# Benchmarking Scene Representation and Encoding Methods for Behavior Prediction

## Motivation and Project Goal

Behavior prediction is crucial for autonomous systems interacting with humans, such as autonomous vehicles. To make accurate predictions in highly interactive and complex traffic scenes, it is important that the prediction model fully utilizes the context information inherited in the input, including the rich semantic information of HD-maps and the social interaction among road participants. To enable sufficient modeling capacity to encode all the useful information, various scene representation methods and encoder architectures have been proposed. To learn scene representations with rich map information, some widely adopted approaches include: 1) representing scenes as rasterized top-down semantic images and encoding using convolutional neural networks (CNNs) [1-3]; 2) representing scenes as vectorized graphs and encoding using graph neural networks (GNNs) [4-6]. To encode social context, attention-based [7] and GNN-based [8] models have been proposed. Furthermore, some approaches adopt a representation unifying HD-maps and agents, so that the model learns to simultaneously aggregate all the context information from the scene [9,10]. While those sophisticated scene representation and encoding methods lead to state-of-the-art prediction performance, there lacks principled studies on whether those proposed methods can indeed enable the prediction model to effectively encode useful context information from the input. In fact, some recent studies have shown that prediction models could ignore social context even if those state-of-the-art encoding modules are used [11,12]. In this project, we would like to fill in the research gap, by initiating the first comprehensive study to benchmark different scene representation and encoding methods for behavior prediction. Specifically, we want to develop a systematic evaluation scheme to quantitatively evaluate and compare the encoding efficiency of context information across different methods. The resulting evaluation scheme could be a complement to the current performance-based evaluation metrics. Also, by evaluating those state-of-the-art models, we can get inspiration for new representation methods, encoding models, and training methods to improve encoding efficiency.

## Challenges and Technical Approach

The largest challenge lies in defining the metrics to evaluate the learnt representations. There are no ground-truth labels for representations. Prior works mainly rely on the performance of the downstream prediction task to compare different representation methods [9]. However, merely comparing the average prediction accuracy is not sufficient to understand the behavior of the encoding module. Alternatively, we may utilize techniques from explainable AI (XAI) to help us analyze the encoding module and evaluate the learnt representation. For instance, we may conduct sensitivity analysis via methods such as Shapley values [11] and saliency maps [14] to quantify the contribution of each input feature to the model output. However, it is not sufficient to just learn the contribution of input features. Not every input feature contains useful information. To establish a practical and informative benchmark, we need to carefully select test cases, where there exists features with critical semantic information that the model ought to understand

Shapley values – a method from coalitional game theory – tells us how to fairly distribute the “payout” among the features (https://christophm.githu

a saliency map is an image that highlights the region on which people's eyes focus first. The goal of a saliency map is to reflect the degree of importance of a pixel to and utilize, for instance, stop lines and traffic lights at intersections, merging vehicles at roundabouts and highway entrances.

**Related Prior Work**

Our research group has explored different aspects and published plenty of research papers in behavior prediction. In particular, we have explored self-supervised learning methods for representation learning in behavior prediction [15, 16], where uni-modal and cross-modal contrastive learning methods are developed to pre-train map and trajectory representations. We showed that the pre-trained representations lead to better prediction performance. We have also investigated scene representation methods for transferable prediction models [17]. Works that are closely related to this project are [12, 18], where XAI techniques are developed and utilized to understand and diagnose the behavior of prediction models. In [12], we investigated whether VAE-based prediction models properly encode social interaction, by using a GNN layer with sparse graph attention. In [18], we studied the inherent temporal causality of conditional behavior prediction models with the help of Shapley values.

**Work Plan**

1. Select and implement state-of-art prediction models with different representation and encoding methods for benchmarking.
2. Design and implement evaluation schemes and metrics based on techniques and metrics from XAI and representation learning.
3. Construct testing dataset through mining existing large-scale traffic datasets.
4. Benchmark the selected prediction models and representation methods. Analyze the evaluation results to find the advantages and disadvantages of different methods and the common limitations of current approaches.
5. Explore new representation methods, encoding modules, and training methods to improve encoding efficiency based on the benchmark results.

## References

[1] Chai, Yuning, et al. "Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction." *arXiv preprint arXiv:1910.05449* (2019).

[2] Cui, Henggang, et al. "Multimodal trajectory predictions for autonomous driving using deep convolutional networks." *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.

[3] Salzmann, Tim, et al. "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data." *European Conference on Computer Vision (ECCV)*. Springer, Cham, 2020.

[4] Gilles, Thomas, et al. "Gohome: Graph-oriented heatmap output for future motion estimation." *arXiv preprint arXiv:2109.01827* (2021).

[5] Gilles, Thomas, et al. "THOMAS: Trajectory Heatmap Output with learned Multi-Agent Sampling." *arXiv preprint arXiv:2110.06607* (2021).

[6] Liang, Ming, et al. "Learning lane graph representations for motion forecasting." *European Conference on Computer Vision (ECCV)*. Springer, Cham, 2020.

[7] Alahi, Alexandre, et al. "Social lstm: Human trajectory prediction in crowded spaces." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[8] Mohamed, Abduallah, et al. "Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
[9] Gao, Jiyang, et al. "Vectornet: Encoding hd maps and agent dynamics from vectorized representation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.
[10] Deo, Nachiket, Eric Wolff, and Oscar Beijbom. "Multimodal trajectory prediction conditioned on lane-graph traversals." *Conference on Robot Learning*. PMLR, 2022.
[11] Makansi, Osama, et al. "You Mostly Walk Alone: Analyzing Feature Attribution in Trajectory Prediction." *arXiv preprint arXiv:2110.05304* (2021).
[12] Tang, Chen, Wei Zhan, and Masayoshi Tomizuka. "Exploring Social Posterior Collapse in Variational Autoencoder for Interaction Modeling." *Advances in Neural Information Processing Systems* 34 (2021).
[14] Kim, Jinkyu, and John Canny. "Interpretable learning for self-driving cars by visualizing causal attention." *Proceedings of the IEEE international conference on computer vision*. 2017.
[15] Ma, Hengbo, et al. "Multi-agent driving behavior prediction across different scenarios with self-supervised domain knowledge." *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021.
[16] Xu, Chenfeng, et al. "Pretram: Self-supervised pre-training via connecting trajectory and map." *2022 European Conference on Computer Vision (ECCV),* under review.
[17] Hu, Yeping, Wei Zhan, and Masayoshi Tomizuka. "Scenario-transferable semantic graph reasoning for interaction-aware probabilistic prediction." *arXiv preprint arXiv:2004.03053*(2020).
[18] Tang, Chen, Wei Zhan, and Masayoshi Tomizuka, "Interventional behavior prediction: Avoiding overly con- fident anticipation in interactive prediction," *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, under review.