SUPPLEMENTARY MATERIAL
Importance-Guided Interpretability and Pruning for Video Transformers in Driver Action Recognition

A. Ablation Study on Importance Metric

To determine the most effective metric for quantifying layer relevance, we conducted an ablation study comparing JSD, ΔProb, and the R score, which is the average of the first two metrics. As shown in Tab. 1, the accuracy values across different cases, particularly in the DAI dataset, are quite similar regardless of the metric used. This consistency is especially evident when pruning high-relevance layers, where all three metrics yield nearly identical accuracy scores across different datasets and architectures. This is because all three metrics effectively identify and agree on the high-relevance layers. However, a slight difference emerges when removing low-relevance layers: JSD tends to perform the worst, while the R score delivers the best results, closely followed by the ΔProb metric. These findings suggest that while all metrics are reliable for identifying critical layers, the R score offers a slight edge in preserving accuracy when pruning less important layers.

Model	Pruned Layer Relevance	Importance Metric	UCF101	DAI
TimeSformer (3 layers pruned)	Lowest	JSD	57.88	52.75
		ΔProb	58.21	52.75
		R	58.43	52.75
	Highest	JSD	1.24	2.17
		ΔProb	1.24	2.17
		R	1.24	2.17
Video Swin (6 layers pruned)	Lowest	JSD	83.08	88.04
		ΔProb	85.12	88.04
		R	86.07	88.04
	Highest	JSD	4.62	3.26
		ΔProb	4.62	3.26
		R	4.62	3.26
MViT (5 layers pruned)	Lowest	JSD	90.11	86.15
		ΔProb	91.11	86.96
		R	91.11	86.96
	Highest	JSD	30.11	25.00
		ΔProb	28.07	25.00
		R	28.07	25.00

Table 1. Comparison of Top-1 accuracy outcomes when pruning layers based on different importance metrics (JSD, ΔProb, and R score) across various datasets and architectures.

B. Specialized Head Generalization

Within the DAI dataset, we also identify layer 22, head 4 as important due to its consistently high relevance score across all test samples. This head consistently attends to the driver's head, regardless of the video class (see first row in Fig. 1). Interestingly, the attention patterns of this head can generalize across datasets. For instance, in UCF101, layer 22, head 7 is the most important for the "PlayingViolin" class, as it seeks violins (see lower part, second column in Fig. 1). However, when we apply the model trained on DAI, with layer 22, head 4 specialized in focusing on the head of the person, it successfully generalizes (see lower part, last column in Fig. 1).

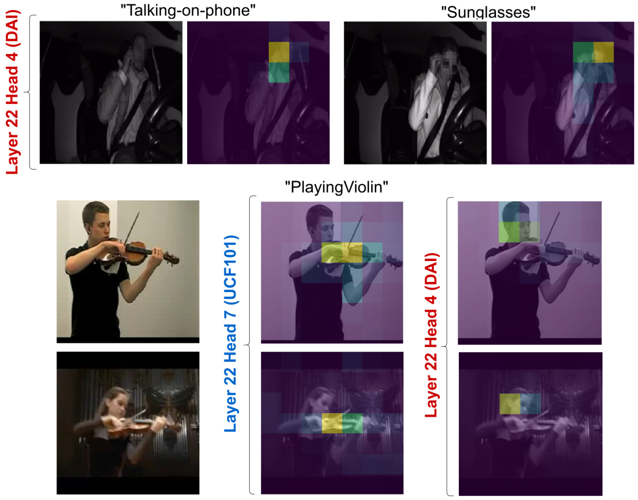

Figure 1. Attention maps of a specialized head using Video Swin that generalizes across datasets. In the upper part (first row), a specialized head on DAI consistently focuses on the driver's head, regardless of the video class. Additionally, in the lower part, second column, we show a head trained on UCF101 to detect violins. For the same violin samples, the DAI-trained head successfully generalizes by detecting the person's head, as shown in the lower part, last column. Colormap: viridis (dark blue: low attention, bright yellow: high attention). Best viewed in color.