Can I Trust My Trajectory Prediction Model?

Franz Motzkus[1,2,*], Christian Schlauch[1,3], Sebastian Bernhard[1] and Ute Schmid[2]

[1]AUMOVIO, Max-Urich-Str.3, 13355, Germany
[2]Universität Bamberg, Kapuzinerstraße 16, 96047 Bamberg, Germany
[3]Karlsruhe Institute for Technology, Kaiserstraße 12, 76131 Karlsruhe, Germany

Abstract

Explainability methods allow for inspecting model-internal mechanisms and thus provide transparency for previously black-box AI models. With increased interpretability, an AI developer can reveal and mitigate model errors with informed model improvements. Thus, the provided model transparency can lead to safer, more reliable models and enable fairness evaluations, the foundation for trust in AI. While explainable AI (xAI) research has made significant contributions, most approaches remain confined to baseline datasets and models, limiting their applicability in real-world domains. In autonomous driving, predicting drivable trajectories is key to anticipating traffic scenarios. Current black-box systems lack transparency, which amplifies with model complexity. We use Sparse Autoencoders (SAEs) to probe the latent space of a Wayformer trajectory predictor, exposing model internals as structured, interpretable features. Our study highlights the role of concepts, potential manipulation risks, and shows how SAEs improve transparency and support trustworthy AI development.

Keywords

Sparse Autoencoders, Explainable Trajectory Prediction, Applied xAI, Explainability

1. Introduction

Explainable AI (xAI) methods aim to make black-box AI models more interpretable, establishing transparency, unveiling spurious behavior, and enabling the targeted mitigation of erroneous behavior. Increasing the transparency of previously black-box behavior and enabling inspection and manipulation of internal representations provides the foundation for semantically testable and trustworthy systems.

Transparency is central to trustworthy AI, as it allows stakeholders to understand reasoning processes, audit models for biases and failure modes, and ensure compliance and accountability. Interpretable features also help identify vulnerabilities, strengthening robustness against adversarial use, while exposing ethically questionable correlations supports systematic fairness testing. Ultimately, transparency fosters trust by enabling AI systems to be semantically tested, validated, and explained in human terms.

Such trustworthiness is especially important in safety-critical applications like autonomous driving. Given its inherent complexity, many autonomous driving systems are decomposed into modular subtasks such as perception, trajectory prediction, and planning [1]. The perception module models the state of the environment, upon which the prediction module forecasts the future behavior of the surrounding agents. These predictions then serve as input to the planning module, which aims to generate a safe and comfortable trajectory for the autonomous vehicle.

In these modular autonomous driving systems, trajectory prediction models play a central role for maneuver planning and collision avoidance, with potentially fatal consequences in cases of failure. Although it was shown that trajectory predictors can over-rely on specific input features, ignoring critical contextual cues such as surrounding agents [2], most state-of-the-art models remain fundamentally black-box systems. This crucial downside impedes the root cause discovery of failure cases, which are currently handled by retraining the model on new data. Advances in the explainability of perception

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.

[*]Corresponding author.

✉ franz.walter.motzkus@aumovio.com (F. Motzkus); christian.schlauch@aumovio.com (C. Schlauch); sebastian.2.bernhard@aumovio.com (S. Bernhard); ute.schmid@uni-bamberg.de (U. Schmid)
0009-0009-4362-7907 (F. Motzkus); 0000-0002-6970-0094 (C. Schlauch); 0000-0002-7194-7539 (S. Bernhard); 0000-0002-1301-0326 (U. Schmid)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-12-16

models, such as attention map visualizations [3], offer only limited insights into trajectory predictors, since the abstract and complex feature interactions remain difficult to interpret.

In this work, we explore the latent space of the Wayformer trajectory prediction model [4] as a popular representative for other models in the trajectory prediction domain. We investigate how concepts are encoded in the Wayformer's latent space and how they are represented in the trajectory prediction domain. We use Sparse Autoencoders (SAEs) – an unsupervised learning technique that aims to find disentangled, human-interpretable features, as they have recently shown to find interesting insights in model-internal encodings [5]. With feature disentanglement, we seek to obtain structured insights into the semantic model encodings, like steering decisions or encoded scenarios. By selectively manipulating individual features in the SAE's latent space, we observe measurable changes in the predicted output, demonstrating a direct causal influence and exposing a significant security risk in safety-critical applications like autonomous driving. Finally, we explore the objective how this xAI-based approach can contribute to the development of transparent and trustworthy autonomous systems.

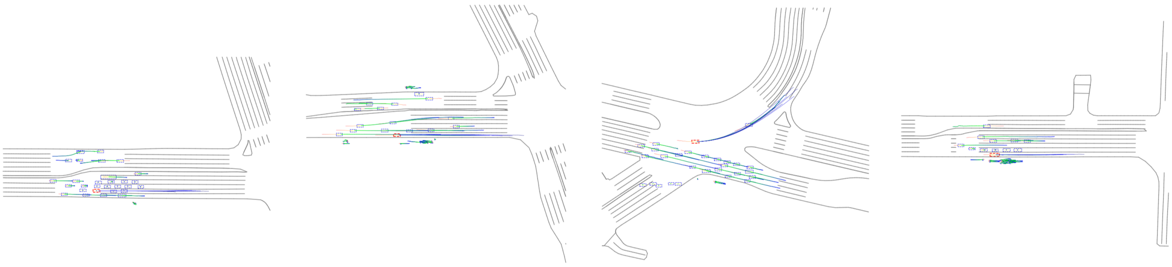


Figure 1: For the trajectory prediction problem, an abstract representation of a traffic scenario in bird's-eye view is provided to a model. The single parts, such as road marks, traffic guidance, agents, and their history, form individual inputs, but can be jointly visualized as shown. The model predicts a set of future trajectories for the ego vehicle (red), including their probability.

2. Background

2.1. Trajectory Prediction

The tested Wayformer model is a state-of-the-art marginal trajectory predictor, which simplifies the problem of modelling the behavior of all present agents by only considering a single target agent at a time, see Figure 1. More formally, it aims to estimate the states $x_{i,t|1:N} = (x_{i,t|1}, .., x_{i,t|N})$ of a target agent i at a timepoint t over a sequence of N time steps into the future, given a scene context as input. This scene context $c_t = \{x_{i,t|H:1}, x_{j,t|H:1}, m_t, t_t\}$ consists of the observed states $x_{i,t|H:1}$ of the target i and $x_{j,t|H:1}$ of its surrounding agents $j \neq i$ over a history length H, as well as the road's topology m_t (drivable area and lane markings) and dynamic traffic guidance g_t (traffic lights or speed limits), centered around the target's current state. The evolution of the target's future states is inherently multi-modal, reflecting the multitude of possible maneuvers and interactions with other agents in the scene. Wayformer approximates this evolution using a Gaussian mixture model (GMM) decoding strategy

$$p(x_{i,t|1:N}|c_t) \approx \sum_{m=1}^{K} \pi_{m,t}(c_t)\, \mathcal{N}(x_{i,t|1:N} \mid \mu_{m,t|1:N}, \Sigma_{m,t|1:N}),$$

where each mixture component or mode m represents a distinct predicted trajectory. Its architecture consists of a transformer-based early-fusion encoder with a transformer-based decoder using compute-efficient latent query encodings. The two output heads estimate the mode probability $\pi_{m,t}$ and the parameters of the Gaussian mixtures $\mu_{m,t|1:N}$ and $\Sigma_{m,t|1:N}$, respectively. The encoder-decoder transformer architecture and GMM decoding strategy in Wayformer is commonly used in state-of-the-art models on public benchmarks [6, 7, 8].

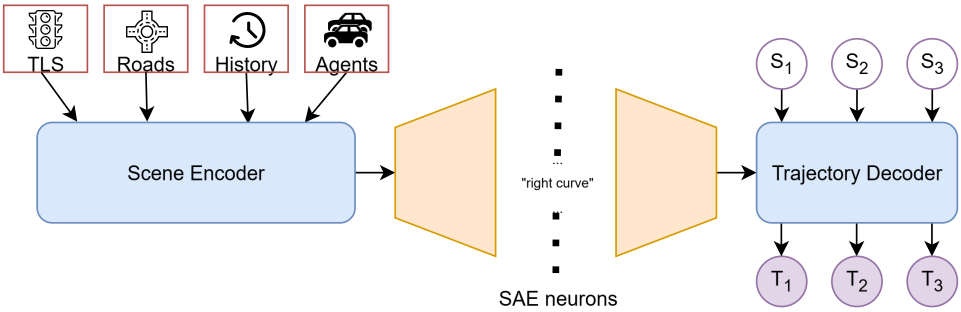


Figure 2: Visualization of the overall model structure. The SAE is applied to the scene context embedding produced by the scene encoder. After applying the SAE, the reconstructed scene context is forwarded through the trajectory decoder together with the seeds S_1-$S_3$to output the predicted trajectories T_1-T_3.

2.2. Explainability for Trajectory Prediction

Previous explainability approaches focus on the input attribution in trajectory predictors. Makansi et al. [2] use Shapley values to show that trajectory predictors are only marginally influenced by most of the scene context features, such as the road topology or other agents' states, which casts doubt on the predictors' abilities to capture safety-critical interactions. This inspired new model designs [9, 10, 11] that aim to make the feature attribution or embedding spaces in prediction models more interpretable. However, explainability methods are still rarely applied to state-of-the-art prediction models, even though approaches in the perception of autonomous driving systems showcase their potential to capture unusual and potentially safety-critical situations [3].

2.3. Sparse Autoencoders

Sparse Autoencoders (SAEs) have lately come to attention by applying the known principle of putting sparsity constraints on higher-dimensional data representations [12]. In latent representations with polysemantic neurons, SAEs can disentangle the semantic directions in the data, unveiling the internal concept directions in monosemantic neurons [13, 14]. In higher-dimensional space, each feature direction can be represented as a mono-semantic neuron, while adding sparsity ensures the reduction of the sample representation to as few neurons as possible, avoiding multi-semanticity in neurons. While an applied sparsity constraint may lead to vanishing weights, sparsity can be enforced by a top-k activation in the latent space instead [15]. Further improvements by initializing the decoder as the inverse of the encoder and including the activation of dead neurons in the loss functions lead to better-performing large-scale SAEs [16].

The processing step of the SAE can be formalized as:

$$\begin{aligned} z &= RELU(W_{enc}x) \\ \hat{z} &= TOPK(z) \\ \hat{x} &= W_{dec}\hat{z} \end{aligned} \tag{1}$$

with $W_{enc} \in R^{n \times d}$ and $W_{dec} \in R^{d \times n}$. The training loss is defined as the weighted sum of the reconstruction loss and the sparsity loss $L = \parallel x - \hat{x} \parallel_2^2 + \lambda \parallel z \parallel_1$.

3. Hyperparameters and Training

We train a Nano version of the Wayformer model [4] with half the number of layers and hidden size, as the full-sized model has been shown to overfit on in-distribution test sets [17]. We train the Wayformer-Nano jointly on the public NuScenes [18], Argoverse2 [19], Waymo-Open [20], and Shifts [21] datasets using the standard loss composed of a classification loss on the mode probabilities and a regression loss on the mixture parameters.

In the Wayformer model, the scene context embedding describes a compressed representation between the scene encoder and the trajectory decoder, describing a dense understanding of the scenario. We train a Sparse Autoencoder (SAE) on the flattened scene context embeddings to extract monosemantic, interpretable feature directions. The sparse SAE features may then encode certain road or agent characteristics that lead to specified driving behavior. The semantics of a feature are extracted with Activation Maximization [22], allowing for later modification of a feature like "left turn" within the SAE. Certain driving behaviors can then be triggered according to the modifiable features in the SAE.

The SAE consists of two linear layers with a latent representation of 49152 neurons, denoting an expansion factor of 4. Multiple sizes of the latent dimension have been tested, while the stated setting provides a reasonable trade-off between sufficient disentanglement by producing first dead neurons – thus being large enough –, and the demand for preferably small dimensionality to maintain explainability. We apply a top-k filter on the SAE activations with $k = 128$. We train the SAE for 50 epochs and quantitatively test its reconstruction quality. We measure the MSE between the original scene context embedding and the reconstruction of the SAE, as well as the influence on the predicted trajectory in terms of the minADE to the top-1 prediction, as the SAE naturally adds an error to the overall processing of the Wayformer. The minADE is the minimal average displacement error among all predictions towards a target trajectory, which is in our case the predicted top-1 trajectory of the unmodified model. We report an MSE of 0.006, showing a very low reconstruction error, and a mean minADE score of 0.263, denoting a mean deviation of 26 cm, which is low in comparison to the average minADE towards the ground truth prediction (135 cm). After the training, 26999 active neurons remain, while the rest of the neurons never fire.

4. Experiments

We evaluate the SAE encodings representing the Wayformer's latent space in two steps. First, we inspect high-activating neurons and describe how to systematically assign semantic meaning to the neurons. Then, we exploit the extracted knowledge to manipulate the trajectory prediction process.

4.1. Semantic Concepts in Trajectory Prediction

Our trained SAE provides a latent space representation of sparse disentangled features represented as SAE neurons. With the activation maximization approach [22], the top-activating samples per neuron can be extracted to reveal the semantic encoding of a neuron. Whereas in standard image settings, a foundation model can be used to assign a semantic label to each neuron [23], we can only refer to a manual inspection as a knowledge source. Figure 3 shows the visualization of the highest-activating inputs for a random selection of high-activating neurons from the SAE encoding. The similarity in the provided samples is interpreted as the information encoded in the respective neuron. However, the semantic feature encoding is occasionally ambiguous, and no distinct property can be assigned. Anyhow, hypotheses about a neuron's semantic encoding can be postulated and later refined. For neuron 21952 in Figure 3, a highly dense traffic scenario could be encoded with many surrounding cars and pedestrians on the sidewalk. Neuron 7828 potentially encodes a slight right turn to follow the selected lane, while neuron 23899 indicates that it is specific to a single scenario, highlighting a risk of overfitting the data. This information can already be extracted and used for multiple use cases without further testing, although manually testing every neuron in the SAE is tedious and costly.

To counter the expansion in the latent space of the SAE, neurons can be automatically grouped to reduce the complexity of the search space. Figure 4 shows the correlation matrix between non-zero neurons of the trained SAE. A clear disentanglement of the features can be seen, as most neurons are only correlated to a few other neurons, highlighting unique feature encodings. Hierarchical feature clustering can be applied to the correlation matrix, grouping similar neurons together. A human inspector can then decide on a granularity level to inspect the clusters from a coarse to the neuron level. Assuming common features causing a vehicle to make a turn in our use case, these features are expected to be grouped together in the same cluster hierarchy. Depending on the desired feature granularity,

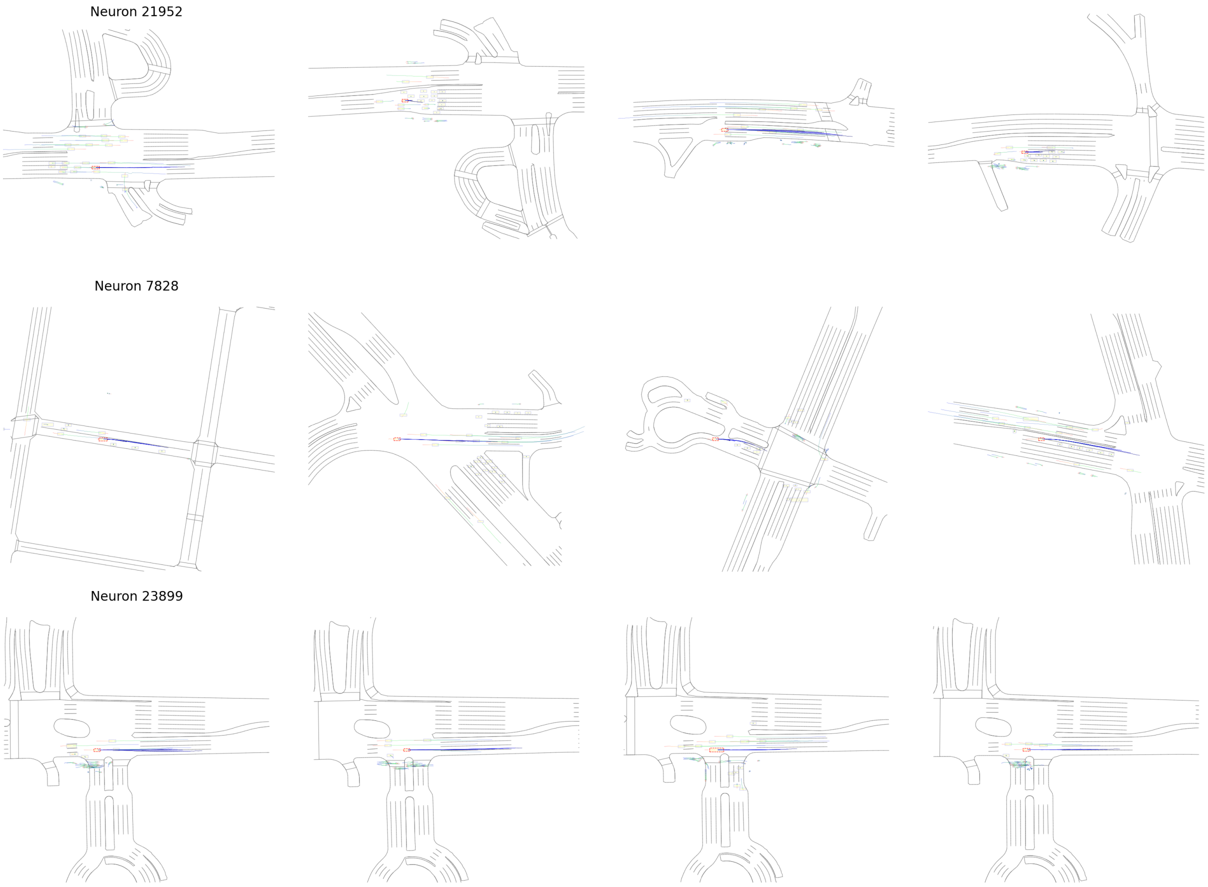


Figure 3: The semantic role of a neuron can be explored through activation maximization, which identifies the inputs that most strongly activate it. By visualizing and comparing these high-activation samples, the underlying feature encoded by the neuron can be interpreted and assigned semantic meaning. E.g., for neuron 7828, scenarios with a car slightly turning right to follow the lane are shown.

the SAE neurons in this group can be further evaluated, or the next-finer level of the hierarchy can be explored.

When inspecting a selected neuron's highest correlating neurons, a clear alignment in terms of the predicted output is observable. These correlated neurons implement complementary aspects related to a specific driving maneuver. The predicted action can further be evaluated in terms of causality with regard to the extracted aspects, as a left turn may be encoded by multiple neurons, with differences in which aspects the neurons encode and how much influence their activation has on the "left turn" action. Hereby, false predictions can potentially be evaluated with regard to the underlying feature activations.

As some vague interpretation is currently necessary for assigning semantics to single neurons, further methods need to be developed to specify the single encoded feature. In the next section, we show how to verify the assigned semantics quantitatively by testing the downstream manipulation.

4.2. Downstream Prediction Manipulation

By identifying semantically meaningful neurons in the SAE latent space and deriving their influence on the predicted output, we can selectively activate specific neurons to steer the model's predictions. This mechanism can be used to verify the neuron encoding and its influence on the prediction, but it can also be used to manipulate the downstream predictions. Using the SAE decoder, we construct a simple downstream feature manipulation pipeline. The SAE encoding of a scene context is manipulated by perturbing one or multiple neuron activations before reconstructing and forwarding the manipulated scene context to the Wayformer decoder. The resulting changes in the trajectory predictions reveal the

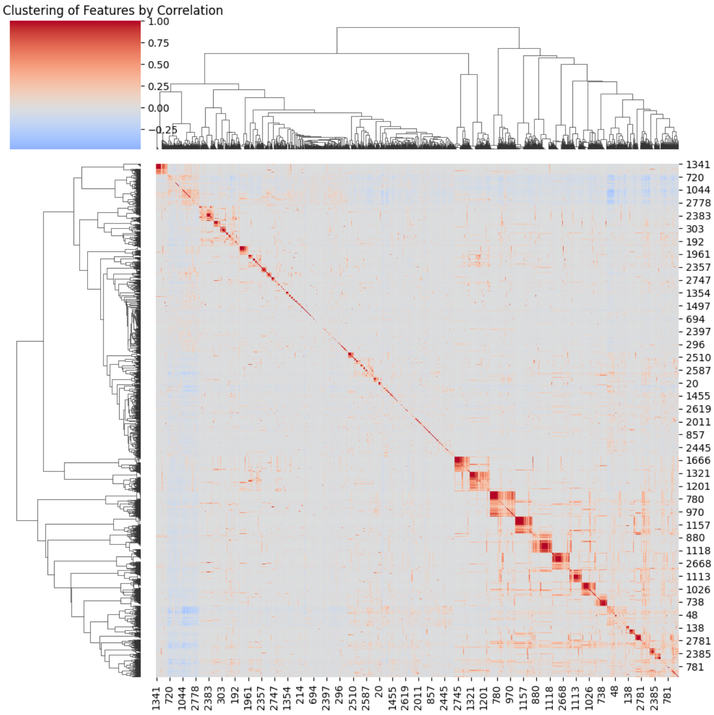


Figure 4: The correlation matrix shows the correlation between neurons, whereas each neuron encodes a unique characteristic. The hierarchical feature clustering sorts the SAE neurons into semantically meaningful clusters.

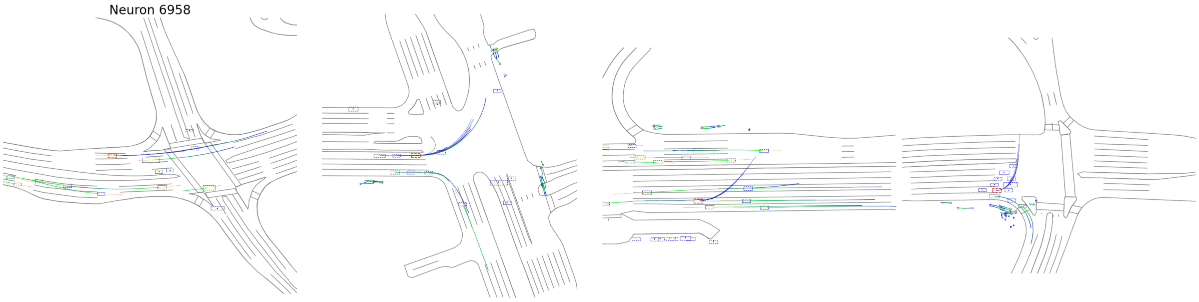


Figure 5: Visualization of the highest-activating samples for the SAE neuron 6958. All depicted samples show a left turn of the ego vehicle, drawing a strong association of the neuron with left turns.

influence of the feature manipulation.

As illustrated in Figure 6, the targeted manipulation of the SAE representation can induce a desired model prediction. By increasing the activation of neuron 6958, which is associated with left turn trajectories, we can alter the Wayformer predictions for random samples to shift towards a left turn for the ego vehicle.

To quantify this effect, we conduct an experiment comparing the manipulated predictions to the top-1 predicted trajectory of the original Wayformer model using the minADE score. Specifically, we perturb the activation value of a single SAE neuron and compute the minADE between the original and manipulated predictions. As shown in Figure 7, neuron 6958, which is responsible for left turns, causes a significantly higher deviation in the manipulated predictions than other neurons, confirming its strong influence on the steering direction. This shows that certain features encode direction-specific behaviors, from which steering-specific neurons can be actively manipulated towards a desired prediction. With input perturbations to activate such neurons, these encodings can potentially be exploited, forming a safety risk. Understanding the relation between these features and their embedding in the overall model processing requires further investigation. Thorough quantitative benchmarking is currently not possible due to missing ground truth explanations and fine-grained concept-based annotations.

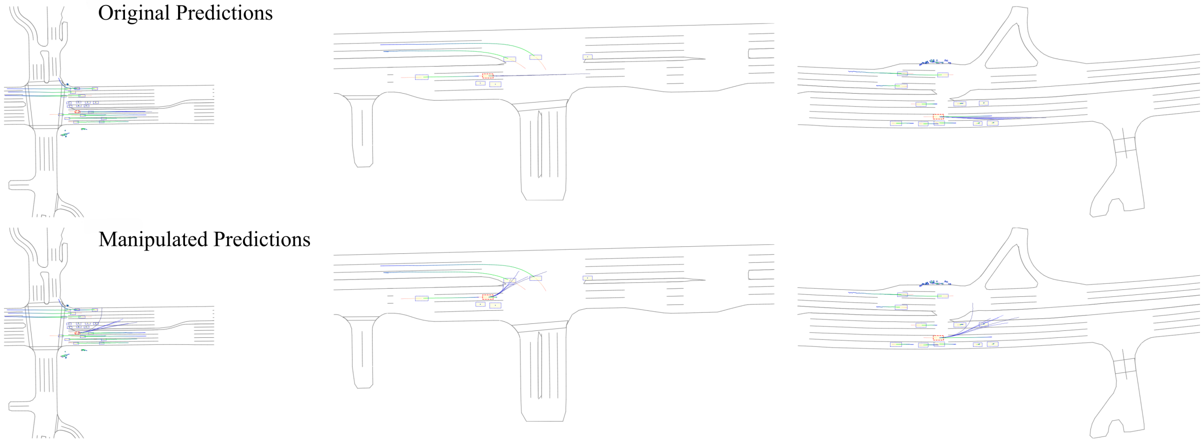
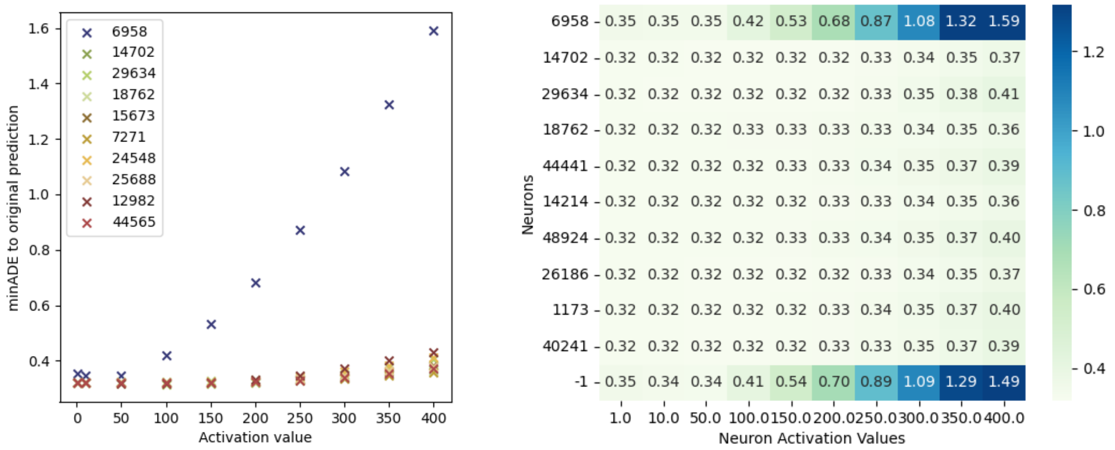


Figure 6: Comparison of Wayformer predictions with and w/o SAE neuron manipulations on neuron 6958, which is associated with left turns. Independent of the scenario, the predictions with a manipulated neuron activation show a strong tendency to left turns. Thus, the driving prediction can be heavily influenced by a single neuron.



Neurons	1.0	10.0	50.0	100.0	150.0	200.0	250.0	300.0	350.0	400.0
6958	0.35	0.35	0.35	0.42	0.53	0.68	0.87	1.08	1.32	1.59
14702	0.32	0.32	0.32	0.32	0.32	0.32	0.33	0.34	0.35	0.37
29634	0.32	0.32	0.32	0.32	0.32	0.32	0.33	0.35	0.38	0.41
18762	0.32	0.32	0.32	0.33	0.33	0.33	0.33	0.34	0.35	0.36
44441	0.32	0.32	0.32	0.32	0.33	0.33	0.34	0.35	0.37	0.39
14214	0.32	0.32	0.32	0.32	0.32	0.33	0.33	0.34	0.35	0.36
48924	0.32	0.32	0.32	0.32	0.33	0.33	0.34	0.35	0.37	0.40
26186	0.32	0.32	0.32	0.32	0.32	0.32	0.33	0.34	0.35	0.37
1173	0.32	0.32	0.32	0.32	0.32	0.33	0.34	0.35	0.37	0.40
40241	0.32	0.32	0.32	0.32	0.32	0.33	0.33	0.35	0.37	0.39
-1	0.35	0.34	0.34	0.41	0.54	0.70	0.89	1.09	1.29	1.49

Neuron Activation Values

Figure 7: Quantitative comparison between the manipulation of different neurons. The minADE of the original prediction to the manipulated prediction is measured over the dataset for multiple replacement values. Neuron -1 denotes the group of highly correlating neurons, including neuron 6958.

5. Conclusion

Most AI-based trajectory prediction models operate as black boxes, optimizing solely on average performance scores while lacking transparency. This limits failure analysis and hinders verification, both of which are essential for safe deployment in autonomous driving. Our approach addresses this gap by using SAEs for revealing interpretable features in the latent space of trajectory prediction models. For our exemplary used Wayformer, we reveal sparse, disentangled feature directions within the previously opaque scene context encoding, enabling structured analysis and controlled manipulation of model internals for desired trajectory outputs by an expert. We demonstrate how the activation of a single SAE feature can shift the predicted trajectories, which can even be actively steered towards desired directions. When the feature can be inadvertently triggered through spurious or adversarial input modifications, this demonstrates a potential safety risk, underscoring the need for a better understanding of the model processing. To our knowledge, this is one of the first approaches to increase the interpretability of the latent space in trajectory prediction models. Our findings highlight the necessity of applying latent space interpretability methods to trajectory prediction models to support model verification, targeted testing, and failure analysis, which are essential components for promoting trust in AI-driven systems.

Acknowledgments

The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Energy within the project “Safe AI Engineering – Sicherheitsargumentation befähigendes AI Engineering über den gesamten Lebenszyklus einer KI-Funktion" and "“NXT GEN AI METHODS – Generative Methoden für Perzeption, Prädiktion und Planung". The authors would like to thank the consortia for the successful cooperation.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly for grammar and spelling checks. Further, the authors used GPT to support the formulation of single sentences, which have been edited towards the final version. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] S. Hagedorn, M. Hallgarten, M. Stoll, A. P. Condurache, The integration of prediction and planning in deep learning automated driving systems: A review, IEEE Transactions on Intelligent Vehicles (2024) 1–17. doi:10.1109/TIV.2024.3459071.

[2] O. Makansi, J. V. Kügelgen, F. Locatello, P. V. Gehler, D. Janzing, T. Brox, B. Schölkopf, You mostly walk alone: Analyzing feature attribution in trajectory prediction, in: International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=POxF-LEqnF.

[3] A. Stocco, P. J. Nunes, M. d’Amorim, P. Tonella, Thirdeye: Attention maps for safe autonomous driving systems, in: 37th IEEE/ACM International Conference on Automated Software Engineering, ASE 2022, Rochester, MI, USA, October 10-14, 2022, ACM, 2022, pp. 102:1–102:12. URL: https://doi.org/10.1145/3551349.3556968. doi:10.1145/3551349.3556968.

[4] N. Nayakanti, R. Al-Rfou, A. Zhou, K. Goel, K. S. Refaat, B. Sapp, Wayformer: Motion forecasting via simple efficient attention networks, in: 2023 IEEE International Conference on Robotics and Automation (ICRA), 2023, pp. 2980–2987. doi:10.1109/ICRA48891.2023.10160609.

[5] J. Ferrando, O. B. Obeso, S. Rajamanoharan, N. Nanda, Do i know this entity? knowledge awareness and hallucinations in language models, in: The Thirteenth International Conference on Learning Representations, 2025. URL: https://openreview.net/forum?id=WCRQFlji2q.

[6] S. Shi, L. Jiang, D. Dai, B. Schiele, Motion transformer with global intention localization and local movement refinement, Advances in Neural Information Processing Systems 35 (2022).

[7] R. Girgis, F. Golemo, F. Codevilla, M. Weiss, J. A. D’Souza, S. E. Kahou, F. Heide, C. Pal, Latent variable sequential set transformers for joint multi-agent motion prediction, in: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, OpenReview.net, 2022. URL: https://openreview.net/forum?id=Dup_dDqkZC5.

[8] J. Sun, S. Yan, X. Song, Qcnet: query context network for salient object detection of automatic surface inspection, Vis. Comput. 39 (2023) 4391–4403. URL: https://doi.org/10.1007/s00371-022-02597-w. doi:10.1007/S00371-022-02597-W.

[9] P. Xu, J. Hayet, I. Karamouzas, Socialvae: Human trajectory prediction using timewise latents, in: S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, T. Hassner (Eds.), Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV, volume 13664 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 511–528. URL: https://doi.org/10.1007/978-3-031-19772-7_30. doi:10.1007/978-3-031-19772-7_30.

[10] C. Wong, Z. Zou, B. Xia, X. You, Resonance: Learning to predict social-aware pedestrian trajectories as co-vibrations, CoRR abs/2412.02447 (2024). URL: https://doi.org/10.48550/arXiv.2412.02447. doi:10.48550/ARXIV.2412.02447.

[11] Y. Tang, W. Ma, Intent: Trajectory prediction framework with intention-guided contrastive clustering, 2025. arXiv:2503.04952.

[12] B. A. Olshausen, D. J. Field, Sparse coding with an overcomplete basis set: A strategy employed by v1?, Vision Research 37 (1997) 3311–3325. URL: https://www.sciencedirect.com/science/article/pii/S0042698997001697. doi:https://doi.org/10.1016/S0042-6989(97)00169-7.

[13] T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, Z. Hatfield-Dodds, A. Tamkin, K. Nguyen, B. McLean, J. E. Burke, T. Hume, S. Carter, T. Henighan, C. Olah, Towards monosemanticity: Decomposing language models with dictionary learning, Transformer Circuits Thread (2023). Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

[14] R. Huben, H. Cunningham, L. R. Smith, A. Ewart, L. Sharkey, Sparse autoencoders find highly interpretable features in language models, in: The Twelfth International Conference on Learning Representations, 2024. URL: https://openreview.net/forum?id=F76bwRSLeK.

[15] A. Makhzani, B. Frey, k-sparse autoencoders, 2014. URL: https://arxiv.org/abs/1312.5663. arXiv:1312.5663.

[16] L. Gao, T. D. la Tour, H. Tillman, G. Goh, R. Troll, A. Radford, I. Sutskever, J. Leike, J. Wu, Scaling and evaluating sparse autoencoders, in: The Thirteenth International Conference on Learning Representations, 2025. URL: https://openreview.net/forum?id=tcsZt9ZNKD.

[17] M. Bouzidi, C. Schlauch, N. Scheuerer, Y. Yao, N. Klein, D. Göhring, J. Reichardt, Closing the loop: Motion prediction models beyond open-loop benchmarks, CoRR abs/2505.05638 (2025). URL: https://doi.org/10.48550/arXiv.2505.05638. doi:10.48550/ARXIV.2505.05638. arXiv:2505.05638.

[18] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, O. Beijbom, nuscenes: A multimodal dataset for autonomous driving, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[19] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, D. Ramanan, P. Carr, J. Hays, Argoverse 2: Next generation datasets for self-driving perception and forecasting, in: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021), 2021.

[20] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou, Z. Yang, A. Chouard, P. Sun, J. Ngiam, V. Vasudevan, A. McCauley, J. Shlens, D. Anguelov, Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9710–9719.

[21] A. Malinin, N. Band, Y. Gal, M. Gales, A. Ganshin, G. Chesnokov, A. Noskov, A. Ploskonosov, L. Prokhorenkova, I. Provilkov, V. Raina, V. Raina, D. Roginskiy, M. Shmatova, P. Tigas, B. Yangel, Shifts: A dataset of real distributional shift across multiple large-scale tasks, in: J. Vanschoren, S. Yeung (Eds.), Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, volume 1, 2021. URL: https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/ad61ab143223efbc24c7d2583be69251-Paper-round2.pdf.

[22] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, J. Clune, Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 29, Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/5d79099fcdf499f12b79770834c0164a-Paper.pdf.

[23] M. Dreyer, J. Berend, T. Labarta, J. Vielhaben, T. Wiegand, S. Lapuschkin, W. Samek, Mechanistic understanding and validation of large AI models with semanticlens, CoRR abs/2501.05398 (2025). URL: https://doi.org/10.48550/arXiv.2501.05398. doi:10.48550/ARXIV.2501.05398. arXiv:2501.05398.