# Interpretable Video Transformers in Imitation Learning of Human Driving

**Andrew Dai** [1] **Wenliang Qiu** [1] **Bidisha Ghosh** [1]

## Abstract

Transformers applied to high-level vision tasks showcase impressive performance due to the use of self-attention sublayers for computing affinity weights across tokens corresponding to image patches. A simple Vision Transformer encoder can also be trained with video clip inputs from popular driving datasets in a weakly supervised imitation learning task, framed as predicting future human driving actions as a time series sequence over a prediction horizon. In this paper, we propose this task as a simple, scalable method for autonomous vehicle planning to match human driving behaviour. We demonstrate initial results for this method, along with model visualizations for interpreting features in video inputs that contribute to sequence predictions.

## 1. Introduction

Autonomous driving for robust operation across all possible domains is an active area of research in applied robotics. Perception, motion planning, and actuation are encompassing tasks in self-driving systems, yet recent advances in fields that contribute to improved performance in these aspects (Yurtsever et al., 2020) have not led to successes in fully autonomous driving.

Learning from demonstrations (LfD) is a methodology that can be applied to transfer a driving policy from example demonstrations by expert human drivers. Extensive work has been carried out in this field, as covered in this survey (Argall et al., 2009). An imitation learning approach would enable the direct mapping of RGB image pixels (e.g. from dashcams) to a driving action in an end-to-end supervised learning manner. The availability of many publicly available driving datasets (Geiger et al., 2013; Yu et al., 2020; Sun et al., 2020) can enable successful transfer of driving knowledge to autonomous vehicles. This is especially useful for enabling policies to behave more like human drivers, as well as learn useful visual features that correspond to actions from a wide variety of driving scenarios. Several works have applied imitation learning from camera inputs towards the control of autonomous vehicles (Bojarski et al., 2016; Xu et al., 2017).

While learning-based methods are promising for use in autonomous driving, explainability of predictions from learned driving policies remains a significant challenge. Visualization of input parts that the model tends to focus on and determine policy outputs is important for debugging the learning algorithm when applied to real-world driving. Implicit awareness of interactions between the ego-vehicle and surrounding environment, or interactions between surrounding static and dynamic objects by the driving model from experience, would be ideal to visualize. This would aid in the design of more interpretable methods that would otherwise be black box model approaches to autonomous driving.

Transformers (Vaswani et al., 2017) have proven to be an effective tool for computer vision tasks such as object detection (Carion et al., 2020), image classification (Dosovitskiy et al., 2020) and image processing (Chen et al., 2020). Pairwise affinity attention weights add more significance to certain parts of inputs in model predictions. Furthermore, making input tokens correspond to square image patches instead of pixels (Carion et al., 2020; Dosovitskiy et al., 2020) solved the problem of quadratic memory cost with respect to input sequence length. Transformers can also be used in tasks involving video input sequences (Dwibedi et al., 2020; Bertasius et al., 2021). In addition to solving these tasks, self-attention score maps can also be used as a component towards interpretability of Transformer models.

In this paper, we present a simple approach to imitation learning in the context of autonomous driving, using a Vision Transformer encoder and MLP head that learns to predict an output sequence of 'Stop' and 'Go' labels (representing future throttle actions) from short, fixed, video-clip inputs observed before a current timestep. The model is trained to predict the movement of the ego-vehicle multiple timesteps ahead in the future within a short fixed-time horizon, by observing image frames recorded in recent past. In addition, we show empirical results in Transformer in-

[1]Department of Civil, Structural and Environmental Engineering, Trinity College Dublin, Ireland. Correspondence to: Andrew Dai <adai@tcd.ie>.

$3^{rd}$ *ICML Workshop on Human in the Loop Learning*, 2021. 

terpretability by computing a heatmap that shows relative attention scores for the video input towards the output prediction.

Unlike a full end-to-end controller, this framework can be used in a motion planner without modelling of driving scenarios with multiple possible actions (e.g. in junctions) as part of the learning process. The aims of this work are: 1) to design a simple imitation learning model that can be implemented for autonomous vehicle planning and automatic braking, 2) to improve scalability and ease-of-use for a system that can be used for online/active learning and fine-tuning with an expert human driver, and 3) to explore the feasibility of visualization techniques for Transformers applied to imitation learning with noisy labels.

The paper is organized as follows. Section 2 covers related work. Section 3 outlines our framework methodology. Section 4 highlights results on prediction performance. Section 5 describes the visualization process and example results. Section 6 concludes.

## 2. Related Work

Imitation learning from pixels for autonomous driving was first demonstrated with the ALVINN system (Pomerleau, 1989). It uses a small end-to-end neural network to steer a car trained on lane following. (Bojarski et al., 2016) followed up on this approach and demonstrated successful lateral control in highways and empty lanes using a CNN architecture. During test-time, it maps single image frames to steering outputs, and does not utilize temporal context in its predictions. (Xu et al., 2017) demonstrates end-to-end learning of a driving policy from a large-scale driving dataset, and framed as a future egomotion prediction task from video in discrete and continuous output cases. Using a FCN-LSTM architecture, it uses semantic segmentation as a side task to improve FCN feature extraction of image frames from video, which serves as input to an LSTM along with previous egomotion. It is trained more closely to align with human driving in complex urban driving scenarios. However, recurrent architectures are still limited in terms of information flow for output predictions.

Interpretability of vision-based systems is relevant to our understanding of how a black box model behaves in real world applications. Existing methods for interpreting models in computer vision typically focus on generating heat maps to find some connections between the output prediction and features from the input signal that contributed to it. A comparison of visualization techniques was made in (Chefer et al., 2021), and they carried out experiments to test some of these methods for Vision Transformers in image classification tasks. For video understanding (Bertasius et al., 2021), visualization of an attention map is also done across frames in an input video clip. However, the proposed results in interpretability from these works are limited to debugging the model to ensure objects desired in the output labels are highlighted by saliency visualizations. In (Bojarski et al., 2016), they showcase results on salient features from CNN feature activation maps. (Kim & Canny, 2017) focuses on interpretability, and uses an attention heat map prediction with causal filtering to explain image patches relevant to model predictions for vehicle steering control. They also use this map as an input to the driving policy. However, the wider goal of explainability of black box prediction models for developers and end users remains a weak point of these tools, as emphasized in the literature (Mittelstadt et al., 2019). This is especially true for end-to-end driving models, where it is shown that existing methods for model interpretability lack sophisticated ability in providing feedback on spatiotemporal features that explain the reasoning behind future action predictions. Our experiments showcase attention-based visualizations for finding relevancy of image patches in video clips to Transformer model predictions, and highlight challenges in explainability for imitation learning models.

## 3. Method

The framework used in our offline learning experiments is shown in Figure 1. We propose a method for imitation learning on high-level future egomotion prediction (with 'Stop' and 'Go' labels), as is also done in (Xu et al., 2017). The model makes predictions using spatiotemporal reasoning on fixed-length video clips using joint space-time attention. This framework allows for potential planning algorithms to use these outputs for detecting risky driving scenarios in any application, while being more scalable for offline learning on a wide range of datasets before fine-tuning using active learning with human drivers operating a vehicle with automatic braking. Self-attention layers enable further interpretability of features in video clips that are most relevant to the prediction.

The input to the framework is a series of consecutive RGB frames from driving dataset video clips. We use F input frames as the backward context of the current timestep. The prediction output is based on the information presented in the backward context alone. The KITTI Raw dataset (Geiger et al., 2013) is used with a training-validation split fixed across experiments. Egomotion data synced with video frames at a 10Hz sampling rate is used to generate ground truth labels for future prediction.

The ResNet18 backbone (He et al., 2016) is pretrained using the self-supervised method presented in (Jabri et al., 2020), and is frozen during training on the driving task. Following them, we also extract features from the modified `res3` block, giving a 256-channel feature vector (downsampled

by a factor of 8) from video clip inputs. We use a Vision Transformer (ViT) encoder (Dosovitskiy et al., 2020) that uses model parameters following BERT-Small (Turc et al., 2019), and adding dropout of 0.1. The input to this is the feature vector as a 256-channel series of consecutive frames. The MLP head is a sequence classifier using binary cross entropy loss across the ground truth action labels for a given clip. It has a hidden layer of size 1024 with ReLU activation, and outputs logit probabilities of the 'Go' labels $y_N$ from 1 to N forward timesteps in the output sequence.

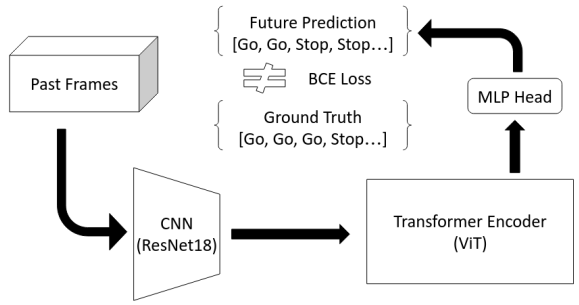



*Figure 1.* Overview of our Transformer-based framework for the 'Stop' and 'Go' driving task.

## 4. Transformer Predictions in Imitation Learning

Following our approach, we define observed backward context frames of length F = 6, and predict a sequence of future egomotion labels with N = 8 elements. Frames are downsized to 416x128 pixels using bilinear interpolation. The train-val split is fixed with a 90:10 ratio. We assign 'Stop' labels on frames where acceleration is under -2 m/s or velocity is under 2 m/s, making the proportion of these labels about 10% across all video clips. All other frames are assigned 'Go' labels. Here, we present results on future prediction task performance.

### 4.1. Classification Performance

After training, the framework was evaluated on the validation split to measure performance. To compare the effect of temporal reasoning on this task, we also tested an identical model with F = 1. Results on precision, recall and F1 Score are calculated and shown in Table 1. In addition, ROC curve false positive results at different true positive thresholds are shown in Table 2.

Based on the first performance metrics, the model with F = 1 appears to have slightly higher performance than the model with F = 6. However, it is observed that the model that predicts from still frames tends to make a significantly higher number of false 'Stop' predictions compared to the model that uses video observations. In addition, ROC results show significantly higher classification performance when F = 6 across all thresholds.

*Table 1.* Classification performance of our models with varying input video frame length on KITTI dataset.

| MODEL | F1 | PRECISION | RECALL |
|---|---|---|---|
| F = 1 | 0.966 | 0.945 | 0.940 |
| F = 6 | 0.973 | 0.850 | 0.967 |

*Table 2.* ROC false positive rate of our models at true positive (TP) thresholds.

| MODEL | TP THRESHOLD | | |
|---|---|---|---|
| | 0.95 | 0.98 | 0.99 |
| F = 1 | 0.014 | 0.029 | 0.043 |
| F = 6 | 0.009 | 0.018 | 0.027 |

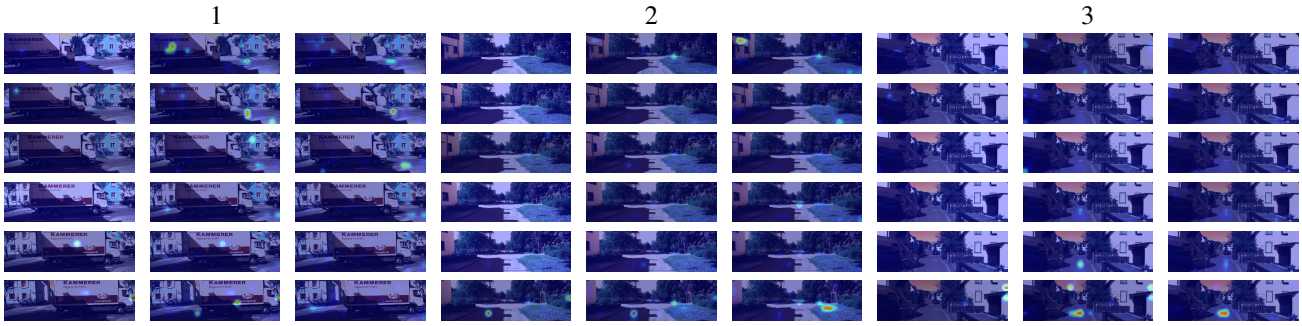
### 4.2. Multilabel Predictions

Since the model can output both 'Stop' and 'Go' as multilabel states in the same sample, various egomotion states can be predicted at significant events during braking or acceleration. This enables higher level interpretability in terms of the types of scenarios the framework can handle. Predictions can also indicate time-to-braking (or acceleration) in different scenarios depending on the timesteps in which labels appear, as well as the duration of braking or acceleration. Samples with labels containing both 'Stop' and 'Go' make up approximately 4% of the dataset. Approximately 45% of the labels are 'Stop'. Performance on these label predictions are shown in Table 3.

Based on the results, prediction of these types of labels are more difficult, partly due to the smaller proportion of these examples in the dataset, as well as potential noise in ground truth labels. Both models show similar relative performance differences as with the results in Table 1, but false 'Stop' predictions here are a significant problem in the case where only a single frame is used as input.

*Table 3.* Classification performance of our models with varying input video frame length on KITTI dataset on samples containing both 'Stop' and 'Go' labels.

| MODEL | F1 | PRECISION | RECALL |
|---|---|---|---|
| F = 1 | 0.781 | 0.660 | 0.868 |
| F = 6 | 0.770 | 0.768 | 0.742 |




(a) raw (b) mean (c) max (a) raw (b) mean (c) max (a) raw (b) mean (c) max

*Figure 2.* Comparison of visualization techniques for interpreting model attention (with respect to the classification token) on 3 example video clips. (a) Raw attention of the last layer. (b) Rollout attention using mean of heads (Abnar & Zuidema, 2020). (c) Rollout attention using maximum of heads (Gildenblat, 2020).

## 5. Interpretability of Imitation Learning Model

To obtain explanations for parts of video clips that contribute to the prediction, attention maps from the Transformers are used. We define a Transformer block $b$, starting from the output block at layer 1, and input block at layer $B$. The Transformer input sequence of tokens has length $n$, with token dimension of size $d_k$. A pairwise attention map for each head is obtained from the softmax dot product of the following:

$$\mathbf{A}^{(b)} = softmax(\frac{\mathbf{Q}^{(b)} \cdot \mathbf{K}^{(b)^T}}{\sqrt{d_k}}) \tag{1}$$

where $\mathbf{A}^{(b)}$ is the pairwise attention map at block $b$, and $\mathbf{Q}^{(b)}, \mathbf{K}^{(b)} \in R^{h \times n \times d_k}$ are the query and key matrices.

Attention between the `[cls]` token and patch tokens are relevant to the output, but this alone is not adequate for interpretability of the model. Rollout attention (Abnar & Zuidema, 2020) is used as a simple method of quantifying the propagation of information flow in terms of the attention maps in each layer. Following their method, we account for the residual sum using the identity matrix and calculate rollout attention from the input layer to the output layer as follows:

$$\hat{\mathbf{A}}^{(b)} = I + \mathbf{A}_{\text{mean}} \tag{2}$$

$$rollout = \hat{\mathbf{A}}^{(1)} \cdot \hat{\mathbf{A}}^{(2)} \cdot \ldots \cdot \hat{\mathbf{A}}^{(B)} \tag{3}$$

where $A_{\text{mean}}$ is the attention map from calculating the mean across all heads,

It was observed in (Gildenblat, 2020) that obtaining the maximum scores from all heads (rather than the mean attention $A_{\text{mean}}$) and zeroing a percentage of the lowest scores in the combined attention map $A_{\max}$ before calculating rollout attention resulted in a more informative visualization with respect to what the model focuses on in the image classification task. Hence, we show salient heatmaps over example input video clips using this approach to observe the practicality of this form of explainability in imitation learning models. The first row of the rollout attention map (excluding the first element) is extracted and reshaped to produce a heatmap over an observed video clip. We show results of this max rollout in Figure 2, including visualizations of the raw attention map in the last layer, and visualizations of mean rollout as proposed in (Abnar & Zuidema, 2020).

Interpreting the model using the raw attention map is more challenging compared to max rollout or mean rollout. Resultant heatmaps show sparser visual activity, while also providing a limited view of how the model as a whole interprets input clips. Rollout attention shows a more informative view of information propagation across the model using attention maps. However, max rollout visualizations appear less noisy, and is the able to highlight stronger attention on the vehicle in the second example clip. Features that contribute to the prediction output can be interpreted from these examples, and provides a tool for understanding potential objects that are useful for making predictions. However, a popular visualization method such as rollout is still lacking in explaining the abstract causal process that leads to predictions from raw pixels.

## 6. Conclusion

In this work, we have shown that our approach to end-to-end imitation learning when applied to autonomous driving is simple enough to be applied to a wide variety of driving scenarios for safe motion planning. In addition, we propose the application of Vision Transformers as a useful tool for modelling interpretable spatiotemporal features for driving within fixed video context observations. While it is possible to implement this framework as an active learning system using a dashcam and speedometer readings, it is still reliant on accurate egomotion data that may pose a challenge to training on some data.

## References


Abnar, S. and Zuidema, W. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

Argall, B. D., Chernova, S., Veloso, M., and Browning, B. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

Bertasius, G., Wang, H., and Torresani, L. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229. Springer, 2020.

Chefer, H., Gur, S., and Wolf, L. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 782–791, 2021.

Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., and Gao, W. Pre-trained image processing transformer. *arXiv preprint arXiv:2012.00364*, 2020.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., and Zisserman, A. Counting out time: Class agnostic video repetition counting in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

Gildenblat, J. Exploring explainability for vision transformers, December 2020. URL https://jacobgil.github.io/deeplearning/vision-transformer-explainability#summary.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jabri, A., Owens, A., and Efros, A. A. Space-time correspondence as a contrastive random walk. *arXiv preprint arXiv:2006.14613*, 2020.

Kim, J. and Canny, J. Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of the IEEE international conference on computer vision*, pp. 2942–2950, 2017.

Mittelstadt, B., Russell, C., and Wachter, S. Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 279–288, 2019.

Pomerleau, D. A. Alvinn: An autonomous land vehicle in a neural network. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA ARTIFICIAL INTELLIGENCE AND PSYCHOLOGY . . . , 1989.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2446–2454, 2020.

Turc, I., Chang, M.-W., Lee, K., and Toutanova, K. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *NIPS*, 2017.

Xu, H., Gao, Y., Yu, F., and Darrell, T. End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2174–2182, 2017.

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., and Darrell, T. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Yurtsever, E., Lambert, J., Carballo, A., and Takeda, K. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8:58443–58469, 2020.