Supplementary Material: TIP: Tabular-Image Pretraining for Multimodal Classification with Incomplete Data

Siyi Du*, Shaoming Zheng, Yinsong Wang, Wenjia Bai, Declan P. O'Regan, and Chen Qin*

Imperial College London, London, UK
{s.du23,s.zheng22,y.wang23,w.bai,declan.oregan,c.qin15}@imperial.ac.uk

Broader Societal Impact: Our proposed TIP, which is optimized using statistical techniques, could potentially perpetuate biases and unfairness present in the training data. For instance, the image-tabular data in the UK Biobank database [4] is mostly collected from white population and healthy subjects [2, 20]. Previous studies have found that deep learning models could potentially learn spurious correlations between cardiac diseases and population characteristics and may result in negative societal impacts when generalizing to other populations [11, 15]. Therefore, further research and deployment based on this model should take into account these issues, along with potential solutions to address biases and unfairness.

A Detailed Data Description

The UK Biobank (UKBB) [4] is used for two cardiovascular disease classification (diagnosis) tasks: coronary artery disease (CAD) and myocardial infarction (Infarction). This dataset comprises 36,167 subjects. For each subject, we utilized mid-ventricle slices of its cardiac magnetic resonance (MR) images at three time phases, *i.e.*, end-systolic (ES) frame, end-diastolic (ED) frame, and a time frame between ED and ES. In addition, we employed 75 tabular features, including 26 categorical features, *e.g.*, alcohol drinker status, and 49 continuous features, *e.g.*, average heart rate, and their detailed information can be found in Tab. 1.

Moreover, we have used a natural image dataset, Data Visual Marketing (DVM) [12], for a car model classification task with 283 classes. We employed 176,414 image-tabular samples from this dataset. As illustrated in Tab. 2, each sample has 17 tabular features in total, including 4 categorical and 13 continuous features. For both UKBB and DVM datasets, we pre-processed their tabular data before using them for our task as in [9]. Additionally, we converted categorical data into ordinal numbers and standardized continuous data using z-score normalization, with a mean value of 0 and standard derivation of 1.

*Corresponding authors.

Table 1: 75 tabular features (26 categorical and 49 continuous) are employed for CAD and Infarction tasks on UKBB. **Cat** denotes whether the feature is categorical, and N_{unq} represents the total number of unique values for each categorical feature.

Tabular Feature	**Cat**	N_{unq}	**Tabular Feature**	**Cat**	N_{unq}
Alcohol drinker status	√	3	LVCO (L/min)	×	-
Alcohol intake frequency	√	6	LVEDV (mL)	×	-
Angina diagnosed by doctor	√	2	LVEF (%)	×	-
Augmentation index for PWA	×	-	LVESV (mL)	×	-
Average heart rate	×	-	LVM (g)	×	-
Basal metabolic rate	×	-	LVSV (mL)	×	-
Blood pressure medication regularly taken	√	2	Number of beats in waveform average for PWA	×	-
Body fat percentage	×	-	Number of days/week of moderate physical activity 10+ minutes	√	8
Body mass index (BMI)	×	-	Number of days/week of vigorous physical activity 10+ minutes	√	8
Body surface area	×	-	Number of days/week walked 10+ minutes	√	8
Cardiac index during PWA	×	-	Oral contraceptive pill or minipill medication regularly taken	√	2
Cardiac index	×	-	Overall health rating	√	4
Cardiac output during PWA	×	-	P duration	×	-
Cardiac output	×	-	Past tobacco smoking	√	4
Central augmentation pressure during PWA	×	-	Peripheral pulse pressure during PWA	×	-
Central pulse pressure during PWA	×	-	Pulse rate	×	-
Central systolic blood pressure during PWA	×	-	Pulse wave Arterial Stiffness index	×	-
Cholesterol lowering medication regularly taken	√	2	QRS duration	×	-
Current tobacco smoking	√	3	RVEDV (mL)	×	-
Diabetes diagnosis	√	2	RVEF (%)	×	-
Diastolic blood pressure	×	-	RVESV (mL)	×	-
Diastolic brachial blood pressure during PWA	×	-	RVSV (mL)	×	-
Duration of moderate activity	×	-	Sex	√	2
Duration of strenuous sports	√	8	Shortness of breath walking on level ground	√	2
Duration of vigorous activity	×	-	Sleep duration	×	-
Duration of walks	×	-	Sleeplessness / insomnia	√	3
End systolic pressure during PWA	×	-	Smoking status	√	3
End systolic pressure index during PWA	×	-	Stroke diagnosed by doctor	√	2
Ever smoked	√	8	Stroke volume during PWA	×	-
Exposure to tobacco smoke at home	×	-	Systolic blood pressure	×	-
Exposure to tobacco smoke outside home	×	-	Systolic brachial blood pressure during PWA	×	-
Falls in the last year	√	3	Total peripheral resistance during PWA	×	-
Heart rate during PWA	×	-	Usual walking pace	×	-
High blood pressure diagnosed by doctor	√	2	Ventricular rate	×	-
Hip circumference	×	-	Waist circumference	×	-
Hormone replacement therapy medication regularly taken	√	2	Weight	×	-
Insulin medication regularly taken	√	2	Whole body fat mass	×	-
Long-standing illness, disability or infirmity	√	2			

B Implementation Details

B.1 Pre-training

We followed the same image and tabular data augmentation techniques during pre-training as in [9]. Specifically, we augmented images through random scaling, rotation, shifting, flipping, Gaussian noise, as well as brightness, saturation, and contrastive changes. After that, all images are resized to 128×128. To speed up the image augmentation process, we used the Albumentations python library [3]. For tabular data augmentation of ITC and ITM, we randomly selected 30% of tabular features in each subject and replaced their values with column-wise randomly selected values. Notice that tabular pre-training algorithms (SCARF [1], VIME [24], and SAINT [18]) have their own tabular augmentations.

The hyper-parameters and training configurations for the self-supervised learning (SSL) image pre-training approaches (SimCLR [5], BYOL [8], SimSiam [6], BarlowTwins [26]) and SSL multimodal pre-training method (MMCL [9]) are the same as those used in [9], which were found using hyper-parameter search.

Table 2: 17 tabular features (4 categorical and 13 continuous) are used for the DVM car model classification task. **Cat** denotes whether the feature is categorical, and N_{unq} represents the total number of unique values for each categorical feature.

Tabular Feature	**Cat**	N_{unq}	**Tabular Feature**	**Cat**	N_{unq}
Advertisement month (Adv_month)	×	-	Height	×	-
Advertisement year (Adv_year)	×	-	Length	×	-
Bodytype	√	13	Price	×	-
Color	√	22	Registration year (Reg_year)	×	-
Number of doors (Door_num)	×	-	Miles runned (Runned_Miles)	×	-
Engine size (Engine_size)	×	-	Number of seats (Seat_num)	×	-
Entry prize (Entry_prize)	×	-	Wheelbase	×	-
Fuel type (Fuel_type)	√	12	Width	×	-
Gearbox	√	3			

We utilized optimal hyper-parameters to pre-train SSL tabular pre-training models and our proposed TIP. Specifically, for each model, we selected the best learning rate from a set of values of $\{3 \times 10^{-3}, 3 \times 10^{-4}, 3 \times 10^{-5}\}$ and the best weight decay from $\{1 \times 10^{-4}, 1.5 \times 10^{-6}\}$, based on its performance on the validation set. All models are deployed on 4 A6000 GPUs and pre-trained for 500 epochs using the Adam optimizer [13]. The learning rate is warmed up linearly for 10 epochs and decayed following a cosine annealing scheduler. The implementation details of TIP and SSL tabular pre-training methods are discussed below.

The Proposed TIP: It utilizes a learning rate of 3×10^{-4} and a weight decay of 1.5×10^{-6} for DVM pre-training and a learning rate of 3×10^{-4} and a weight decay of 1×10^{-4} for UKBB cardiac pre-training.

SCARF [1]: It applies contrastive learning to original tabular data and an augmented view by corrupting a random subset of features. Based on the validation performance, the corruption ratio and the temperature parameter are set to 0.3 and 0.1, respectively. The hidden dimension of SCARF's multi-layer perceptron (MLP) is 512. We utilized a learning rate of 3×10^{-4} and a weight decay of 1.5×10^{-6} for DVM pre-training and a learning rate of 3×10^{-3} and a weight decay of 1×10^{-4} for UKBB cardiac pre-training.

VIME [24]: It predicts the corrupted positions in tabular data and reconstructs their values. Based on the validation performance, the corruption ratio is set to 0.3. The reconstruction loss adjustment parameter α is 2.0 as in [24], and the hidden dimension of VIME's MLP is 512. For pre-training, we utilized a learning rate of 3×10^{-4} and a weight decay of 1.5×10^{-6} for DVM and a learning rate of 3×10^{-3} and a weight decay of 1×10^{-4} for UKBB cardiac dataset.

SAINT [18]: It produces an augmented tabular view through CutMix [25] and mixup [27] and then operates contrastive learning and denoising pre-training. Additionally, SCARF proposed a new transformer-based tabular architecture that performs attention across rows and columns. We utilized a learning rate of 3×10^{-5} and a weight decay of 1.5×10^{-6} for DVM pre-training and a learning rate of 3×10^{-5} and a weight decay of 1×10^{-4} for UKBB cardiac pre-training.

Table 3: Learning rates of DVM, CAD, and Infarction tasks for different models during supervised training or fine-tuning. ❄ means linear probing, and 🔥 represents fully fine-tuning.

Model	DVM Accuracy (%) ↑		CAD AUC (%) ↑		Infarction AUC (%) ↑	
	❄	🔥	❄	🔥	❄	🔥
(a) Supervised Methods						
ResNet-50 [10]	3×10^{-4}		1×10^{-3}		1×10^{-3}	
Concat Fuse (CF) [19]	3×10^{-4}		3×10^{-3}		3×10^{-3}	
Max Fuse (MF) [21]	3×10^{-4}		3×10^{-3}		3×10^{-3}	
Interact Fuse (IF) [7]	3×10^{-4}		3×10^{-3}		3×10^{-3}	
DAFT [23]	3×10^{-4}		3×10^{-3}		3×10^{-3}	
(b) SSL Pre-training Methods						
SimCLR [5]	1×10^{-3}	1×10^{-4}	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}
BYOL [8]	1×10^{-3}	1×10^{-4}	1×10^{-3}	1×10^{-4}	1×10^{-3}	1×10^{-4}
SimSiam [6]	1×10^{-3}	1×10^{-5}	1×10^{-3}	1×10^{-4}	1×10^{-3}	1×10^{-4}
BarlowTwins [26]	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}
SCARF [1]	1×10^{-4}	1×10^{-4}	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}
VIME [24]	1×10^{-4}	1×10^{-4}	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}
SAINT [18]	1×10^{-4}	1×10^{-5}	1×10^{-3}	1×10^{-5}	1×10^{-3}	1×10^{-5}
MMCL [9]	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}
TIP (proposed)	1×10^{-4}	1×10^{-4}	1×10^{-3}	1×10^{-4}	1×10^{-3}	1×10^{-4}

B.2 Fine-tuning

For either fully fine-tuning or linear probing settings of each pre-trained model, we chose the best learning rate from a set of values of $\{3 \times 10^{-2}, 1 \times 10^{-2}, 3 \times 10^{-3}, 1 \times 10^{-3}, 3 \times 10^{-4}, 1 \times 10^{-4}, 3 \times 10^{-5}, 1 \times 10^{-5}\}$ depending on its validation performance. Tab. 3(b) demonstrates the learning rate used for each model. We utilized an Adam optimizer without weight decay and a batch size of 512. To alleviate over-fitting, an early stopping strategy in Pytorch Lightning has been adopted, with a minimal delta (divergence threshold) of 0.0002, a maximal number of epochs of 500, and a patience (stopping threshold) of 10 epochs.

B.3 Supervised Training

We reproduced 1 supervised image approach, ResNet-50, and 4 supervised multimodal algorithms: concatenation fusion (CF) [19], maximum fusion (MF) [21], interactive fusion through channel-wise multiplication (IF) [7], and dynamic affine transform (DAFT) [23]. These multimodal techniques leverage a ResNet-50 as their image encoder for fair comparison. To adapt CF and MF to our task, we used a 2-layer MLP with a hidden dimension of 512 and an output dimension of 2048 as their tabular encoder. For IF, its tabular encoder is a 4-layer MLP, with hidden dimensions of [64, 256, 512, 1024] and an output dimension of 2048. We undertook the same training strategy and learning rate sweep as the fine-tuning process, *i.e.*, an early stopping strategy with a maximum of 500 epochs. We ensure that all supervised models were converged after training. Tab. 3(a) demonstrates the learning rate used for each model.

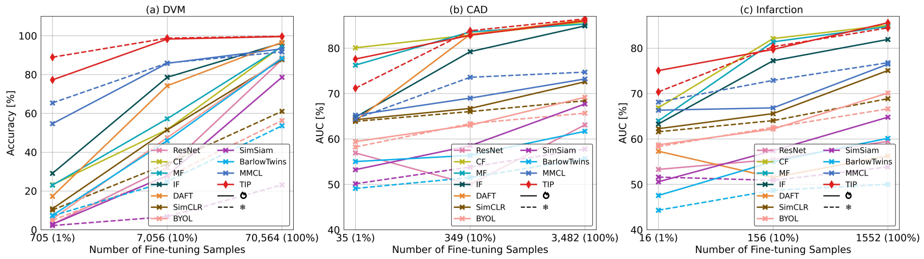

Fig. 1: Overall result comparison with supervised/SSL image/multimodal approaches on various number of fine-tuning samples. 🔥 denotes fully fine-tuning, and ❄ means linear probing. In addition to the results shown in Fig. 3 of the manuscript, we have included the results of BYOL, SimSiam, and BarlowTwins.

C Additional Experiment

C.1 Robustness to Low-data Regimes (Complete Results)

As mentioned in Sec. 4.1 of the manuscript, we propose to assess the performance of TIP and other SOTA methods on low-data regimes (10% and 1% of the original dataset size). Fig. 3 in the manuscript displays only SimCLR's results for SSL image approaches since it showed the best performance among them. We present the complete results of all models in Fig. 1.

C.2 Ablation Study on The Proposed SSL Strategy

We conduct an ablation study to analyze the impact of each SSL pre-training task: image-tabular contrastive learning (ITC), image-tabular matching (ITM), and masked tabular reconstruction (MTR). Tab. 4 demonstrates the results on complete downstream task data. We can obtain the following observations: (1) Compared with supervised TIP w/o any pre-training tasks, adding pre-training tasks improves the model performance, indicating the usefulness of our pre-training strategy. (2) In the linear probing setting, compared with integrated TIP, removing any of our pre-training tasks significantly decreases the model performance, *e.g.*, TIP w/o ITC decreases AUC by 14.08% on Infarction, TIP w/o ITM decreases AUC by 1.61% on CAD, and TIP w/o MTR decreases AUC by 2% on CAD. This indicates that our three pre-training tasks enable the model to learn transferable features and efficiently produce promising results with a few tunable parameters. The competitive results of TIP and TIP w/o ITM or ITC in fully fine-tuning can be attributed to the relatively small pre-training datasets and also the fact that tuning all parameters can moderately alleviate the reliance on integrating all three pre-training tasks.

Moreover, as mentioned in Sec. 4.3 of the manuscript, we study the performance of TIP with and without our SSL pre-training strategy when encountering incomplete downstream task data. Fig. 5 in the manuscript only illustrates the

Table 4: Ablation study of TIP's SSL pre-training tasks on complete data. ❄ means linear probing, and 🔥 represents fully fine-tuning. TIP w/o pre-training (1st row) is trained in a supervised manner, *i.e.*, all of its parameters are trainable in both ❄ and 🔥 columns.

ITC	ITM	MTR	DVM Accuracy (%) ↑		CAD AUC (%) ↑		Infarction AUC (%) ↑	
			❄	🔥	❄	🔥	❄	🔥
			98.57	98.57	86.04	86.04	84.19	84.19
	√	√	98.84	99.14	76.51	**86.89**	70.38	85.72
√		√	99.71	99.53	84.82	86.22	83.71	**85.89**
√	√		99.70	99.56	84.43	86.11	82.91	85.78
√	√	√	**99.72**	**99.56**	**86.43**	86.03	**84.46**	85.58

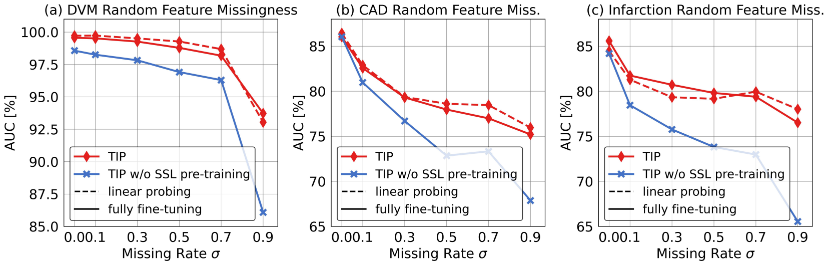

Fig. 2: Results of DVM, CAD, and Infarction tasks comparing TIP with or without the proposed SSL pre-training in the random feature missingness (RFM) scenario with different missing rates. In addition to the DVM's and CAD's results shown in Fig. 5 of the manuscript, we have included the results on Infarction.

results on DVM and CAD due to page limitations. We present the complete results of DVM, CAD, and Infarction in Fig. 2. We observe that our pre-training task enhances the model robustness to missing data across various missing rates on DVM, CAD, and Infarction tasks.

C.3 Effect of TIP's Tabular Encoder

We examine the contributions of our proposed transformer-based tabular encoder to the performance increase compared to supervised multimodal methods. Specifically, we replaced the MLP-based tabular encoder in the supervised multimodal methods (CF and MF) with the tabular encoder from TIP and conducted incomplete data experiments on the DVM classification task. As shown in Fig. 6, TIP's tabular encoder can improve the performance of supervised multimodal methods. However, these methods still lag behind TIP, especially in a high missing rate condition, demonstrating the efficacy of other components of TIP.

C.4 Sensitivity of Masking Ratio

As mentioned in Sec. 4.3 of the manuscript, we evaluated the effect of different masking ratios of the MTR pre-training task. In addition to Tab. 4 in

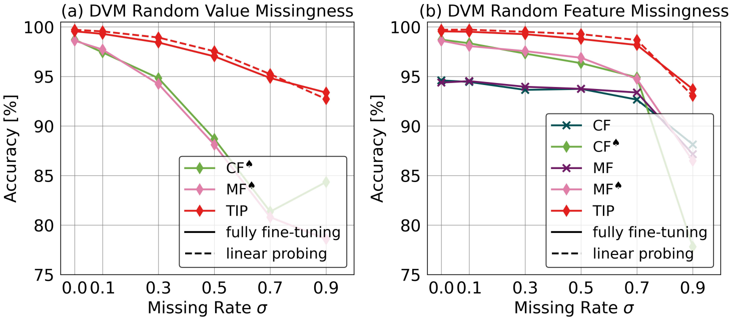

Fig. 3: Results comparing supervised multimodal methods and TIP on the DVM random value missingness (RVM) and random feature missingness (RFM) scenarios. ♠ means using TIP's tabular encoder.

Table 5: TIP's RMSE results on the DVM and UKBB test sets for reconstruction of missing continuous features. σ denotes data missing rate in fine-tuning and inference, and ρ means masking ratio of the MTR pre-training task.

Model	DVM RMSE ↓			UKBB RMSE ↓		
Missing rate σ	0.3	0.5	0.7	0.3	0.5	0.7
$\rho = 0.1$	0.5349	0.6752	0.7871	0.6245	0.6903	0.7851
$\rho = 0.3$	0.4110	0.5128	0.5924	0.6044	0.6538	0.7469
$\rho = 0.5$	**0.3899**	0.4651	0.5055	0.6039	0.6460	0.7106
$\rho = 0.7$	0.3986	**0.4612**	**0.4733**	**0.5963**	**0.6171**	**0.6654**
$\rho = 0.9$	0.4279	0.4800	0.4816	0.6542	0.6696	0.6791

the manuscript, we present the results of missing value reconstruction on two datasets in Tab. 5 and conducted experiments on the DVM classification task using diverse masking ratios (Fig. 4). As shown in Tab. 5, $\rho \in (0.5, 0.7)$ achieve the best reconstruction performance, whereas too high (0.9) or too low (0.1) masking ratios adversely affect model learning. In addition, the higher RMSE in UKBB than that in DVM indicates that there could be some outliers in UKBB. However, compared with SOTA data imputation methods in Tab. 2 of the manuscript, our TIP still achieves the best performance.

As displyed in Fig. 4, TIP has fairly consistent results in data missing scenarios across masking ratios, and moderate ratios (0.3, 0.5, 0.7) are better than extreme ones (0.1, 0.9). The sensitivity of ρ in fully fine-tuning is smaller than that in linear probing. This may be because tuning all the parameters mitigates the reliance on optimal masking ratios.

C.5 Visualization

As mentioned in Sec. 4.3 of the manuscript, we illustrate TIP's attention to tabular features when predicting a specific class in downstream tasks. Fig. 5 shows complete attention scores to all tabular features in downstream CAD task.

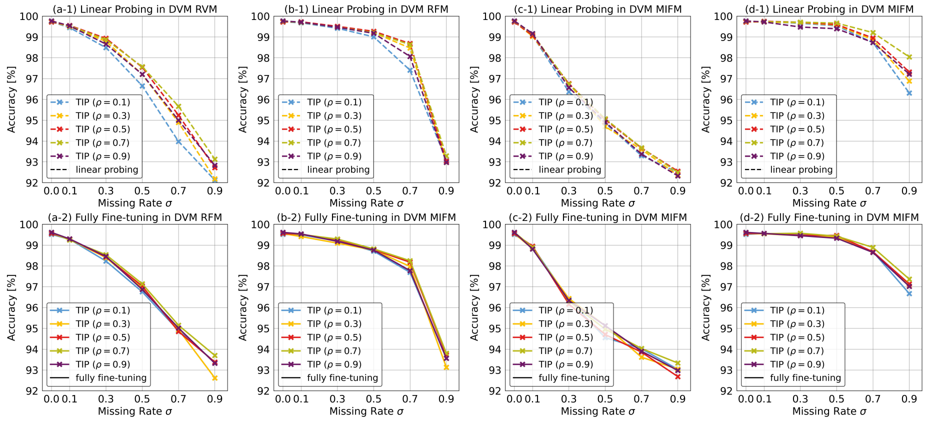

Fig. 4: Results of TIP with different masking ratios ρ on 4 DVM's missing data scenarios: (a) random value missingness (RVM), (b) random feature missingness (RFM), (c) most important feature missingness (MIFM), and (d) least important feature missingness (LIFM). We evaluated linear probing (1st row) and fully fine-tuning (2nd row). Notice that ρ is the masking ratio of the MTR pre-training task, while σ is the missing rate of missing data scenarios.

Our observations are as follows: (1) TIP distinguishes important imaging phenotypes, *e.g.*, the model attends more to left ventricle myocardial mass (LVM) and left ventricle end-diastolic volume (LVEDV) than left ventricle ejection fraction (LVEF), which is consistent with previous cardiac disease studies [2]. (2) TIP attends to critical non-imaging risk factors, *e.g.*, obesity-related features such as waist circumference and whole body fact mass [17, 22]. (3) TIP focuses more on physical measurements, *e.g.*, weight, body fat percentage, and blood pressure. These measurements have demonstrated high correlations with the left ventricular function, which plays an important role in CAD diagnosis [2].

Furthermore, we computed Grad-CAM [14,16] on the cross-attention maps in the 2nd layer of TIP's multimodal interaction module and generated per-token visualization. As displayed in Fig. 6, TIP does not only identify the classification object, but also captures inter-modality relations, *e.g.*, the 'Bodytype' token attends to the entire car, whereas the 'Wheelbase' token mainly focuses on the wheels. This showcases the effectiveness of our multimodal interaction module.

We also visualize some challenging cases of the DVM classification task where TIP still outperforms supervised/SSL image/multimodal algorithms in Fig. 7. The results demonstrate that a single image modality may not provide sufficient information for decision-making, whereas TIP can effectively integrate multimodal information to enhance model performance.

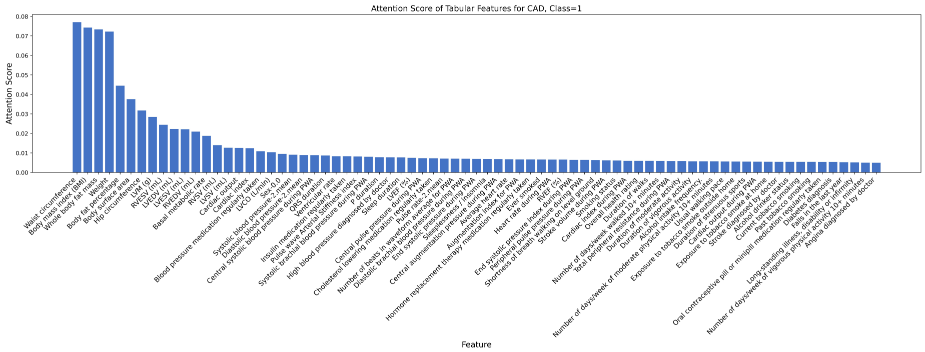

Fig. 5: The [CLS] token's attention scores to tabular features for the True class in the CAD task from the last layer of TIP' tabular encoder.

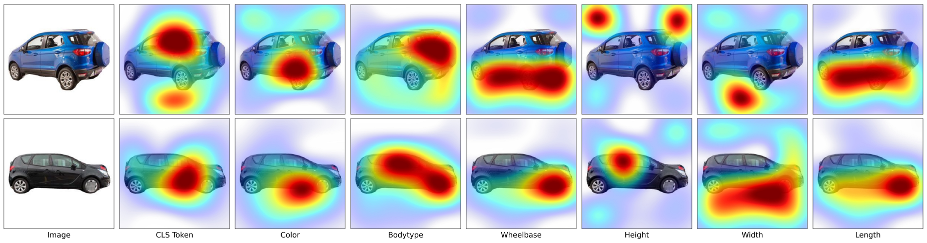

Fig. 6: Grad-CAM visualization on the cross-attention map in the 2nd layer of TIP's multimodal interaction module.

References

1. Bahri, D., Jiang, H., Tay, Y., Metzler, D.: SCARF: Self-supervised contrastive learning using random feature corruption. In: ICLR (2022)
2. Bai, W., Suzuki, H., Huang, J., Francis, C., Wang, S., Tarroni, G., et al.: A population-based phenome-wide association study of cardiac and aortic structure and function. Nature Medicine **26**(10), 1654–1662 (2020)
3. Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A.: Albumentations: Fast and flexible image augmentations. Information **11**(2) (2020)
4. Bycroft, C., Freeman, C., Petkova, D., et al.: The UK Biobank resource with deep phenotyping and genomic data. Nature **562**(7726), 203–209 (2018)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607. PMLR (2020)
6. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR. pp. 15750–15758 (2021)
7. Duanmu, H., Huang, P.B., Brahmavar, S., Lin, S., Ren, T., Kong, J., Wang, F., Duong, T.Q.: Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using deep learning with integrative imaging, molecular and demographic data. In: MICCAI. pp. 242–252. Springer (2020)

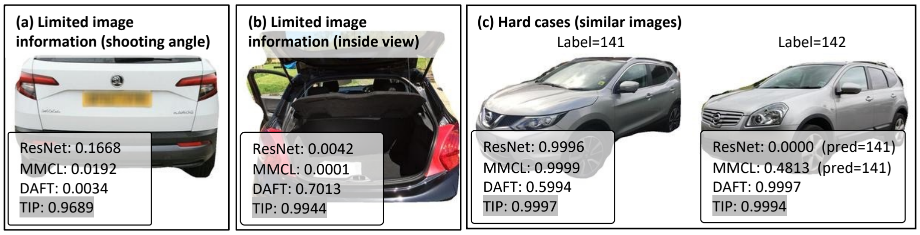

Fig. 7: DVM car visualization of samples and ground-truth class's predictions of TIP and other methods.

8. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., et al.: Bootstrap your own latent-a new approach to self-supervised learning. NIPS **33**, 21271–21284 (2020)
9. Hager, P., Menten, M.J., Rueckert, D.: Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In: CVPR. pp. 23924–23935 (2023)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
11. Ho, F.K., Gray, S.R., Welsh, P., Gill, J.M., Sattar, N., Pell, J.P., Celis-Morales, C.: Ethnic differences in cardiovascular risk: examining differential exposure and susceptibility to risk factors. BMC Medicine **20**(1), 149 (2022)
12. Huang, J., Chen, B., Luo, L., et al.: DVM-CAR: A large-scale automotive dataset for visual marketing research and applications. In: 2022 IEEE International Conference on Big Data (Big Data). pp. 4140–4147. IEEE (2022)
13. Kingma, D.P., Ba, J.: ADAM: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Li, J., Selvaraju, R., Gotmare, A., et al.: Align before fuse: Vision and language representation learning with momentum distillation. NIPS **34**, 9694–9705 (2021)
15. Nayak, A., Hicks, A.J., Morris, A.A.: Understanding the complexity of heart failure risk and treatment in black patients. Circulation: Heart Failure **13**(8), e007264 (2020)
16. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: ICCV. pp. 618–626 (2017)
17. Sniderman, A.D., Thanassoulis, G., Glavinovic, T., Navar, A.M., Pencina, M., Catapano, A., Ference, B.A.: Apolipoprotein B particles and cardiovascular disease: a narrative review. JAMA Cardiology **4**(12), 1287–1295 (2019)
18. Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C.B., Goldstein, T.: SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training. arXiv preprint arXiv:2106.01342 (2021)
19. Spasov, S., Passamonti, L., Duggento, A., Lio, P., Toschi, N., et al.: A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to alzheimer's disease. NeuroImage **189**, 276–287 (2019)
20. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al.: UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Medicine **12**(3), e1001779 (2015)

21. Vale-Silva, L.A., Rohr, K.: Long-term cancer survival prediction using multimodal deep learning. Scientific Reports **11**(1), 13505 (2021)
22. Wilkins, J.T., Li, R.C., Sniderman, A., Chan, C., Lloyd-Jones, D.M.: Discordance between apolipoprotein b and ldl-cholesterol in young adults predicts coronary artery calcification: the cardia study. Journal of the American College of Cardiology **67**(2), 193–201 (2016)
23. Wolf, T.N., Pölsterl, S., et al.: DAFT: a universal module to interweave tabular data and 3D images in CNNs. NeuroImage **260**, 119505 (2022)
24. Yoon, J., Zhang, Y., et al.: VIME: Extending the success of self-and semi-supervised learning to tabular domain. NIPS **33**, 11033–11043 (2020)
25. Yun, S., Han, D., Oh, S.J., et al.: CutMix: Regularization strategy to train strong classifiers with localizable features. In: ICCV. pp. 6023–6032 (2019)
26. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow Twins: Self-supervised learning via redundancy reduction. In: ICML. pp. 12310–12320. PMLR (2021)
27. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: ICLR (2018)