Supplementary Material for AIDE: A Vision-Driven Multi-View, Multi-Modal, Multi-Tasking Dataset for Assistive Driving Perception

In this Supplementary Material, we will further detail the following aspects omitted in the main paper. Here, DER, DBR, TCR, and VCR stand for Driver Emotion Recognition, Driver Behavior Recognition, Traffic Context Recognition, and Vehicle Condition Recognition, respectively.

1. Coarse-grained Evaluation Taxonomy

Considering the demand for practicality [3] in Driver Monitoring Systems (DMS), we provide three-category evaluations of polarity emotions and two-category evaluations of anomaly behaviors. Specifically, we convert the main fine-grained taxonomies of driver emotion and behavior evaluations into coarse-grained evaluation taxonomies. The coarse-grained emotion evaluation [4] usually involves *positive*, *neutral*, and *negative* emotions. Meanwhile, the coarse-grained behavior evaluation considers mainly *normal* driving and diverse secondary (*i.e.*, *abnormal*) behaviors that are likely to cause traffic accidents. The new taxonomies are shown in Table 1. The coarse-grained accuracy (CG-Acc) and the F1 score (CG-F1) results refer to the comparison experiments of *Table 2 in the main paper*.

Table 1. The fine- to coarse-grained taxonomies on polarity emotions and anomaly behaviors. "CG" and "FG" stand for coarse-grained and fine-grained, respectively.

Evaluation Guidelines	CG-Taxonomy	FG-Taxonomy
Driver Emotion Recognition Task		
Polarity Emotion	Positive	Happiness
	Neutral	Peace
	Negative	Anxiety, Weariness, Anger
Driver Behavior Recognition Task		
Anomaly Behavior	Normal	Normal Driving
	Abnormal	Smoking, Making Phone, Looking Around, Dozing Off, Talking, Body Movement

2. Spatio-temporal Embedding

In the 2D and 2D + Timing patterns, we add spatial and temporal embeddings to maintain the necessary spatio-temporal correlations for networks dealing with skeleton-based modalities, respectively. Here, we present the details of both embeddings.

Table 2. Experimental results of spatio-temporal embedding necessity. "2DT" means "2D + Timing" pattern. "w/" and "w/o" are short for with and without, respectively. "SE" and "TE" stand for spatial embedding and temporal embedding, respectively.

Pattern	MLP Configuration	DER	DBR	TCR	VCR
		Acc	Acc	Acc	Acc
2D	(2) MLPs w/ SE	**71.26**	**65.35**	**83.74**	**77.12**
	(2) MLPs w/o SE	70.14	64.57	83.69	**77.14**
	(4) MLPs w/ SE	**70.72**	**63.65**	**82.77**	**77.94**
	(4) MLPs w/o SE	70.05	63.44	82.74	77.82
2DT	(7) MLPs w/ TE	**72.65**	**67.08**	**86.63**	**78.46**
	(7) MLPs w/o TE	71.83	66.61	**86.75**	**78.49**
	(9) MLPs w/ TE	**71.12**	**67.15**	**85.13**	**78.58**
	(9) MLPs w/o TE	70.43	66.74	85.11	78.50

Spatial Embedding. We start by constructing a Tensor variable that is consistent with a specific input shape. The values of this Tensor are initialized by order of topological connections among the skeleton points to capture spatial correlations. A linear layer is then introduced to project the dimensions of the Tensor variable three times to match the three channels of each keypoint, including the x-axis coordinate, the y-axis coordinate, and the confidence value. After that, the projected variable is added to the skeleton-based inputs to learn the feature representation together via a Multi-Layer Perceptron (MLP).

Temporal Embedding. Similarly, we first construct a Tensor variable of the same size as the number of frames. The values of this Tensor are initialized by order of the input frames to preserve temporal dependencies. Immediately, the Tensor variable is added to the input stream for each frame along the time dimension to learn the feature representation together via a specific MLP.

Experimental Analysis. To verify the necessity of spatio-temporal embedding, we choose two model combinations in the 2D and 2D + Timing pattern frameworks (Experiments 2, 4, 7, 9 *in Table 2 of the main paper*) to perform the ablation studies, respectively. In this case, the corresponding spatio-temporal embeddings in the network models of the gesture and posture streams are removed to observe the

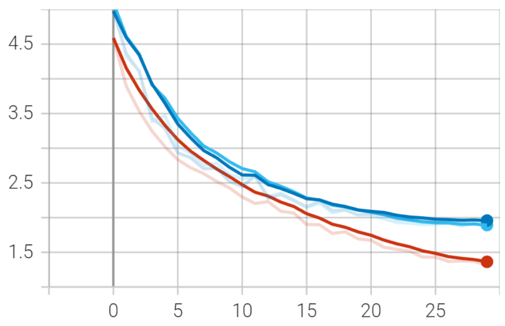

Figure 1. Visualization convergence of the training set (red), validation set (blue), and testing set (cyan) losses.

performance variations of the four perception tasks. From Table 2, the model combinations without the Spatial Embedding (SE) and Temporal Embedding (TE) versions show a significant deterioration in performance, especially on the DER and DBR tasks. This inevitable phenomenon offers two insights. **(i)** The skeleton-based body pose information keeps the human interpretability compared to the intermediate feature maps, which are indispensable for understanding driver states. **(ii)** The proposed spatio-temporal embedding strategies preserve spatial correlations and temporal dynamics in gesture and posture streams, providing informative semantic clues for the DER and DBR tasks.

3. Visualization of Loss Convergence

As a multi-stream architecture, evaluating the convergence of an AIDE-based framework is essential for the DMS. We show in Figure 1 the convergence of different losses for the resource-efficient model combination (13) from *Table 2 in the main paper*. We find that the training, validation, and testing set losses all converge smoothly and then stabilize. This observation suggests that the proposed model framework is easy to optimize and workable.

4. Effect of Pre-training

Table 3 explores the performance gains of our framework for different pre-training strategies. We select diverse model combinations (1, 6, 15) from *Table 2 of the main paper* to provide the following observations. **(i)** First, we evaluate the pre-training effect of the scene stream model. In Experiments (1, 6) of the main paper, we have shown that the pre-trained backbone on a large-scale Places365 dataset [5] can significantly improve the TCR and VCR tasks. A reasonable reason is that the pre-trained backbone contains rich semantic prototypes of scenes to facilitate more effective context representation learning. Here, we provide the vanilla and pre-trained versions on ImageNet [1] of the scene stream model for Experiments (1, 6). As shown in the upper part of Table 3, the scratch-trained backbones show the worst results on the TCR and VCR tasks. The pre-trained backbones on Places365 bring better gains than the improvements from pre-training on ImageNet. These findings suggest that feature representations based on scene semantics are more expressive than those based on object semantics. **(ii)** Furthermore, we provide pre-training results in the 3D pattern via the Holistic Video Understanding (HVU) dataset [2]. We select Experiment (15) to perform the new experiment, where the 3D-CNN backbones of the face, body, and scene streams are pre-trained on HVU. The results in the bottom part of Table 3 show that HVU enables promising improvements in driving monitoring performance for the four tasks, benefiting from its comprehensive properties related to scenes, objects, actions, events, attributes, and concepts.

Table 3. Experimental results for pre-trained backbones from different streams. “w/” and “w/o” are short for the with and without, respectively. “S”, “F”, and “B” represent the scene, face, and body, respectively.

ID	Stream	Config	DER	DBR	TCR	VCR
			F1	F1	F1	F1
(1)	S	w/ Places365 Pre-training	**63.06**	**59.52**	**86.63**	**77.27**
	S	w/ ImageNet Pre-training	63.02	**59.55**	86.24	76.86
	S	w/o Pre-training	**63.12**	59.49	85.86	76.15
(6)	S	w/ Places365 Pre-training	**67.14**	**64.45**	**89.66**	**77.94**
	S	w/ ImageNet Pre-training	**67.18**	64.43	89.05	77.21
	S	w/o Pre-training	67.05	**64.53**	88.19	76.77
(15)	S + F + B	w/ HVU Pre-training	**65.63**	**65.41**	**88.37**	**80.53**
	S + F + B	w/o Pre-training	65.15	64.57	87.26	79.71

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 2

[2] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large scale holistic video understanding. In *European Conference on Computer Vision (ECCV)*, pages 593–610, 2020. 2

[3] Iuliia Kotseruba and John K Tsotsos. Attention for vision-based assistive and automated driving: A review of algorithms and datasets. *IEEE Transactions on Intelligent Transportation Systems*, 2022. 1

[4] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016. 1

[5] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017. 2