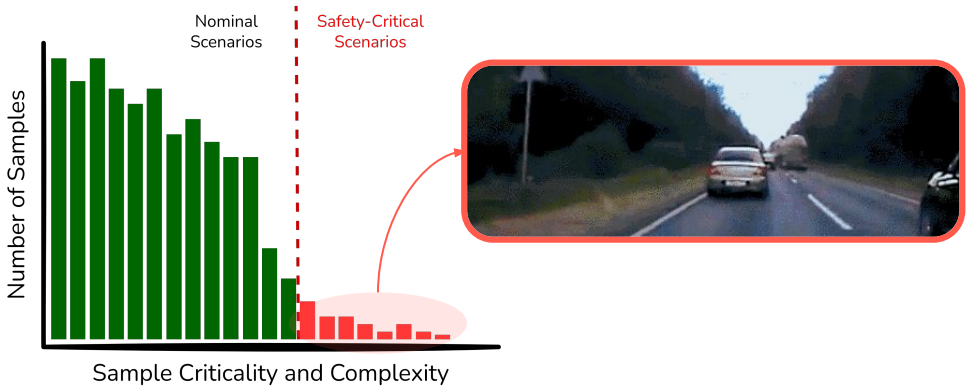
- A scenario characterization paradigm for trajectory datasets
- An out-of-distribution robustness benchmark
- Bonus: applied in industry!

STACK

Assessing the Generalizability of Autonomous Vehicles (AV)

The "Curse of Rarity":

- Safety-critical scenarios in real-world datasets are rare.
- Directly leveraging real-world datasets to train and validate robust policies is challenging.



Ding, Wenhao, Chejian Xu, Mansur Arief, Haohong Lin, Bo Li, and Ding Zhao. *A survey on safety-critical driving scenario generation—a methodological perspective.* IEEE Transactions on Intelligent Transportation Systems 24, no. 7 (2023): 6971-6988.

Assessing the Generalizability of AVs

Existing methodologies:

On-road testing [7]

\+ Realistic
\- Potentially dangerous

Scenario Generation [6]

\+ Not dangerous
\- Potentially unrealistic

[6] W. Ding, C. Xu, M. Arief, H. Lin, B. Li, and D. Zhao. *A survey on safety-critical driving scenario generation—a methodological perspective.* IEEE Transactions on Intelligent Transportation Systems, 2023.
[7] W. Huang, K. Wang, Y. Lv, and F. Zhu. *Autonomous vehicles testing methods review* in 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2016, pp. 163–168.

Assessing the Generalizability of AVs – An Alternative

Broadening the concept of safety:

Safety-relevance

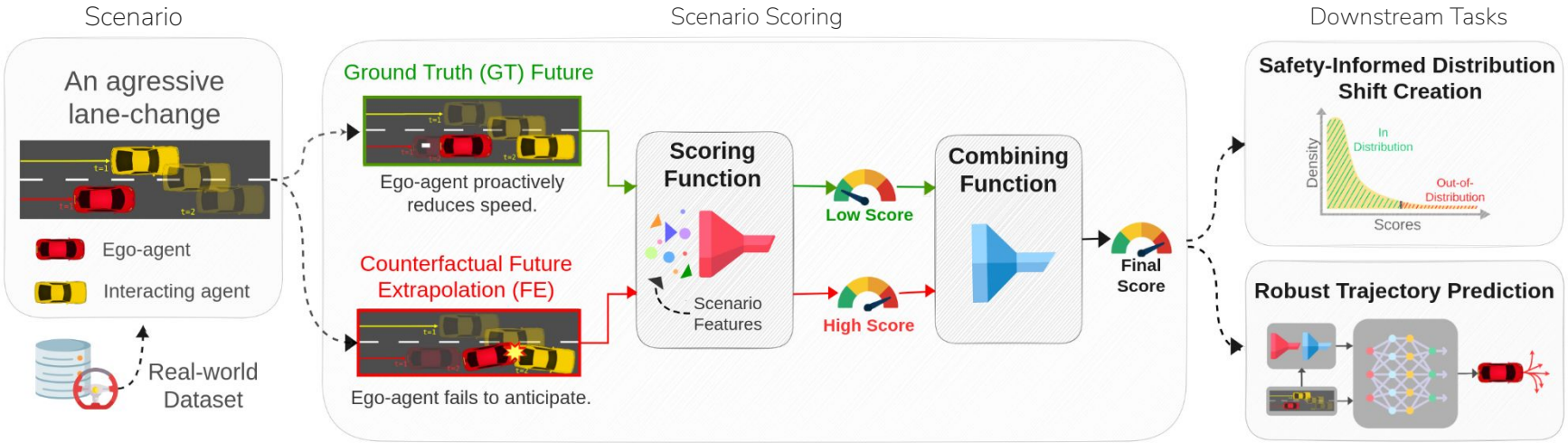
Acting proactively to avoid safety criticality, e.g., swerving, braking.

Acting near safety-critically, e.g., recklessly or distractedly.

Source: https://www.pbh2.com/wtf/close-call-gifs/5/

SafeShift

A scenario characterization framework through counterfactual probing for identifying and studying safety-relevance in trajectory datasets via safety-informed priors.

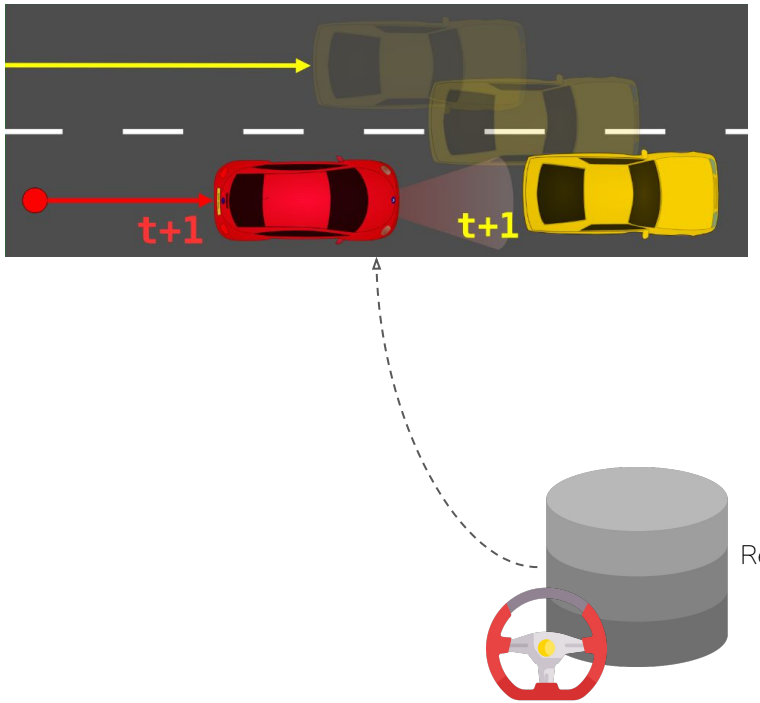


A Toy Example

Scenario: The yellow vehicle performs an aggressive lane change in front of the red vehicle.

Real Outcome: The proactive red vehicle anticipates and slows down to allow the lane change.

*→How could've this scenario **gone wrong**?*

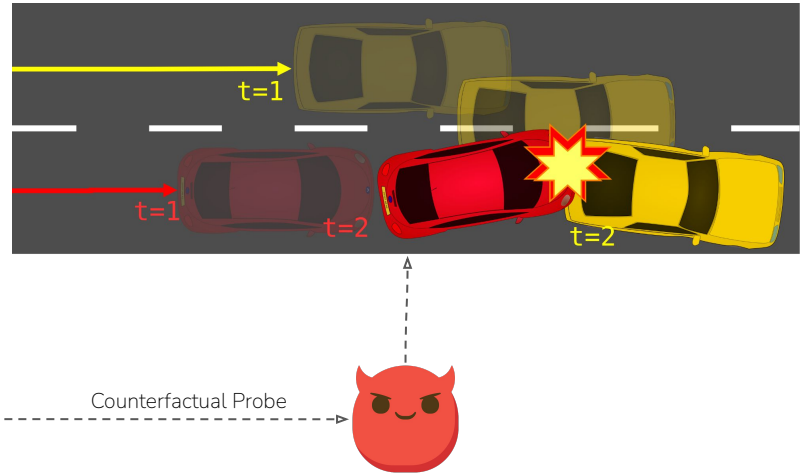


A Toy Example

Scenario: The yellow vehicle performs an aggressive lane change in front of the red vehicle.

~~**Real Outcome:** The proactive red vehicle anticipates and slows down to allow the lane change.~~

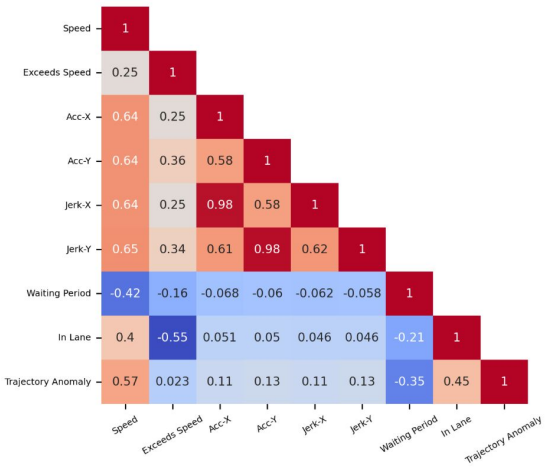
→ *What if... the red vehicle is distracted, doesn't anticipate the lane change?*



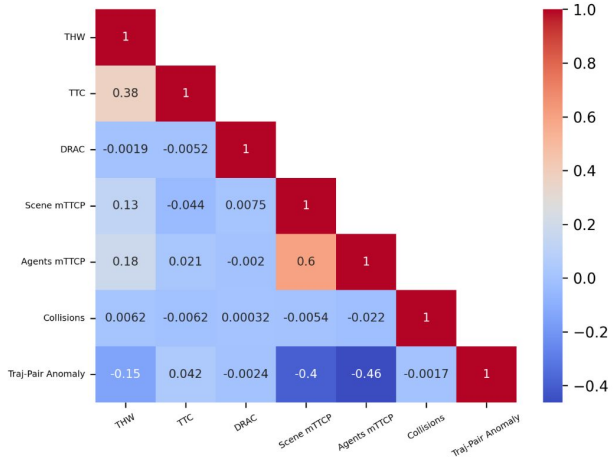
The score of the counterfactual probe incurred a high cost, thus, the scenario is safety-relevant.

Scenario Features

Correlation matrices for individual and interaction features.



(a) Feature Correlation for Individual State Features



(b) Feature Correlation for Interaction State Features

Scenario Scores

The score combines per-agent individual and social features:

$$\texttt{IndScore}^{(i)} = \mathbf{W}_{ind} \cdot \left[\max_t(v_t^{(i)}) \mid v \in \mathbf{V}_{ind} \right]$$

$$\texttt{SocScore}^{(i,j)} = \mathbf{W}_{soc} \cdot \left[\max_t(v_t^{(i,j)}) \mid v \in \mathbf{V}_{soc} \right]$$

$$\texttt{TrajScore}^{(i)} = \texttt{IndScore}^{(i)} + \sum_{j \neq i} \texttt{SocScore}^{(i,j)} \quad \text{(For all agents)}$$

SceneScore

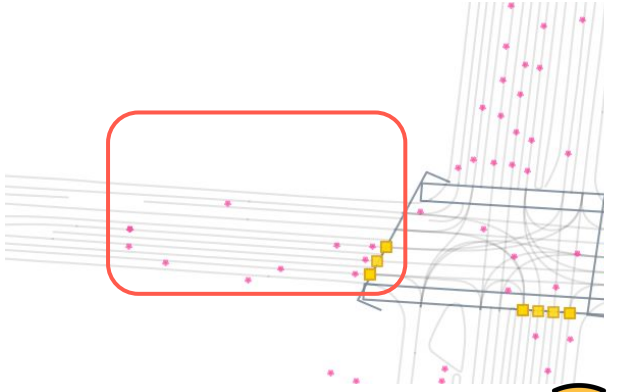
Table 6.2: Trajectory scoring variations.

Variation	IndScore	SocScore
Ground Truth (GT)	$\mathbf{X}_{GT}^{(i)}$	$(\mathbf{X}_{GT}^{(i)}, \mathbf{X}_{GT}^{(j)})$
Future Extrapolated (FE)	$\mathbf{X}_{FE}^{(i)}$	$(\mathbf{X}_{FE}^{(i)}, \mathbf{X}_{FE}^{(j)})$
Asymmetric (AS)	$\mathbf{X}_{FE}^{(i)}$	$(\mathbf{X}_{FE}^{(i)}, \mathbf{X}_{GT}^{(j)})$
Combined (CO)	$\max(\texttt{TrajScore}_{\texttt{GT}}, \texttt{TrajScore}_{\texttt{FE}})$	
Asymmetric Combined (AC)	$\max(\texttt{TrajScore}_{\texttt{GT}}, \texttt{TrajScore}_{\texttt{AS}})$	

A Real Example

Original scenario:

Black vehicle sees the standing vehicle ahead and **slows down** to avoid collision.



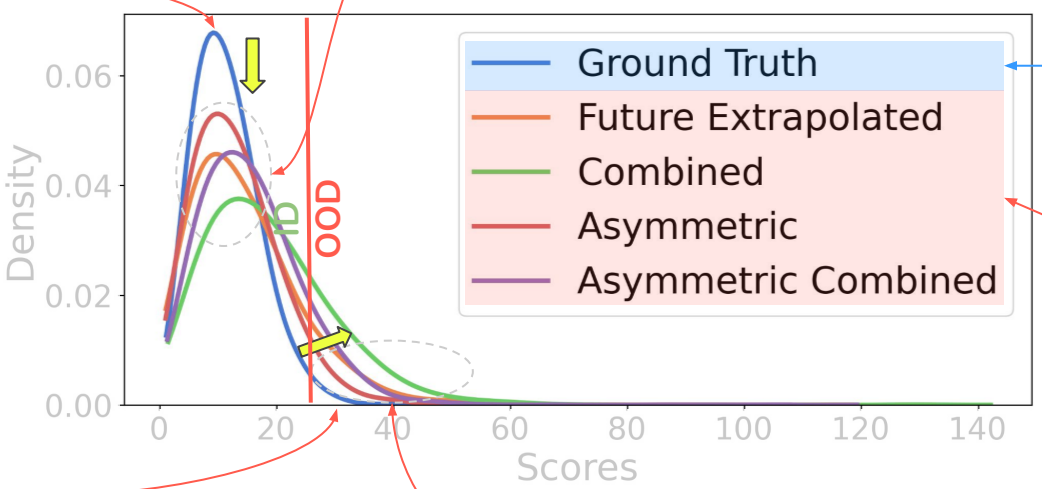
Counterfactual scenario:

Distracted black vehicle does not see the standing vehicle**, keeps going** until it collides.

Note: Scenario obtained from a subset of the Waymo Open Motion Dataset (WOMD)

Expanding the Long-tail

Density function across scores for different scoring strategies.



SafeShift wants to pull down the nominal mode

Ground Truth: High density mode representing low-scored (nominal) scenarios.

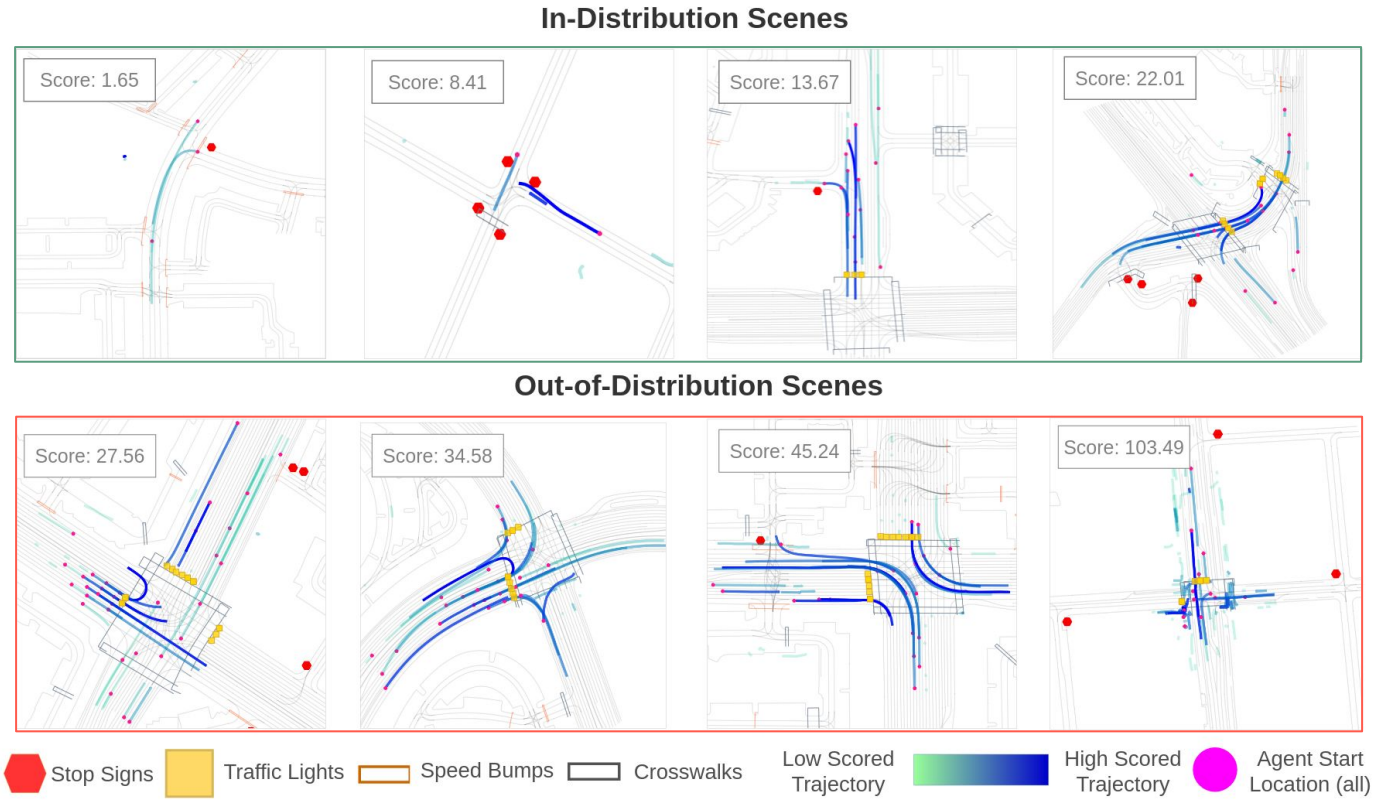
Ground Truth: Low density long-tail representing high-scored scenarios

SafeShift

SafeShift +Counterfactual Probing (i.e., mining for safety-relevant scenarios)

...to push up the long-tail density and obtain an out-of-distribution set.

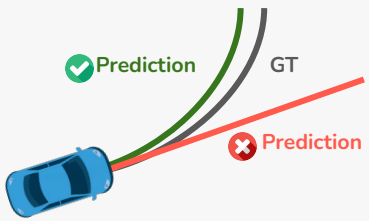
In-Distribution vs Out-of-Distribution Results



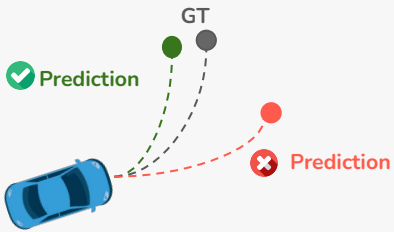
Metrics

We analyze widely accepted trajectory prediction metrics in autonomous driving:

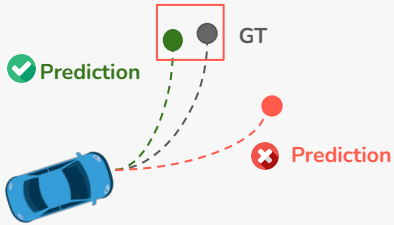
Average Displacement Error (ADE)



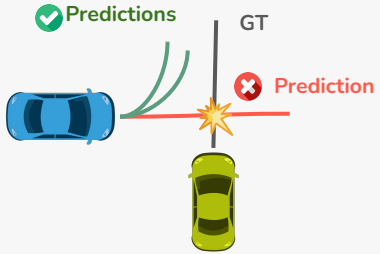
Final Displacement Error (FDE)



Mean Average Precision (mAP)



Collision Rate (CR)



In-Distribution vs Out-of-Distribution Results

Want to observe an **increased GT** crash rate in the **test set (ODD)** w.r.t. **GT val set (ID)**

TABLE II: Distribution shift experiments in WOMD [5]. ADE / FDE is in meters. Δ_{val} is the change in test collision rate (CR) from the corresponding val CR. A more drastic **increase** is better.

Data Split	Method	Validation Set (In-Distribution) ADE / FDE	mAP	CR	Testing Set (Out-of-Distribution) ADE / FDE	mAP	CR (Δ_{val})
Uniform	GT	- / -	-	0.008	- / -	-	0.009 (+12.5%)
	MTR [27]	0.73 / 1.58	0.30	0.062	0.73 / 1.59	0.31	0.061 (−1.60%)
	A-VRNN [24]	1.80 / 4.63	0.06	0.057	1.82 / 4.67	0.06	0.058 (+1.80%)
Clusters [18]	GT	- / -	-	0.009	- / -	-	0.007 (−22.2%)
	MTR	0.69 / 1.50	0.35	0.060	0.71 / 1.55	0.33	0.051 (−15.0%)
	A-VRNN	1.79 / 4.59	0.08	0.062	1.82 / 4.70	0.07	0.049 (−21.0%)
Scoring (Ours)	GT	- / -	-	0.005	- / -	-	**0.017 (+240%)**
	MTR	0.72 / 1.59	0.32	0.044	0.74 / 1.59	0.30	**0.100 (+127%)**
	A-VRNN	1.99 / 5.26	0.05	0.042	2.13 / 5.55	0.05	**0.099 (+136%)**

GT: Ground truth tracks

~280 scenarios

~315 scenarios

~245 scenarios

~595 scenarios

Note: experiments done on a subset of ~170k (20%) scenarios from WOMD (~135k for train/val and ~35k for test).

Unremediated Trajectory Prediction

Want to observe models

How can we **remediate** these models to mitigate the collision rates incurred and improve model **generalizability** under OOD conditions?

onding val set (ID)

TABLE II: Distribution shift experiments in WOMD [5]. ADE / FDE is in meters. Δ_{val} is the change in test collision rate (CR) from the corresponding val CR. A more drastic **increase** is better.

Data Split	Method	Validation Set (In-Distribution) ADE / FDE	mAP	CR	Testing Set (Out-of-Distribution) ADE / FDE	mAP	CR (Δ_{val})
Uniform	GT	- / -	-	0.008	- / -	-	0.009 (+12.5%)
	MTR [27]	0.73 / 1.58	0.30	0.062	0.73 / 1.59	0.31	0.061 (−1.60%)
	A-VRNN [24]	1.80 / 4.63	0.06	0.057	1.82 / 4.67	0.06	0.058 (+1.80%)
Clusters [18]	GT	- / -	-	0.009	- / -	-	0.007 (−22.2%)
	MTR	0.69 / 1.50	0.35	0.060	0.71 / 1.55	0.33	0.051 (−15.0%)
	A-VRNN	1.79 / 4.59	0.08	0.062	1.82 / 4.70	0.07	0.049 (−21.0%)
Scoring (Ours)	GT	- / -	-	0.005	- / -	-	**0.017 (+240%)**
	MTR	0.72 / 1.59	0.32	0.044	0.74 / 1.59	0.30	**0.100 (+127%)**
	A-VRNN	1.99 / 5.26	0.05	0.042	2.13 / 5.55	0.05	**0.099 (+136%)**

GT: Ground truth tracks

~2082 scenarios

~1750 scenarios

~3482 scenarios

Note: experiments done on a subset of ~170k (20%) scenarios from *WOMD*.

Generalizable Trajectory Prediction under Out-of-Distribution Conditions

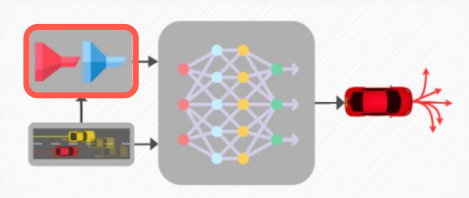
Difficulty-based sample weighting

To encourage the model to not treat all scenarios equally and care about safety-relevant situations.

$$\texttt{WeightedLoss}^{(i)} = \frac{1}{N}\sum_{i}^{N} \texttt{Loss}^{(i)} \cdot \texttt{Score}_{AC}^{(i)}$$

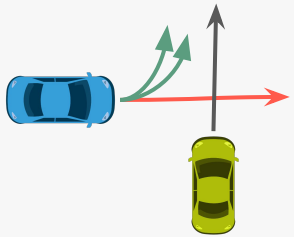
Counterfactual Biasing

To add counterfactual understanding to the model using the extrapolated futures.



Collision Loss

To add collision-awareness by *favour* predictions that do not collide with other agents' GT trajectories.



X. Zhou, O. Wu, W. Zhu, and Z. Liang, ***Understanding difficulty-based sample weighting with a universal difficulty measure*** in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2022, pp. 68–84.

Generalizable Trajectory Prediction under Out-of-Distribution Conditions

Want to **mitigate the models'** crash rates in the test set (ODD) w.r.t. the corresponding val set (ID)

TABLE III: Robust trajectory prediction experiments in WOMD [5]. ADE / FDE is in meters. Δ_{test} is the change in test CR from the *un-remediated* test CR for each method. A more drastic **decrease** is better.

Data Split	Method	Validation Set (In-Distribution) ADE / FDE	mAP	CR	Testing Set (Out-of-Distribution) ADE / FDE	mAP	CR (Δ_{test})
Scoring (Ours)	GT	- / -	-	0.005	- / -	-	0.017 (-)
	MTR	0.72 / 1.59	0.32	0.044	0.74 / 1.59	0.30	0.100 (-)
	MTR + F+ [18]	0.73 / 1.59	0.32	0.043	0.75 / 1.59	0.30	0.099 (−1.00%)
	MTR + Ours	0.83 / 1.80	0.25	0.037	0.89 / 1.91	0.22	**0.086** (**−14.0%**)
	A-VRNN	1.99 / 5.26	0.05	0.042	2.13 / 5.55	0.05	0.099 (-)
	A-VRNN + F+	2.05 / 5.24	0.06	0.041	2.23 / 5.73	0.06	0.103 (+4.04%)
	A-VRNN + Ours	1.76 / 4.61	0.06	0.039	1.91 / 4.94	0.06	**0.093** (**−6.06%**)

GT: Ground truth tracks, **F+**: Frenet+ Strategy [18]

~3482 scenarios

~3130 scenarios

Generalizable Trajectory Prediction under Out-of-Distribution Conditions - Ablations

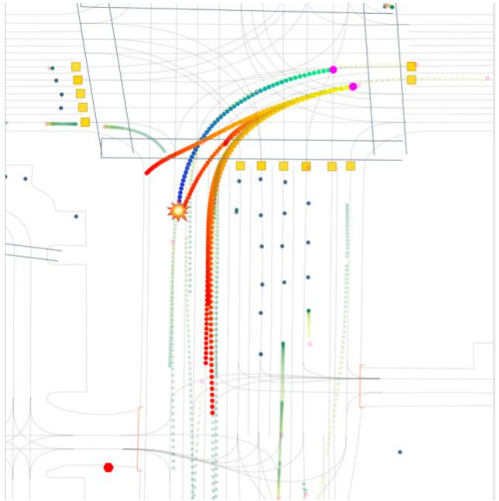
Table 6.6: Remediation strategy ablation study based on our proposed approach in Section 6.3.6 utilizing MTR [65] on WOMD [57]. ADE / FDE is in meters. Δ_{test} is the change in test CR from the *un-remediated* MTR test CR.

Ablation Name	Remediation		Validation Set (In-Distribution)			Testing Set (Out-of-Distribution)		
	SC	CL	ADE / FDE	mAP	CR	ADE / FDE	mAP	CR (Δ_{test})
MTR [65]	-	-	0.72 / 1.59	0.32	0.044	0.74 / 1.59	0.30	0.100 (-)
MTR + Ours (SC only)	✓	-	0.74 / 1.63	0.31	0.046	0.74 / 1.61	0.29	0.103 (+3.00%)
MTR + Ours (CL only)	-	✓	0.81 / 1.77	0.27	0.038	0.88 / 1.92	0.23	0.093 (−7.00%)
MTR + Ours (Full)	✓	✓	0.83 / 1.80	0.25	0.037	0.89 / 1.91	0.22	**0.086 (−14.0%)**

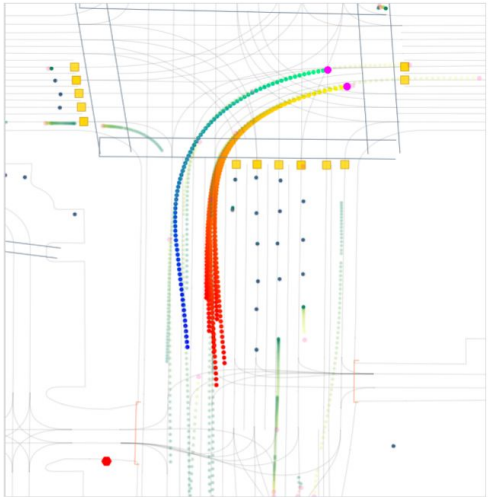
SC: Score incorporation, **CL**: Collision loss objective.

Generalizable Trajectory Prediction under Out-of-Distribution Conditions

Original



Safety-aware



SEAL

Towards Safe Autonomous Driving via **S**kills-**E**nabled **A**dversary **L**earning for Closed-Loop Scenario Generation

IEEE Robotics and Automation Letters, 2025

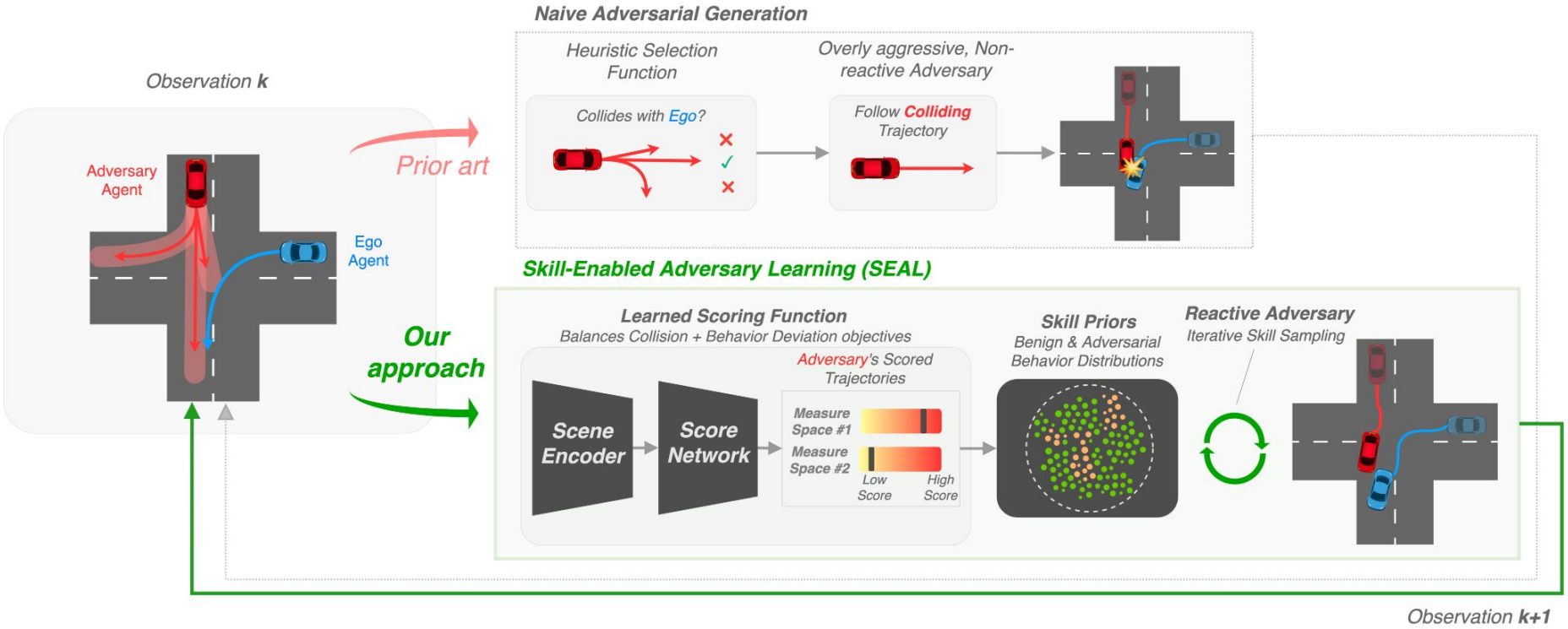
https://navars.xyz/seal

In the context of Trajectory Forecasting:

- Predictors used as a prior for candidate trajectory selection.
- SafeShift scenarios used as an evaluation fairness benchmark.
- Bonus: applied in industry!

STACK

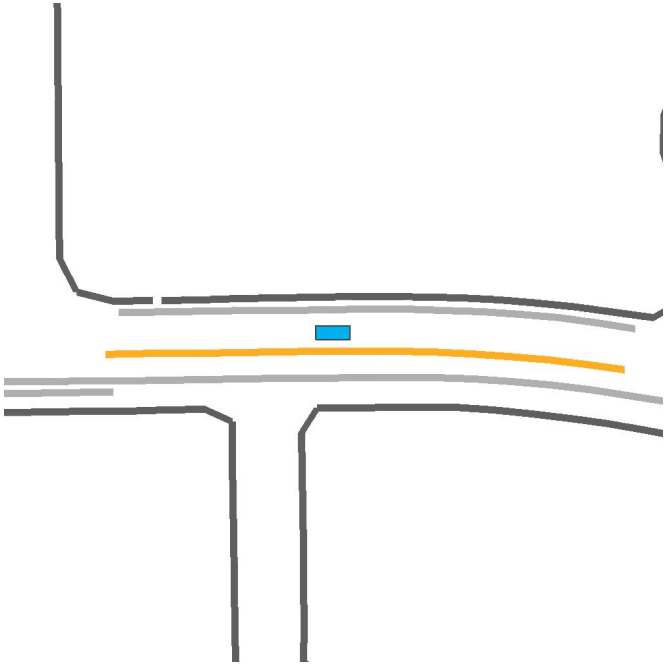
Skill-Enabled Adversary Learning for Scenario Generation



Scenario Realism

SOTA adversarial scenario-generation approaches often struggle to provide useful training stimuli to closed-loop agents:

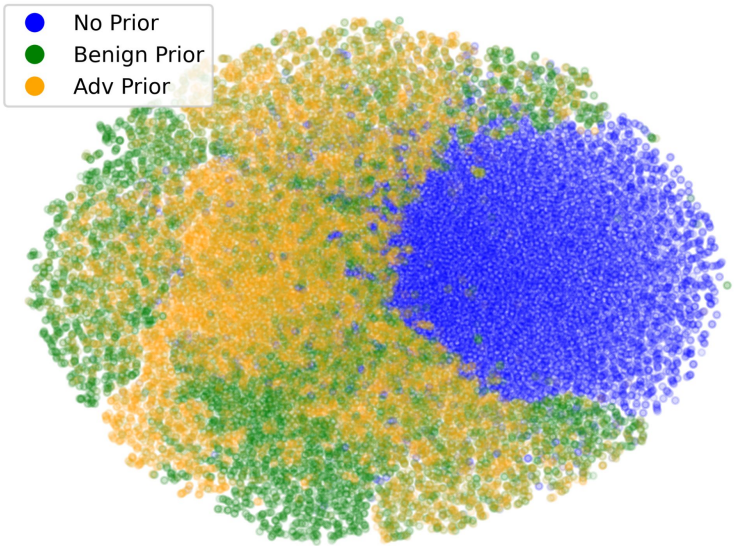
- Limited view of safety-criticality, often **only focused** on optimizing unrealistic and overly-aggressive adversarial behavior
- **Lacking reactivity** to an ego-agent's behavior diversity.



Idea: Skill-Enabled Adversary Learning (SEAL)

SEAL introduces two novel components:

- A learned objective function to anticipate how a reactive ego agent will respond to a candidate adversarial behavior.
- A reactive adversary policy that hierarchically selects human-like skill primitives to increase criticality and maintain realism.



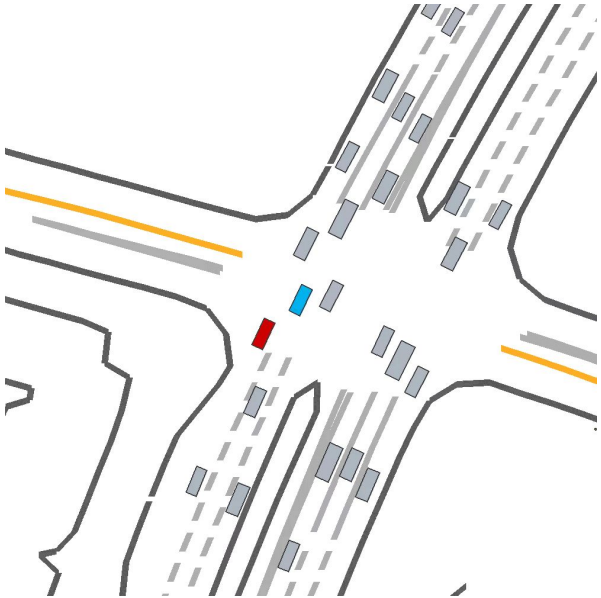
SEAL: Evaluation Fairness

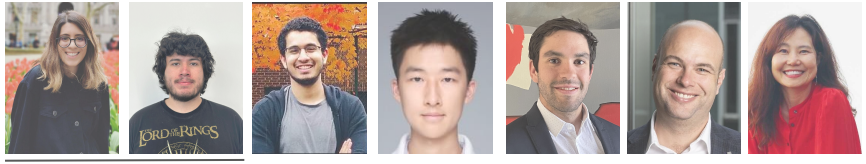
Safety-critical scenario generation approaches commonly evaluate safety-criticality *in-distribution*.

SEAL argues that performance on challenging scenes is ultimately more important.

Therefore, it leverages SafeShift to create a realistic out-of-distribution evaluation setting for scenario generation.

Scenario Generation using SafeShift “hard” scenes





Amelia

A Large Model and Dataset for Airport Surface Movement Forecasting

AIAA Aviation and Ascend Forum, 2024

https://ameliacmu.github.io

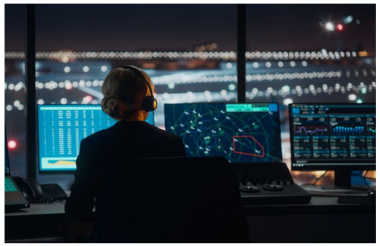
In the context of Trajectory Forecasting:

- A new domain for trajectory forecasting
- A generalizability benchmark
- SafeShift characterization used as a prior for generalizable scene representation
- Bonus: Best Paper Award!

Uptick in close calls at U.S. airports

Airline Close Calls Happen Far More Often Than Previously Known

By Sydney Ember and Emily Steel
Graphics by Leanne Abraham, Eleanor Lutz and Ella Koeze
Aug. 21, 2023



Corodenkoff / Shutterstock

What's Behind the Uptick in Close Calls at U.S. Airports?

An increase in near collisions at U.S. airports should be a "wake-up call," say government officials and aviation experts. What's causing the issue? And what's being done about it?

By Sean Cudahy
December 18, 2023

More Airport Close Calls: This Is Getting Serious

The air travel system is under constant pressure to move more people, faster. The strains are mounting up.

JAMES FALLOWS
APR 27, 2024 - PAID

40 31 Share

21 AUGUST 2023 / SF NEWS / JAY BARMANN

Close Calls on SFO Runways Are Just Tip of the Iceberg In Broader National Trend

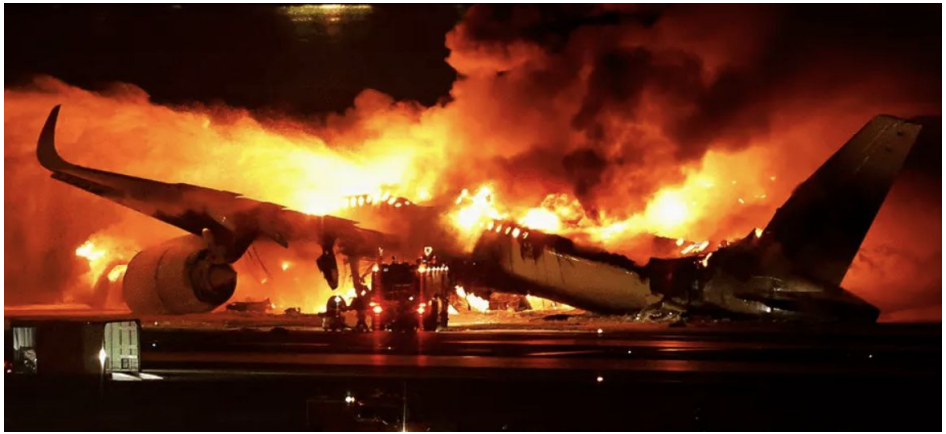
LOCAL

FAA, NTSB investigating several serious close calls amid record-breaking air travel

By Kirstin Garriss, CMG Washington News Bureau
July 05, 2024 at 5:55 pm EDT

A runway collision at Haneda Airport, Japan

January 2nd, 2024



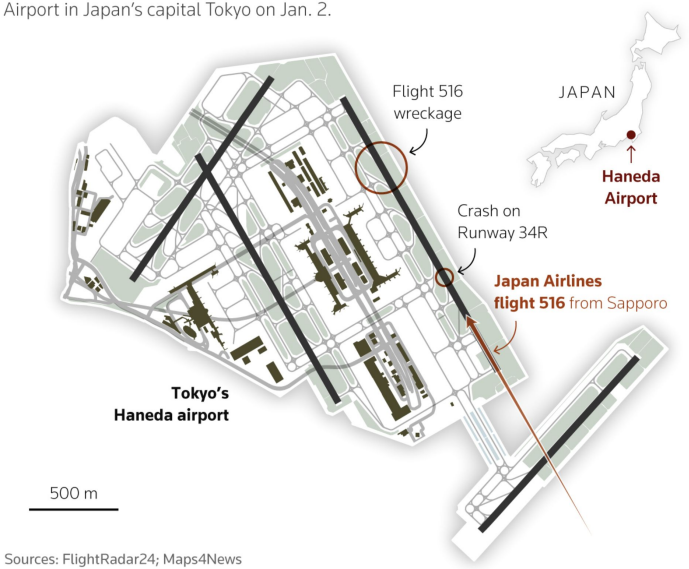
Japan Airlines counts losses from wrecked Tokyo plane

By **Daniel Leussink**, **Kaori Kaneko** and **Lisa Barrington**

January 4, 2024 1:12 PM GMT+5:30 · Updated 2 hours ago

Planes collide at Tokyo airport

Japan Airlines flight 516 collided with a Coast Guard aircraft while landing at Haneda Airport in Japan's capital Tokyo on Jan. 2.



Sources: FlightRadar24; Maps4News

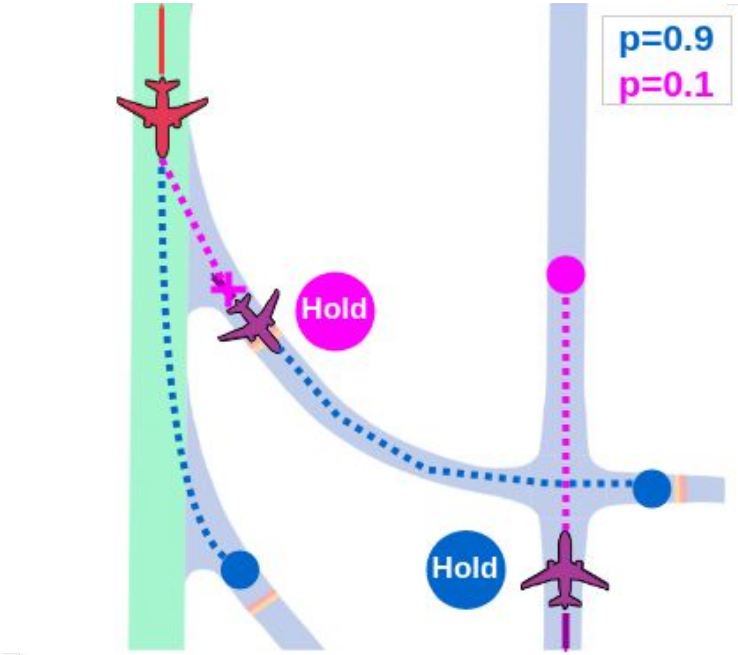
Vijdan Mohammad Kawoosa • Jan. 2, 2024 | REUTERS

Leussink, Daniel, Kaneko, Kaori and Barrington, Lisa. *Japan Airlines counts losses from wrecked Tokyo Plane*. Reuters, 2024: https://www.reuters.com/business/aerospace-defense/japan-airlines-estimates-loss-about-1048-mln-collision-2024-01-03/#:~:text=BIG%20LOSSES,financial%20year%20ending%20March%2031

A Framework for Airport Surface Movement Forecasting

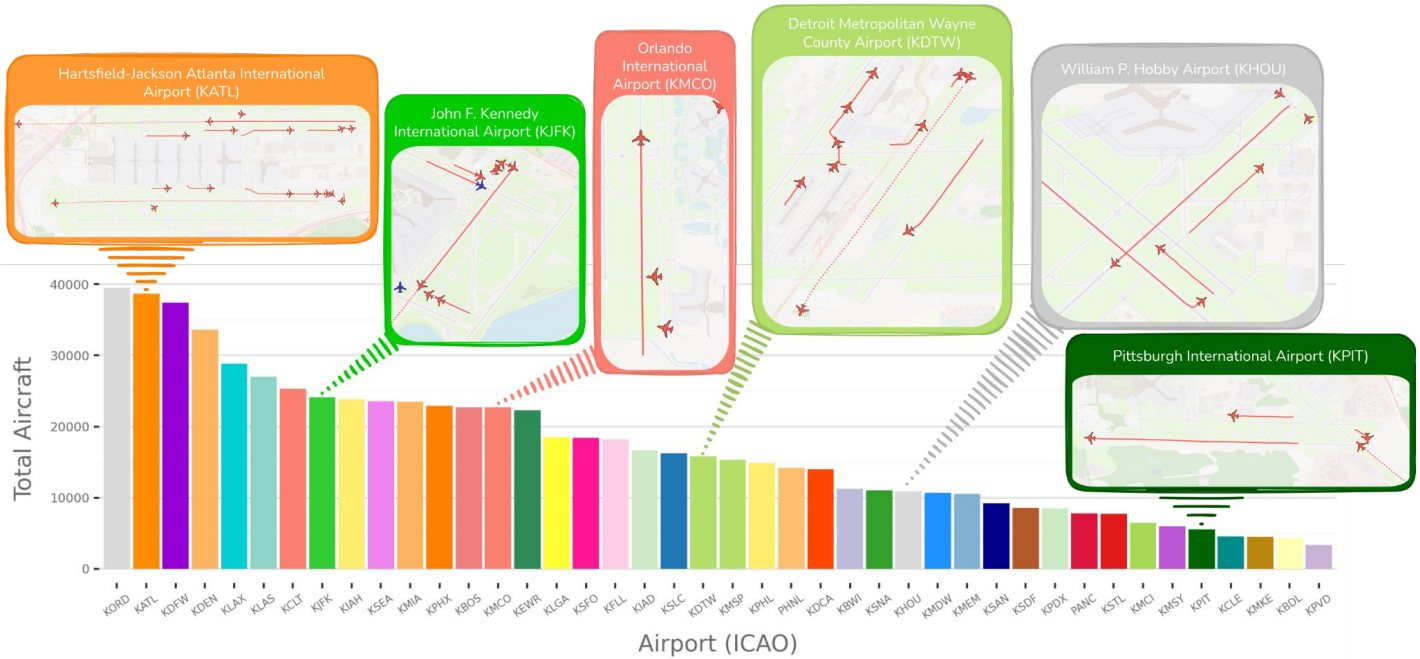
Amelia's objectives:

- Enable data-driven solutions for improving the safety and efficiency of airport operations.
- Enable scale and diversity for ML research in aviation.



Amelia-42: Enabling Dataset Diversity and Scale

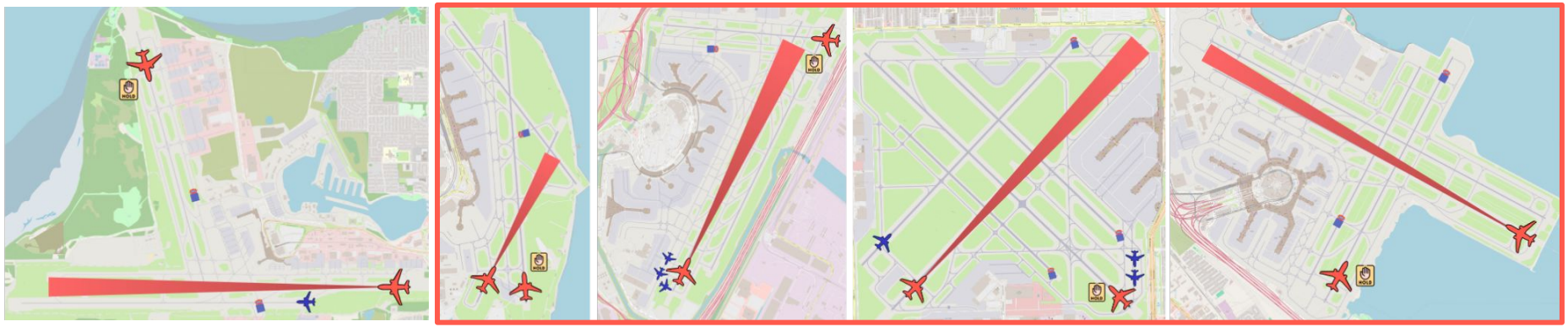
We collect and process a large dataset for airport surface movement covering 42 diverse U.S. airports.

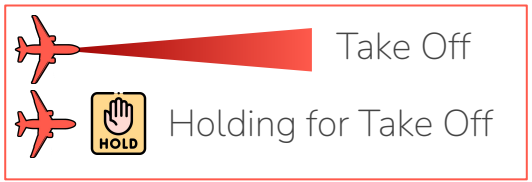


https://huggingface.co/AmeliaCMU/datasets

Amelia-10: Enabling Generalizable Behavior Forecasting

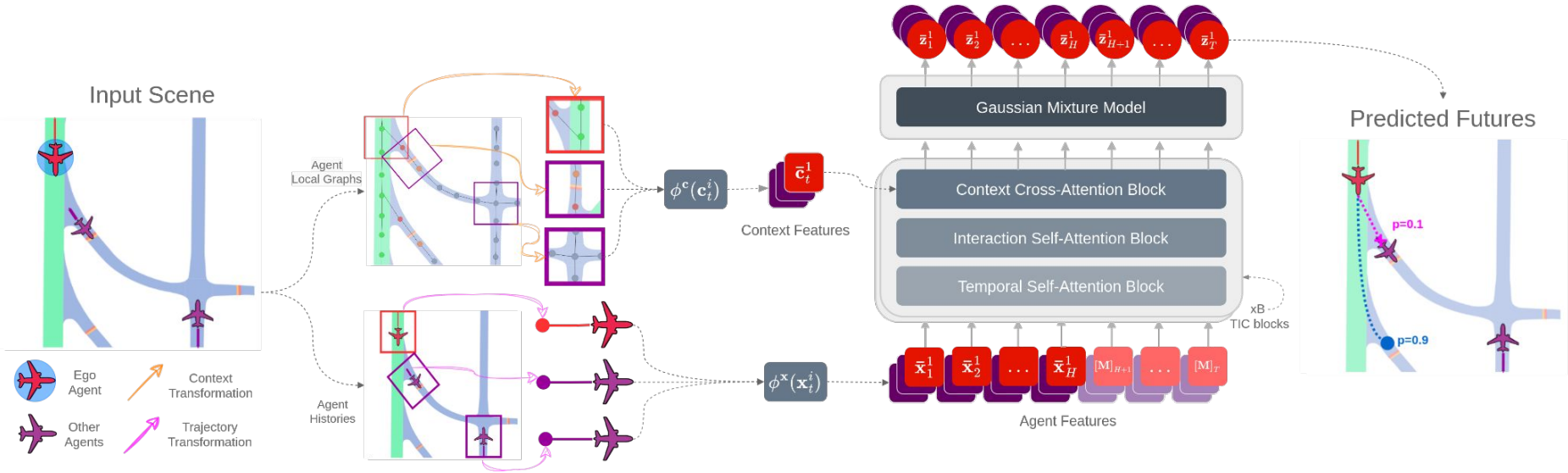
How can we design and evaluate models that generalize across a wide variety of contexts?



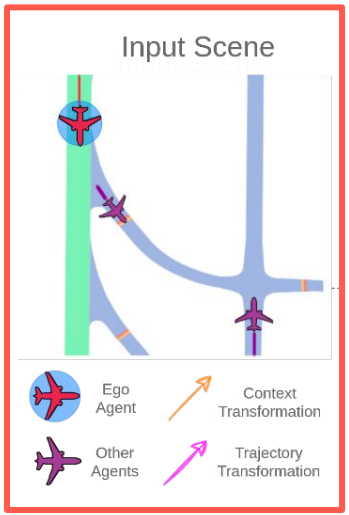


Amelia-TF: Trajectory Forecasting Model

A safety-informed airport surface movement trajectory forecasting baseline.



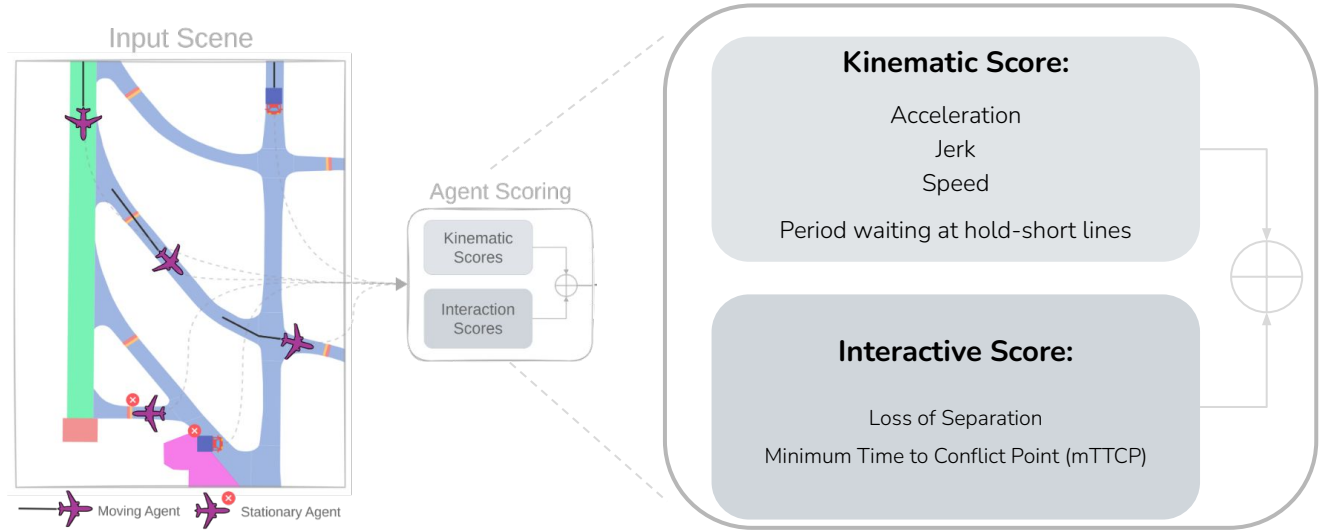
Amelia-TF: Safety-informed Scene Representation



- Want to generalize across a wide variety of contexts
  - **Idea:** Leverage safety-priors to encode the scene:
    - Characterize the degree to which an agent might affect others in the environment→ Safety-relevance.

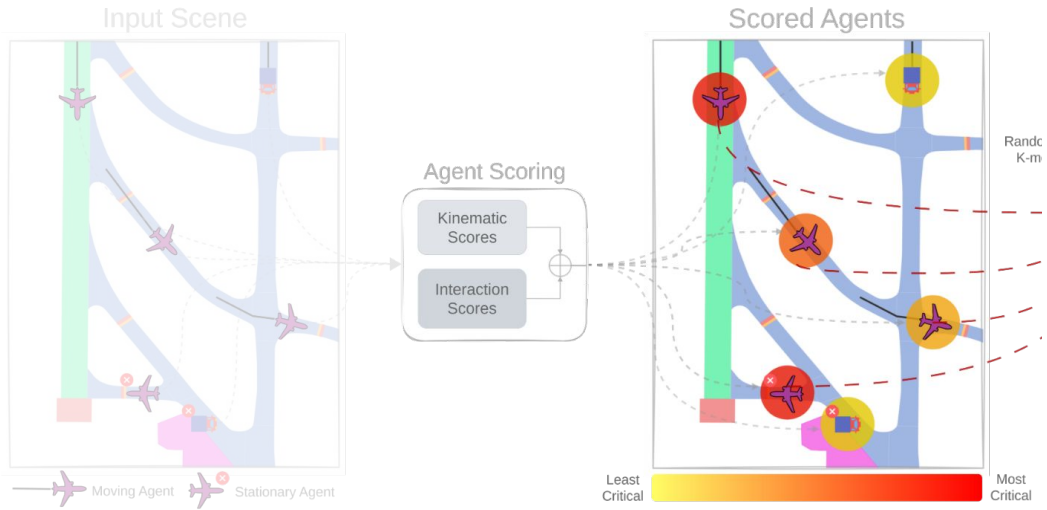
Safety-informed Scene Representation – An Example

To do so, we use an automated scenario characterization scheme to compute each agents kinematic and interactive states.



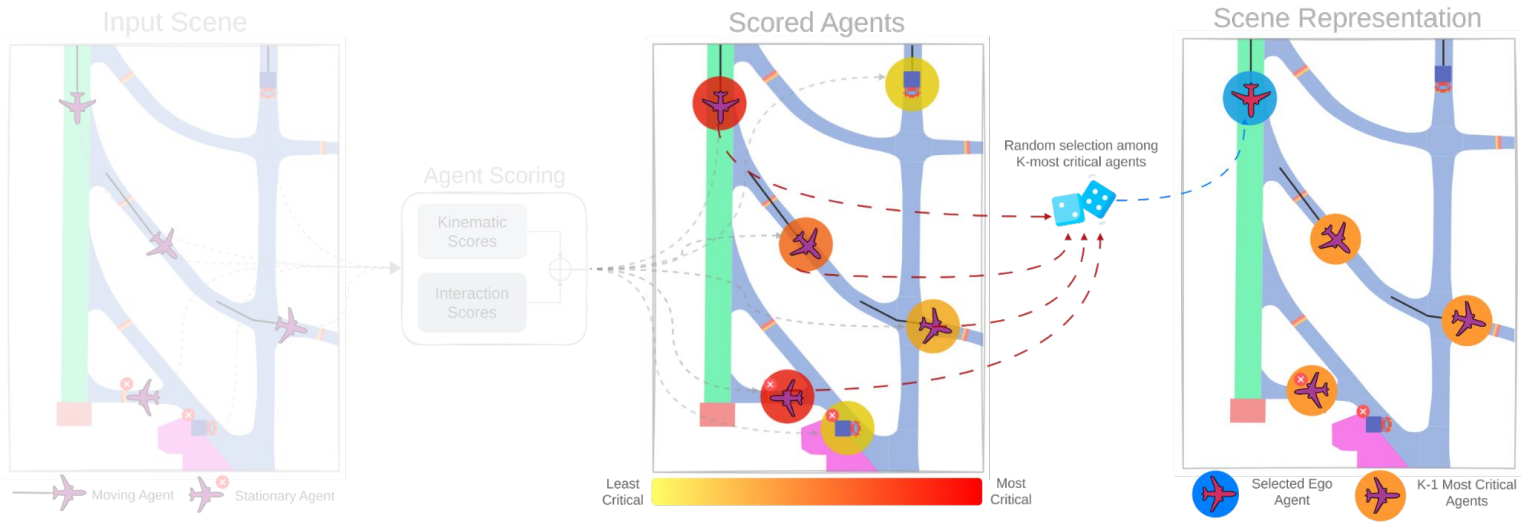
Safety-informed Scene Representation – An Example

We combine these features into a safety score for each agent, representing how relevant it is.



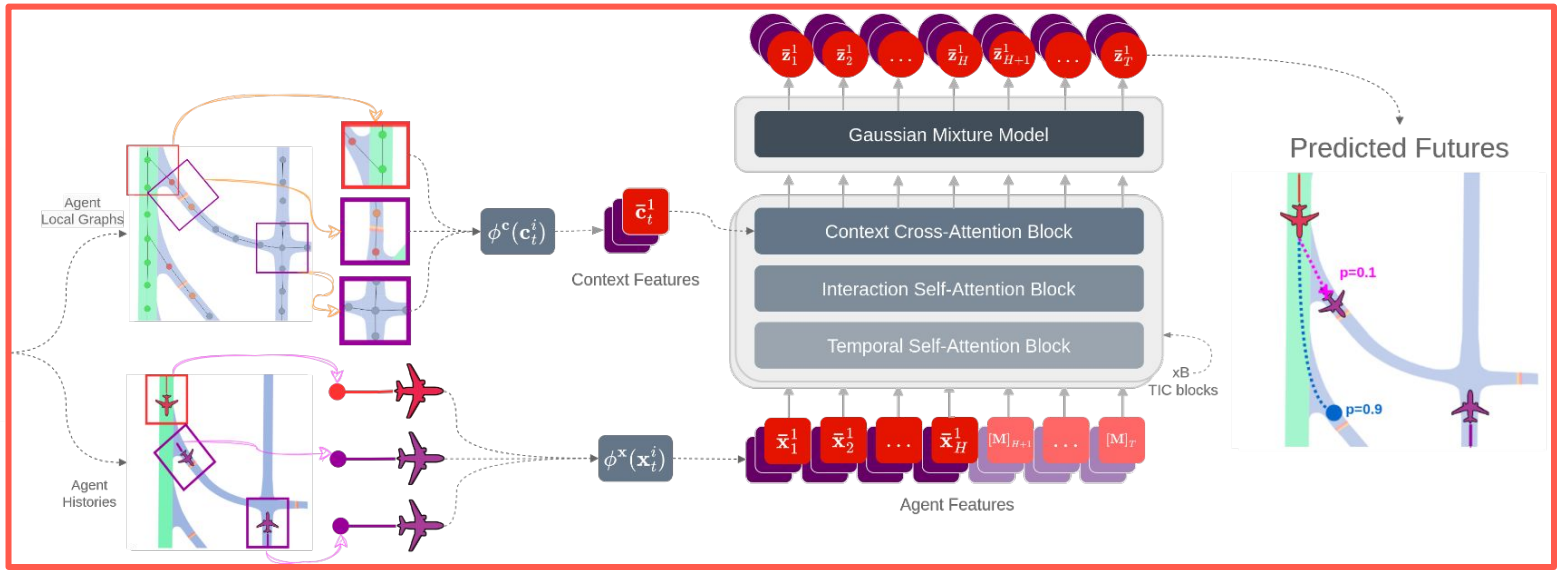
Safety-informed Scene Representation – An Example

We select the K-most relevant agents to represent the scene and an ego-agent within them



Amelia-TF: Scene Encoder and Trajectory Decoder

Local Transformation + Transformer Encoder + GMM Decoder



Ngiam, Jiquan, Benjamin Caine, Vijay Vasudevan, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs et al. *Scene transformer: A unified architecture for predicting multiple agent trajectories.* arXiv preprint arXiv:2106.08417 (2021).

Varadarajan, Balakrishnan, Ahmed Hefny, Avikalp Srivastava, Khaled S. Refaat, Nigamaa Nayakanti, Andre Cornman, Kan Chen et al. *Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. 2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022.

Baseline Trajectory Forecasting Results

Single-Airport; Specialist Models

Table 6 We show results for the *multi-airport* experiment for a prediction horizon of $F = 20$.

Experiment	KMDW mADE / mFDE	KEWR mADE / mFDE	KBOS mADE / mFDE	KSFO mADE / mFDE	KSEA mADE / mFDE	KDCA mADE / mFDE	PANC mADE / mFDE	KLAX mADE / mFDE	KJFK mADE / mFDE	KMSY mADE / mFDE	Average mADE / mFDE
Single-Airport	3.30 / 6.12	6.61 / 12.92	5.58 / 10.90	5.06 / 9.82	9.76 / 18.35	4.74 / 9.22	10.11 / 20.87	11.36 / 20.63	4.58 / 9.52	2.73 / 5.12	6.38 / 12.35
1-Seen	3.30 / 6.12	13.76 / 30.91	11.30 / 25.58	9.68 / 21.73	14.07 / 31.44	7.80 / 16.07	15.00 / 33.75	15.49 / 33.13	7.40 / 16.77	7.43 / 17.11	10.52 / 23.26
2-Seen	3.31 / 6.23	**6.92 / 13.45**	8.17 / 17.63	7.49 / 17.18	10.57 / 22.86	6.04 / 12.47	10.61 / 23.00	12.78 / 26.34	5.21 / 11.15	4.49 / 9.64	7.56 / 15.99
3-Seen	**3.26 / 6.59**	7.25 / 14.20	**6.05 / 12.11**	7.25 / 15.50	9.90 / 20.95	6.16 / 12.74	9.53 / 20.26	10.99 / 21.86	4.96 / 10.54	4.33 / 9.29	6.97 / 14.40
4-Seen	3.52 / 6.74	7.26 / 14.33	6.31 / 12.68	**5.66 / 11.33**	9.79 / 20.42	5.99 / 12.28	9.22 / 19.14	10.70 / 21.16	4.80 / 10.27	4.18 / 9.05	6.74 / 13.74
7-Seen	3.59 / 7.03	7.30 / 14.54	6.59 / 13.59	5.73 / 11.65	**8.30 / 16.71**	**4.55 / 8.82**	**7.48 / 14.94**	9.99 / 19.35	4.66 / 9.96	4.64 / 9.85	**6.28 / 12.64**
All-Seen	3.88 / 7.70	7.87 / 15.80	6.87 / 14.34	6.09 / 12.64	9.03 / 18.35	4.84 / 9.55	8.24 / 16.75	**8.80 / 16.73**	**4.56 / 9.64**	**3.22 / 6.25**	6.34 / 12.77

Each airport's ADE/FDE;
white cell → seen, gray cell → unseen

Average ADE/FDE
across airports

Multi-Airport;
Generalist

NOTE: Airport names are denoted with their *International Civil Aviation Organization (ICAO)* Code

SoRTS

Social **R**obot **T**ree **S**earch for Long-Horizon Navigation in Shared Airspace

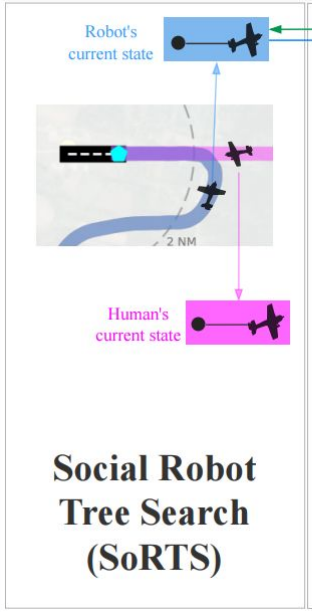
IEEE Robotics and Automation Letters, 2024

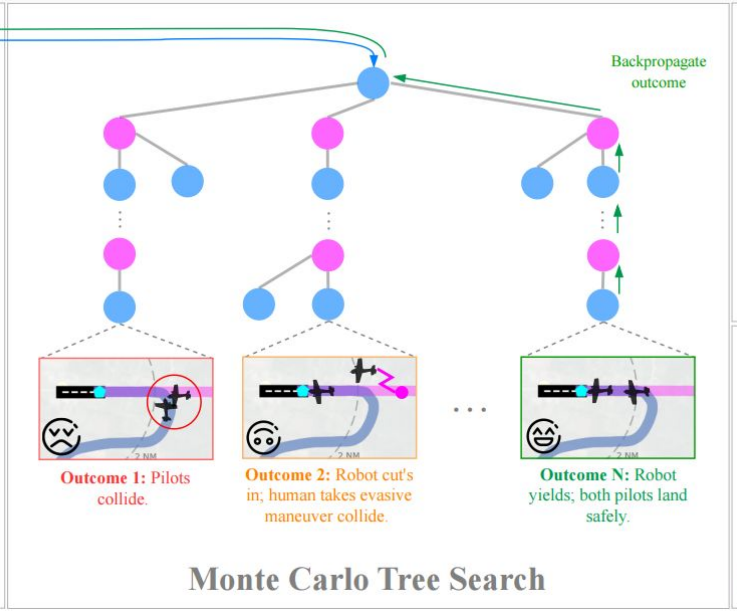
https://navars.xyz/sorts

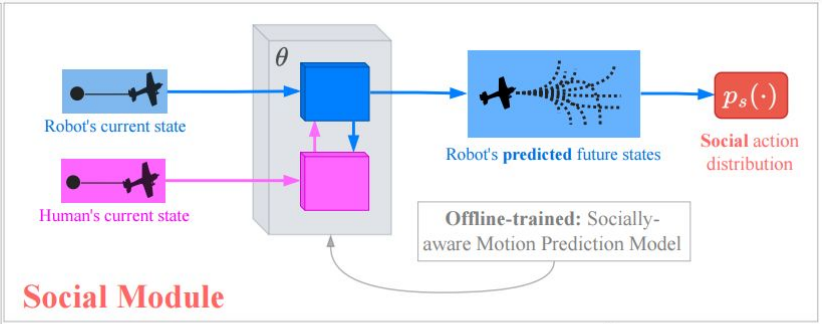
In the context of Trajectory Forecasting:

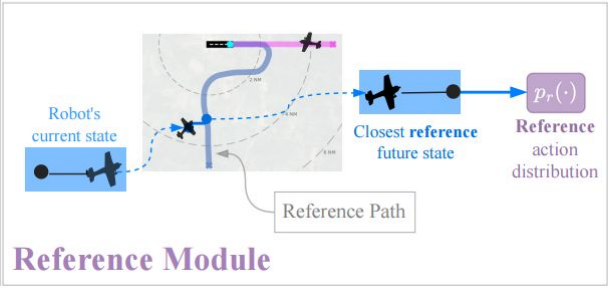
- Deployment of pre-trained models

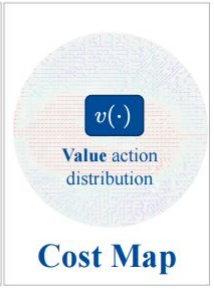
Social Robot Tree Search (SoRTS)







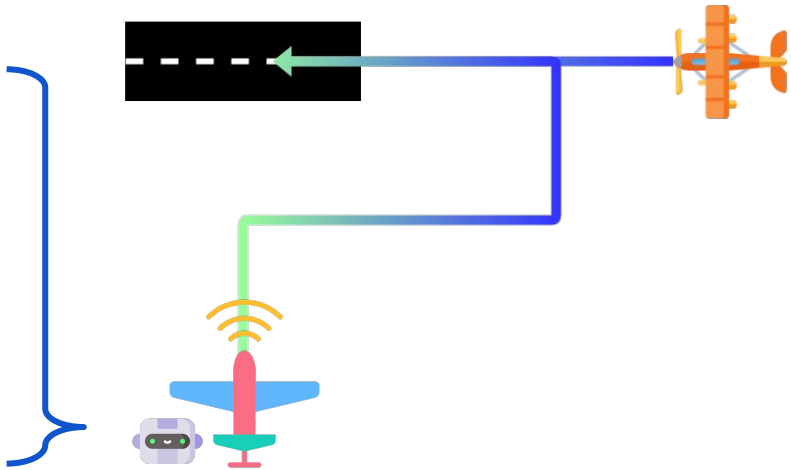




Social Robot Navigation in Shared Airspace

Desired Capabilities:

- Follow navigation norms
- Reason over human behavior and social cues
- Reason over multiple outcomes in the long-term future.

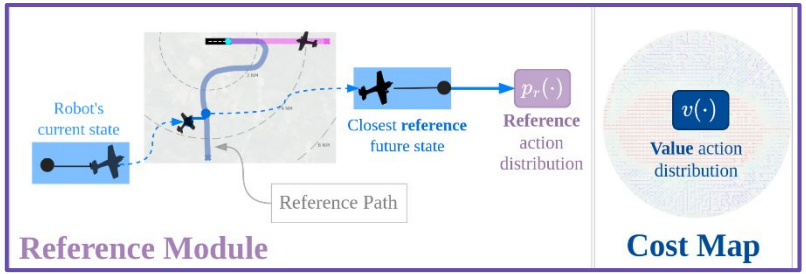


Follow Navigation Norms

Solely focusing on following navigation guidelines may overlook social interactions.

We also provide an agent with a cost map to bias the agent toward more desirable areas.

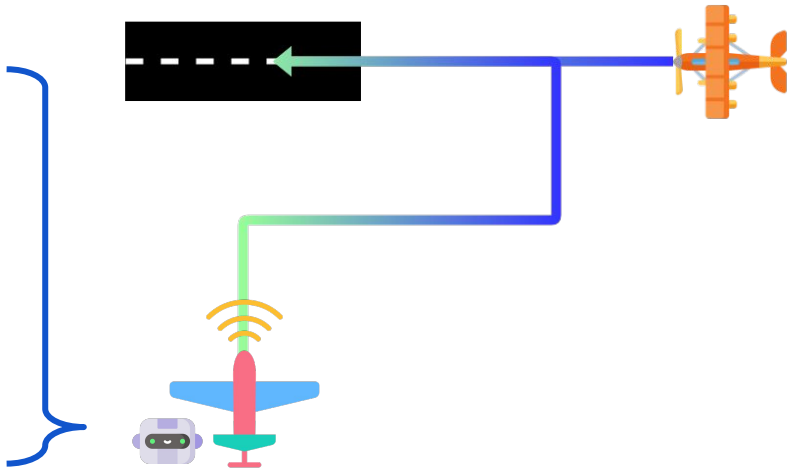
We design a reference module to provide agents with a navigation guideline to follow (e.g., a *flying pattern*) given a start location.



Social Robot Navigation in Shared Airspace

Desired Capabilities:

- Follow navigation norms
- Reason over human behavior and social cues
- Reason over multiple outcomes in the long-term future.

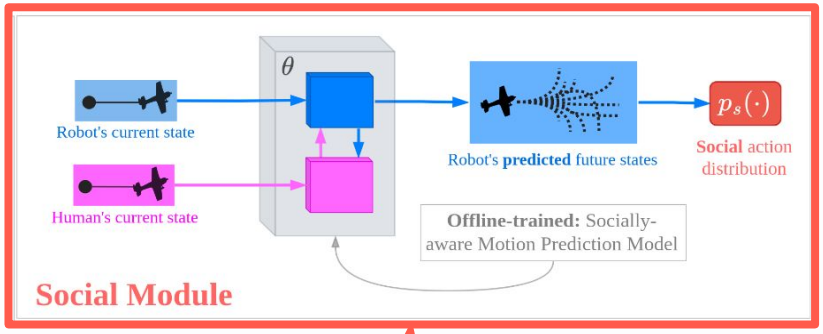


Reasoning over Social Cues

How to balance the ***reference*** vs. ***social*** action distributions?

We design a social module to handle short-horizon social dynamics.

To do so, it leverages a socially-aware trajectory prediction model [21], trained offline on the *TrajAir* [20] dataset.



[20] Patrikar, Jay, et al. "Predicting like a pilot: Dataset and method to predict socially-aware aircraft trajectories in non-towered terminal airspace." 2022 International Conference on Robotics and Automation (ICRA). IEEE, 2022.
[21] I. Navarro and J. Oh, "Social-PatteRNN: Socially-Aware Trajectory Prediction Guided by Motion Patterns," 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 2022.

An Initial Planner

Selects *optimal* next action by simply weighing the reference and the social action distributions for the **current step**,

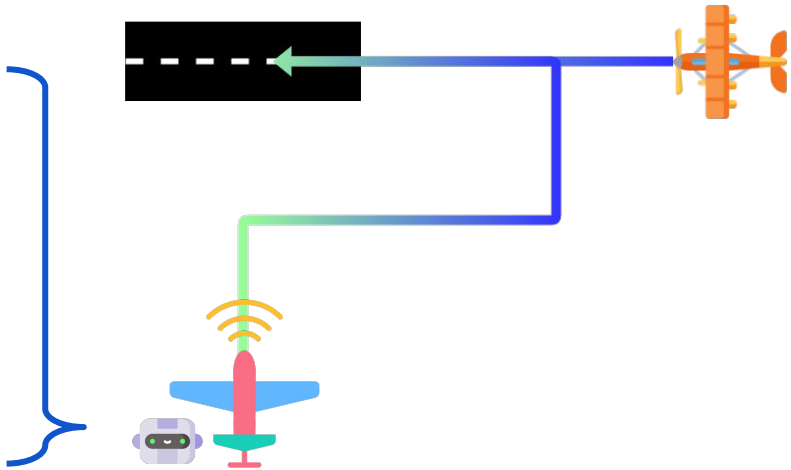
$$\mathbf{a}^* = \arg\max_{\mathbf{a}\in\mathcal{A}} \left[\lambda \cdot p_r(\mathbf{s}_t, \mathbf{a}) + (1-\lambda) \cdot p_s(\mathbf{s}_t, \mathbf{a})\right]$$

However, *short-sighted* reasoning might lead to suboptimal results

Social Robot Navigation in Shared Airspace

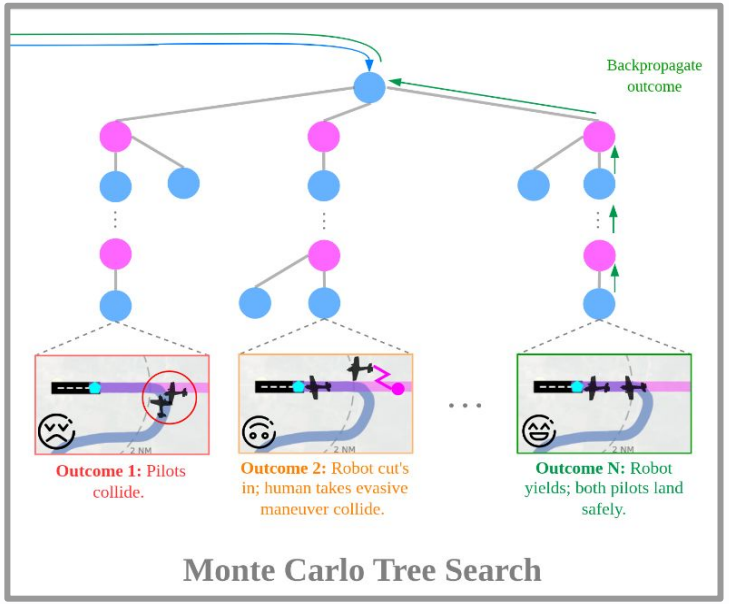
Desired Capabilities:

- Follow navigation norms
- Reason over human behavior and social cues
- Reason over multiple outcomes in the long-term future.



A Long-horizon Planner

We leverage Monte Carlo Tree Search (MCTS) to reason over multiple long-term outcomes.



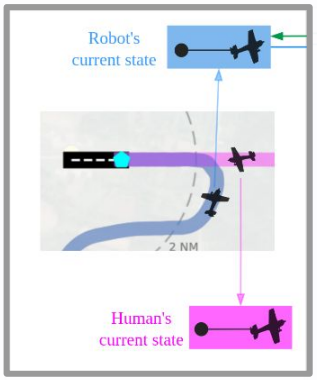
Uses a tree policy that's biased by reference and the social modules.

U = R + Q + S

*Upper Confidence Bound [17]

[17] Kocsis, Levente, and Csaba Szepesvári. "Bandit based monte-carlo planning." European conference on machine learning. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

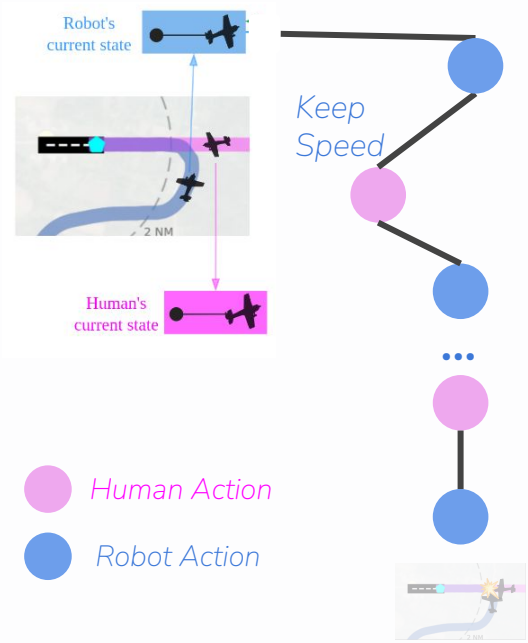
A Toy Example



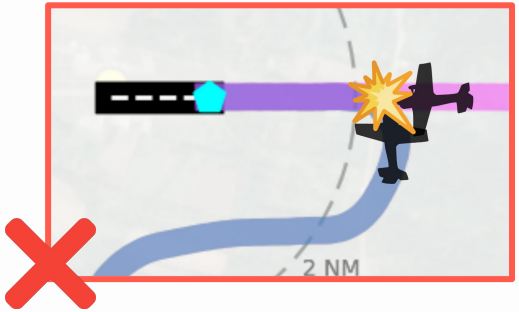
A robot and a human pilot want to land on the same runway, and have to coordinate to do so.

- The robot has multiple decision modalities to choose from in order to complete the task (e.g, slow-down, speed-up, etc).
- The robot has a reference to follow.
- The robot is able to characterize social interactions.

Exploring Possible Outcomes

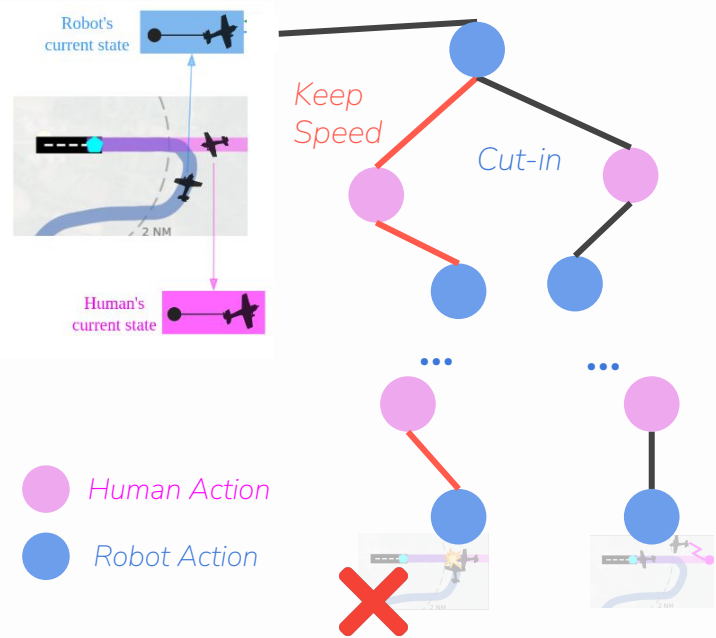


The robot keeps speed and prioritizes the reference.

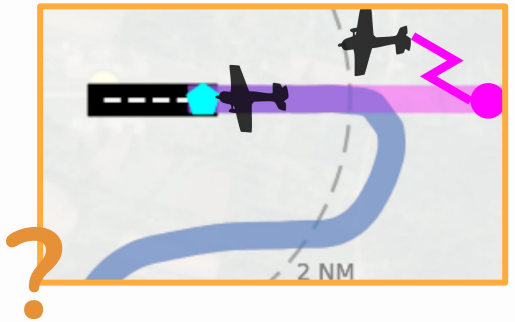


Outcome 1: Collision!

Exploring Possible Outcomes

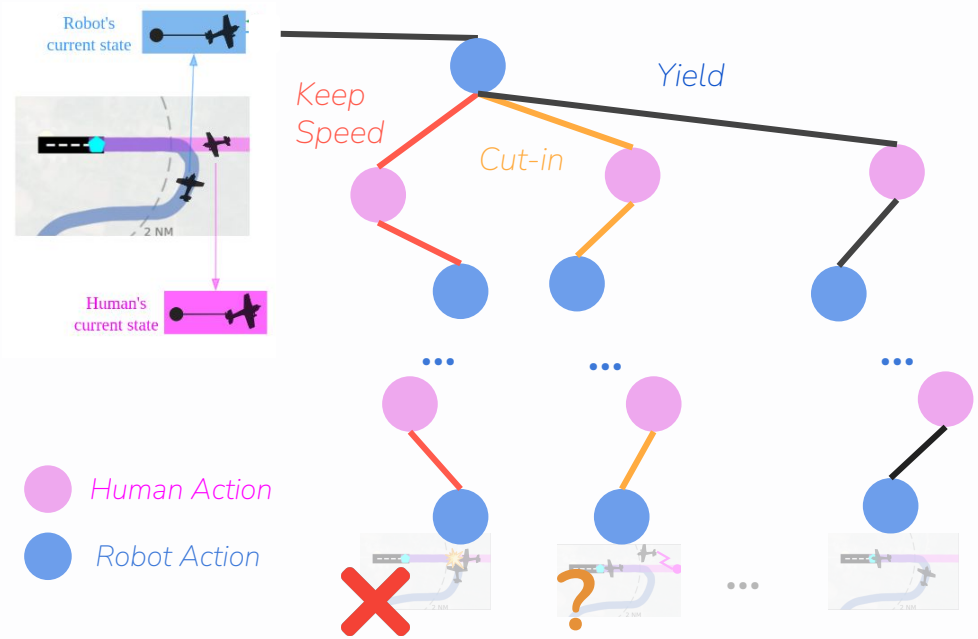


The robot aggressively cuts-in.

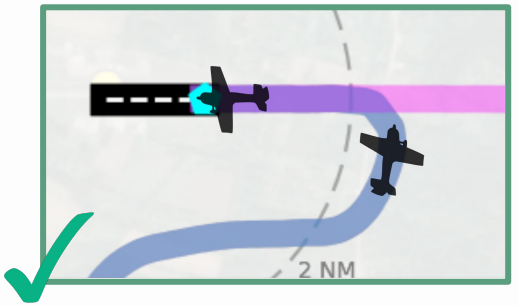


Outcome 2: Human takes an evasive maneuver to avoid a collision.

Exploring Possible Outcomes

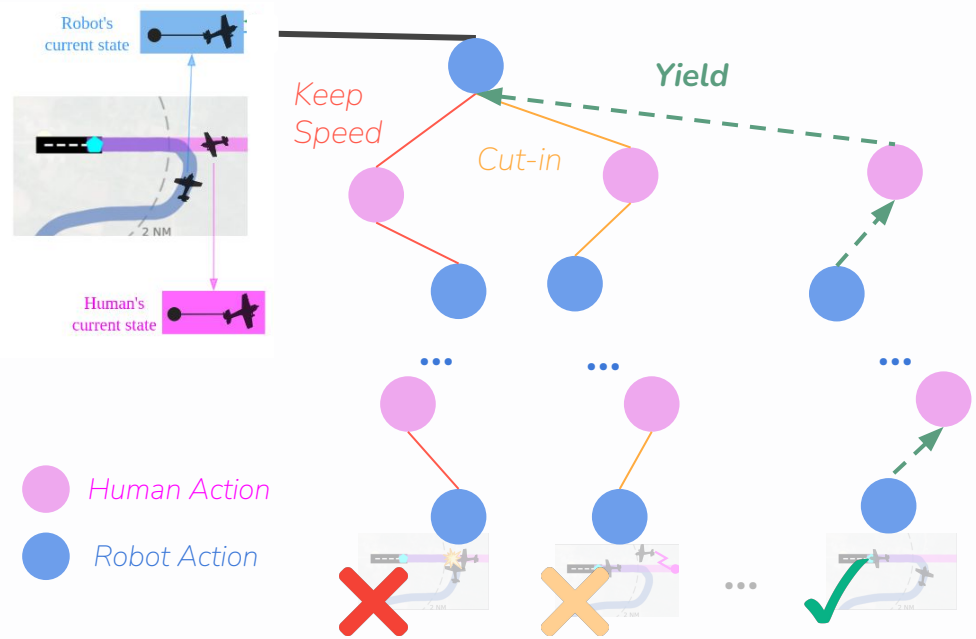


The robot yields to the human.



Outcome N: Both the human and the robot land safely.

Exploring Possible Outcomes



Backpropagate Best Outcome

Recommended Papers: Fundamentals, Datasets, Models

Surveys and Fundamentals:

- Wang, L., Lavoie, M. A., Papais, S., Nisar, B., Chen, Y., Ding, W., ... & Waslander, S. (2025). ***Deployable and Generalizable Motion Prediction: Taxonomy, Open Challenges and Future Directions***. arXiv preprint arXiv:2505.09074.
- Rudenko, A., Palmieri, L., Herman, M., Kitani, K. M., Gavrila, D. M., & Arras, K. O. (2020). ***Human motion trajectory prediction: A survey***. The International Journal of Robotics Research, 39(8), 895-935.

Models:

- Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). ***Social lstm: Human trajectory prediction in crowded spaces.*** In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 961-971).
- Yuan, Y., Weng, X., Ou, Y., & Kitani, K. M. (2021). *Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting*. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9813-9823).
- Ngiam, J., Caine, B., Vasudevan, V., Zhang, Z., Chiang, H. T. L., Ling, J., ... & Shlens, J. (2021). ***Scene transformer: A unified architecture for predicting multiple agent trajectories.*** arXiv preprint arXiv:2106.08417.
- Gao, J., Sun, C., Zhao, H., Shen, Y., Anguelov, D., Li, C., & Schmid, C. (2020). ***Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*** (pp. 11525-11533).
- Shi, S., Jiang, L., Dai, D., & Schiele, B. (2022). ***Motion transformer with global intention localization and local movement refinement. Advances in Neural Information Processing Systems***, 35, 6531-6543.
- Nayakanti, N., Al-Rfou, R., Zhou, A., Goel, K., Refaat, K. S., & Sapp, B. (2022). ***Wayformer: Motion forecasting via simple & efficient attention networks***. arXiv preprint arXiv:2207.05844.
- Feng, L., Bahari, M., Amor, K. M. B., Zablocki, É., Cord, M., & Alahi, A. (2024, September). ***Unitraj: A unified framework for scalable vehicle trajectory prediction.*** In European Conference on Computer Vision (pp. 106-123). Cham: Springer Nature Switzerland.

Datasets:

- Chang, M. F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., ... & Hays, J. (2019). ***Argoverse: 3d tracking and forecasting with rich maps.*** In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8748-8757).
- Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., ... & Anguelov, D. (2021). ***Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset.*** In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9710-9719).
- Navarro, I., Ortega, P., Patrikar, J., Wang, H., Ye, Z., Park, J. H., ... & Scherer, S. (2024). ***AmeliaTF: A Large Model and Dataset for Airport Surface Movement Forecasting.*** In AIAA AVIATION FORUM AND ASCEND 2024 (p. 4251).

Evaluation:

- Ivanovic, B., & Pavone, M. (2021). ***Rethinking trajectory forecasting evaluation.*** arXiv preprint arXiv:2107.10297.
- Li, L., Lin, X., Huang, Y., Zhang, Z., & Hu, J. F. (2024). ***Beyond minimum-of-N: Rethinking the evaluation and methods of pedestrian trajectory prediction.*** IEEE Transactions on Circuits and Systems for Video Technology.
- Weng, E., Hoshino, H., Ramanan, D., & Kitani, K. (2023). ***Joint metrics matter: A better standard for trajectory forecasting.*** In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 20315-20326).

Recommended Papers: Related / Frontier Topics

Scenario Generation, Re-simulation and Closed-loop Evaluation

- Suo, S., Wong, K., Xu, J., Tu, J., Cui, A., Casas, S., & Urtasun, R. (2023). ***Mixsim: A hierarchical framework for mixed reality traffic simulation***. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9622-9631).
- Stoler, B., Navarro, I., Francis, J., & Oh, J. (2025). ***SEAL: Towards safe autonomous driving via skill-enabled adversary learning for closed-loop scenario generation.*** IEEE Robotics and Automation Letters, (99), 1-8.

Domain Adaptation

- Navarro, I., Ortega, P., Patrikar, J., Wang, H., Ye, Z., Park, J. H., ... & Scherer, S. (2024). ***AmeliaTF: A Large Model and Dataset for Airport Surface Movement Forecasting.*** In AIAA AVIATION FORUM AND ASCEND 2024 (p. 4251).
- Diehl, C., Karkus, P., Veer, S., Pavone, M., & Bertram, T. (2025, May). ***Lord: Adapting differentiable driving policies to distribution shifts.*** In 2025 IEEE International Conference on Robotics and Automation (ICRA) (pp. 7036-7043). IEEE.
- Messaoud, K., Cord, M., & Alahi, A. (2025). ***Towards Generalizable Trajectory Prediction using Dual-Level Representation Learning and Adaptive Prompting.*** In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 27564-27574).

Trajectory Selection and Deployment

- Navarro, I., Patrikar, J., Dantas, J. P., Baijal, R., Higgins, I., Scherer, S., & Oh, J. (2024). ***SoRTS: Learned tree search for long horizon social robot navigation.*** IEEE Robotics and Automation Letters, 9(4), 3759-3766.

Scenario Mining

- Stoler*, B., Navarro*, I., Jana, M., Hwang, S., Francis, J., & Oh, J. (2024, June). ***Safeshift: Safety-informed distribution shifts for robust trajectory prediction in autonomous driving.*** In 2024 IEEE Intelligent Vehicles Symposium (IV) (pp. 1179-1186). IEEE.
- Ding, W., Veer, S., Leung, K., Cao, Y., & Pavone, M. (2025). ***Surprise potential as a measure of interactivity in driving scenarios.*** arXiv preprint arXiv:2502.05677.
- Yang, Y., Zhang, Q., Ikemura, K., Batool, N., & Folkesson, J. (2024, June). ***Hard cases detection in motion prediction by vision-language foundation models***. In 2024 IEEE Intelligent Vehicles Symposium (IV) (pp. 2405-2412). IEEE.

Task Failure Detection and Remediation

- Farid, A., Veer, S., Ivanovic, B., Leung, K., & Pavone, M. (2023, March). ***Task-relevant failure detection for trajectory predictors in autonomous vehicles.*** In Conference on Robot Learning (pp. 1959-1969). PMLR.
- Nakamura, K., Tian, T., & Bajcsy, A. (2025, January). ***Not All Errors Are Made Equal: A Regret Metric for Detecting System-level Trajectory Prediction Failures***. In Conference on Robot Learning (pp. 4051-4065). PMLR.
- Saadatnejad, S., Bahari, M., Khorsandi, P., Saneian, M., Moosavi-Dezfooli, S. M., & Alahi, A. (2022). ***Are socially-aware trajectory prediction models really socially-aware?.*** Transportation research part C: emerging technologies, 141, 103705.
- Filos, A., Tigkas, P., McAllister, R., Rhinehart, N., Levine, S., & Gal, Y. (2020, November). ***Can autonomous vehicles identify, recover from, and adapt to distribution shifts?***. In International Conference on Machine Learning (pp. 3145-3153). PMLR.

Autonomous Driving Paradigms

- Fu, H., Zhang, D., Zhao, Z., Cui, J., Liang, D., Zhang, C., ... & Bai, X. (2025). ***Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation***. arXiv preprint arXiv:2503.19755.

Thank you!