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

Siyi Du

Shaoming Zheng

Yinsong Wang

Wenjia Bai

Declan P. O'Regan

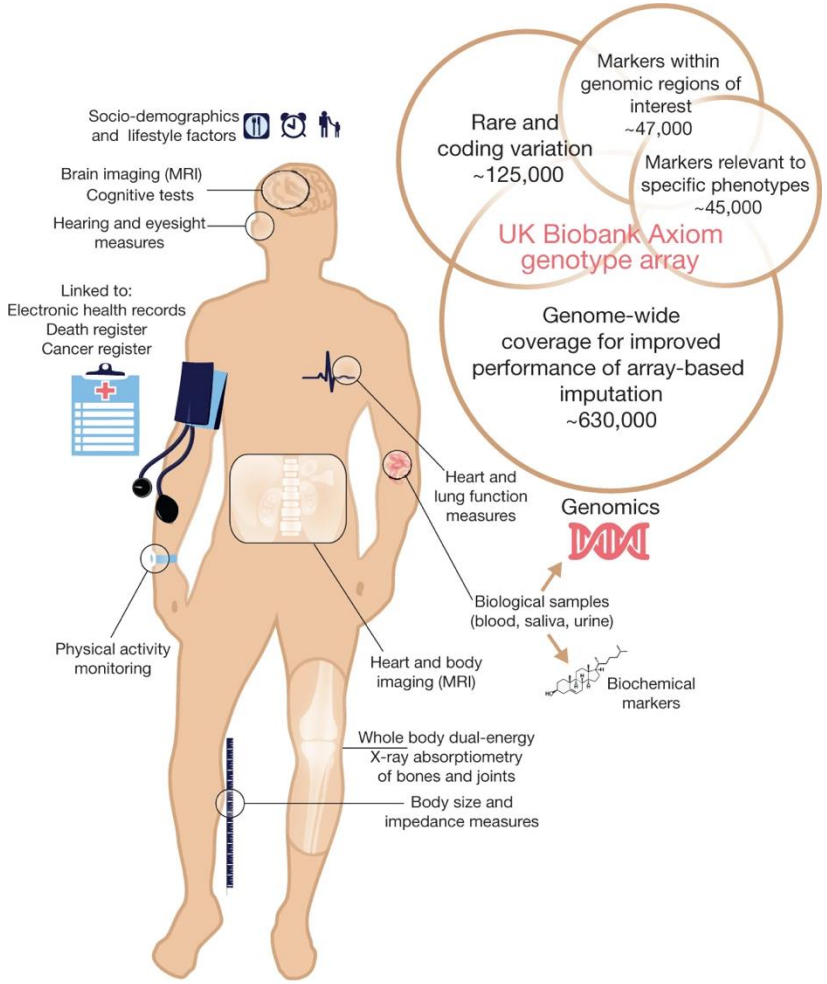
Chen Qin

18th European Conference on Computer Vision ECCV 2024

Milan, Italy

September 29 , 2024

1.1 Image-Tabular Representation Learning



UK Biobank [1]

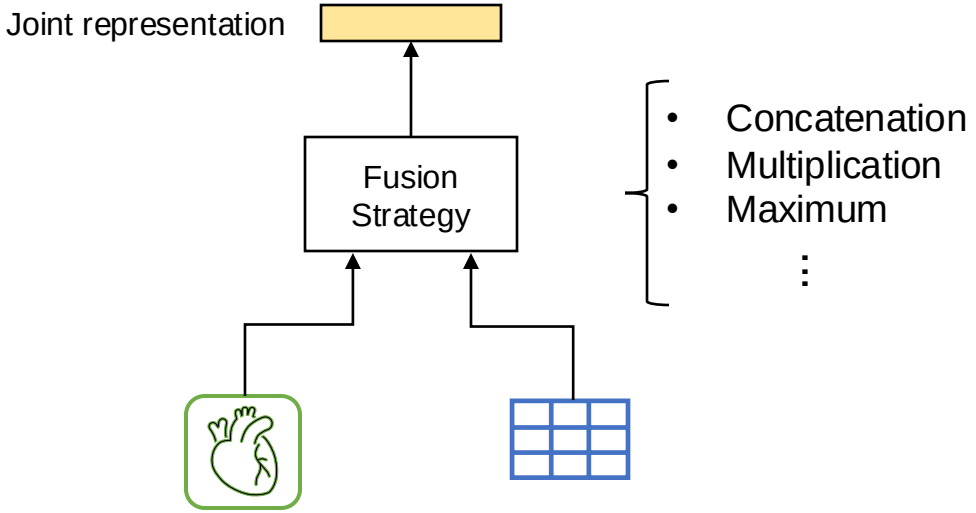
- Structured tables are increasingly available in real-world multimodal datasets.
- Integrating tabular data is crucial in various applications.
- Developing image-tabular representation learning methods is receiving more and more attention.

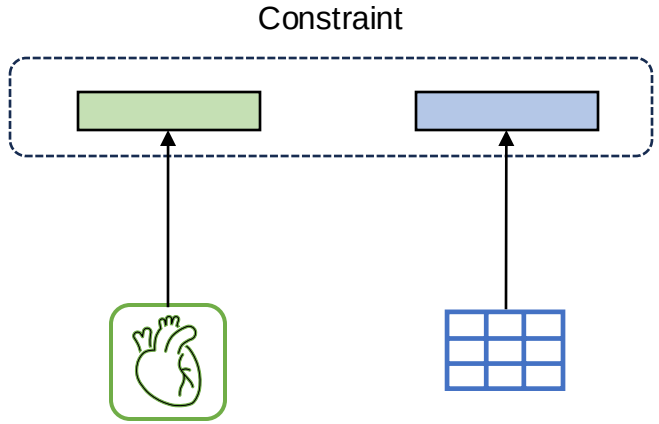
Sex	Alcohol drinking	Diabetes diagnosis	…	Pulse rate	Weight
Male	Current	No	…	69.0	107.8
Female	Never	No	…	67.5	61.6
Male	Current	No	…	64.5	83.7

[1] Bycroft, Clare, et al. "The UK Biobank resource with deep phenotyping and genomic data." *Nature* 562.7726 (2018): 203-209.

1.2 Current Practices and Challenges

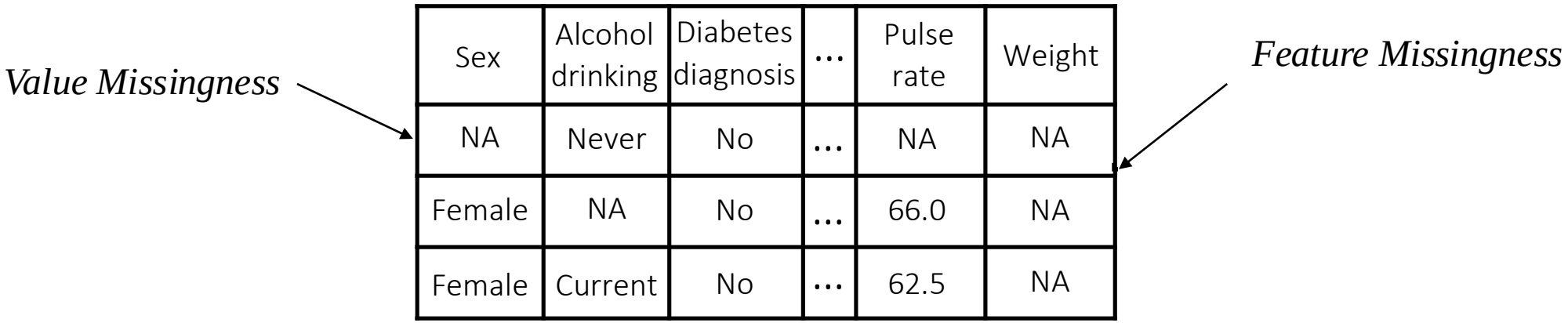
- A few image-tabular representation learning methods have been proposed.
  - Joint representation learning
  - Coordinated representation learning





1.2 Current Practices and Challenges

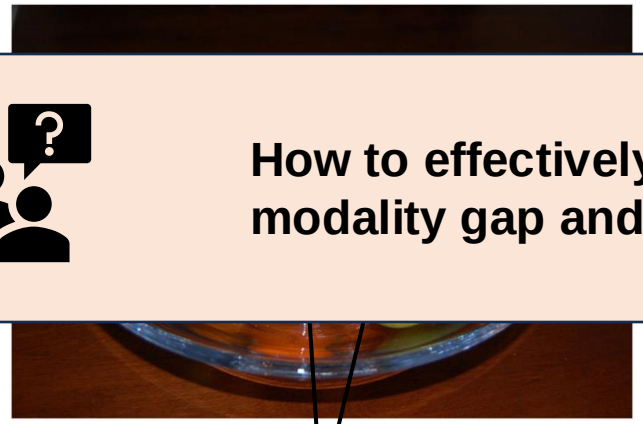
- **However**, current image-tabular approaches encounter two critical challenges
- **Challenge 1: low-quality data**
  - Limited data for a specific downstream task
  - Missing tabular data



1.2 Current Practices and Challenges

➢ **Challenge 2:** Modality disparities

Image and text data



Spatial correlation

A clear bowl containing four oranges, 2 bananas, and one apple.

Tabula data

1. Different data attributes

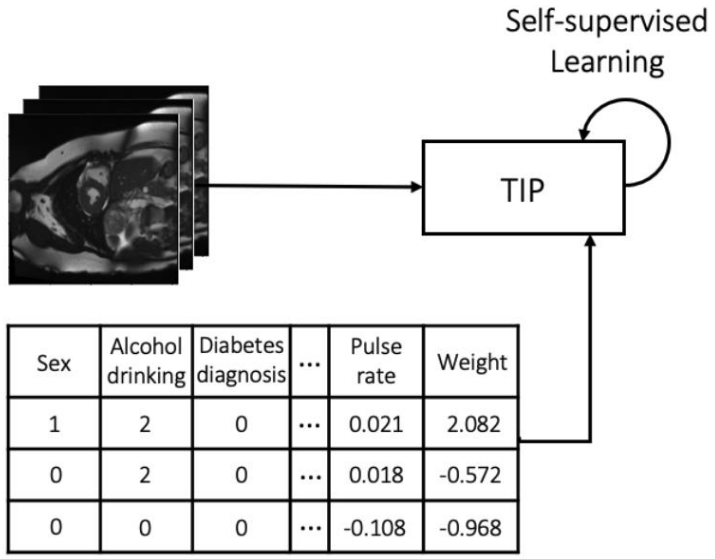
Sex	drinking	diagnosis	...	rate	Weight
Male	Current	No	...	69.0	107.8
Female	Never	No	...	67.5	61.6
Male	Current	No	...	64.5	83.7

How to effectively learn tabular and image representations to bridge the modality gap and handle missing data?

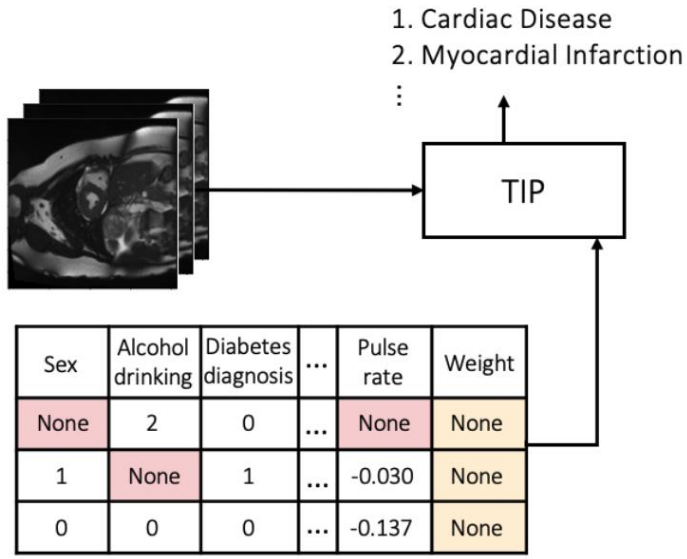
1.3 Our Method -- TIP

We propose **TIP**, a new image-tabular pretraining framework, which

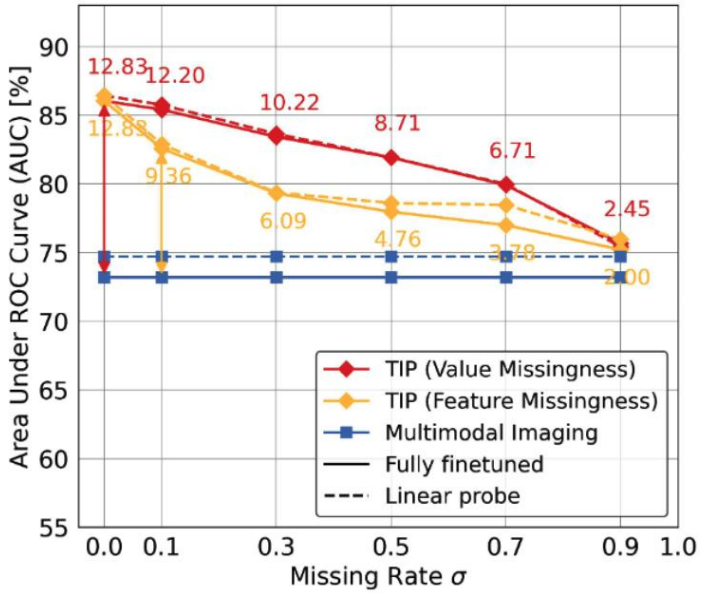
- utilizes a **novel transformer-based architecture** to supports incomplete, heterogeneous tabular data and improve intra- and inter-modality representation learning.
- exploits a **new self-supervised pretraining strategy** to tackle data missingness and extract multimodal information.



(a) Large-scale Pretraining with intact image-tabular pairs



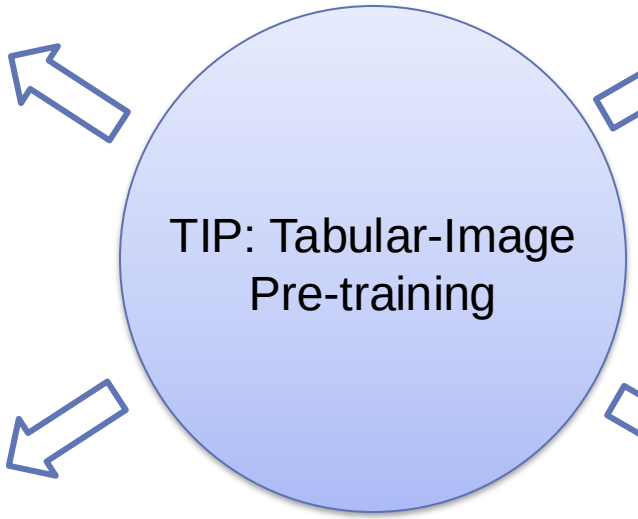
(b) Downstream task finetuning and inference with image and incomplete tabular data



(c) AUC results of cardiac disease diagnosis (CAD) in different missing rate

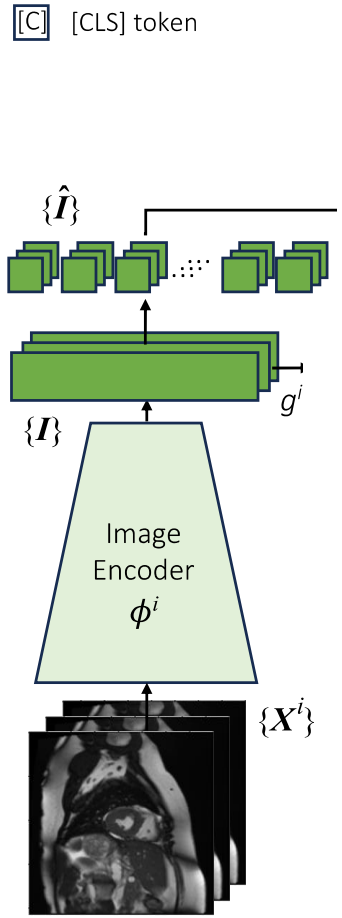
1.4 Contributions

- we are the **first** to propose SSL image-tabular pre-training (TIP) to tackle two challenges and investigate various data missingness
- Investigate **a novel self-supervised pre-training strategy** to capture rich multimodal representations robust to missing data



- Propose **a new multimodal network** for heterogeneous and incomplete data and for inter-modality representation learning
- TIP **significantly outperforms** SOTAs in complete and various incomplete data scenarios

2.1 TIP

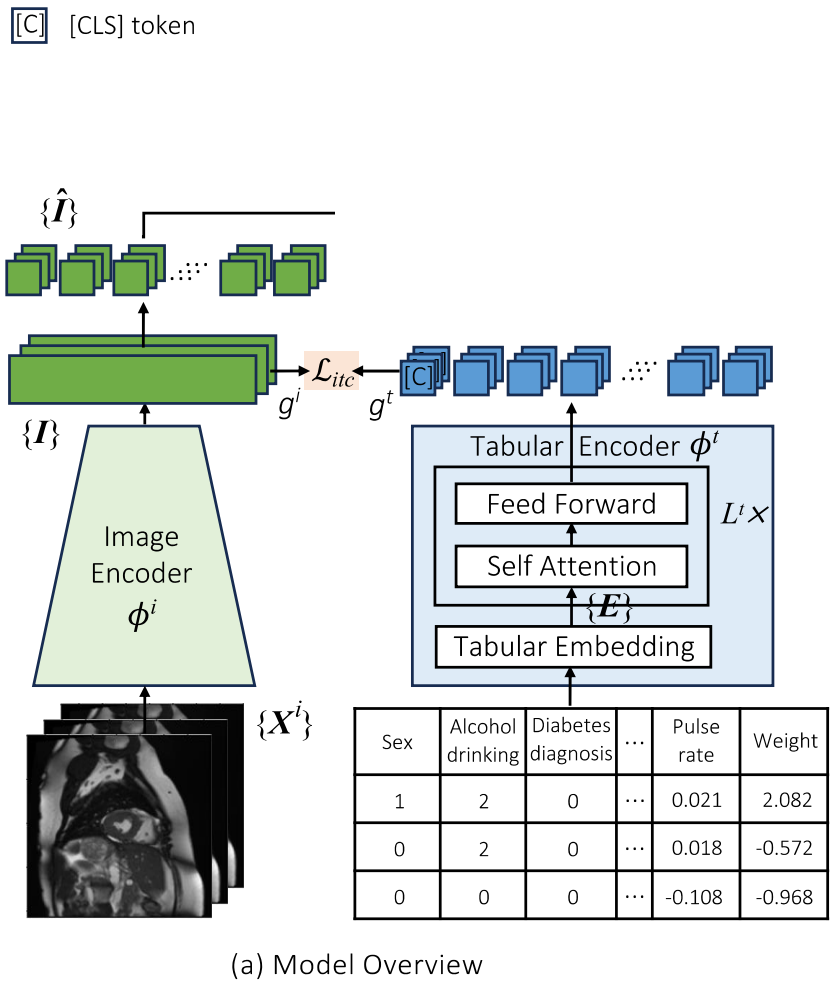


(a) Model Overview

Model Architecture

- Image Encoder: encodes image data

2.1 TIP

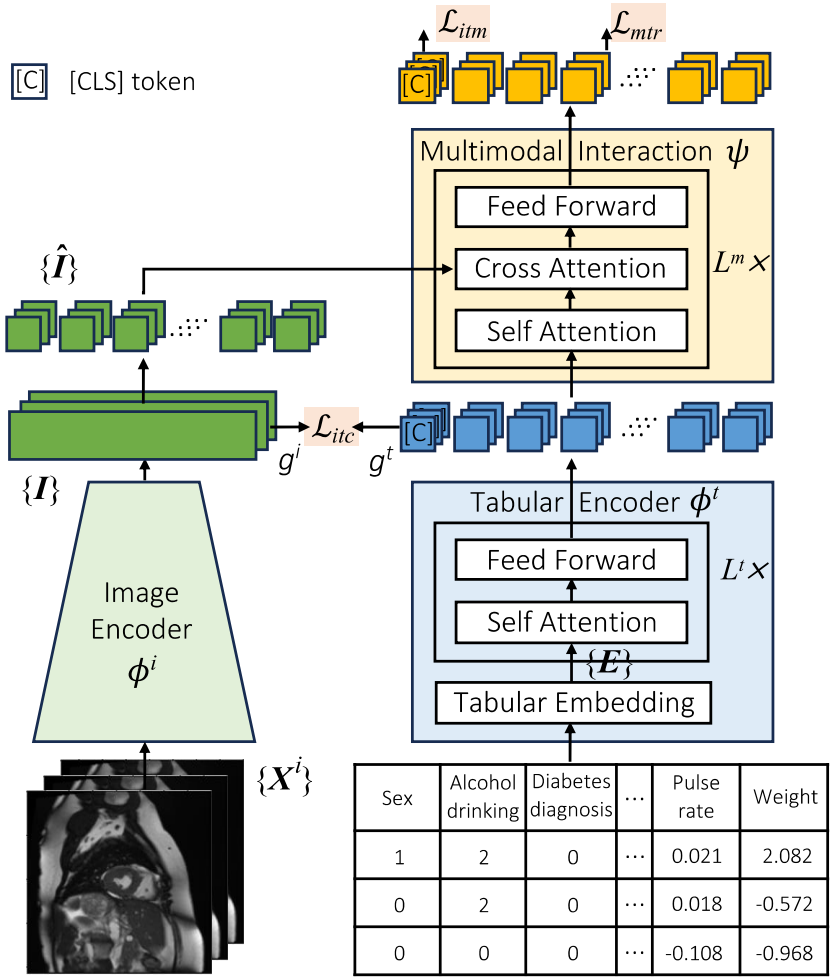


(a) Model Overview

Model Architecture

- Image Encoder: encodes image data
- Tabular Encoder: encodes heterogeneous and incomplete tabular data

2.1 TIP

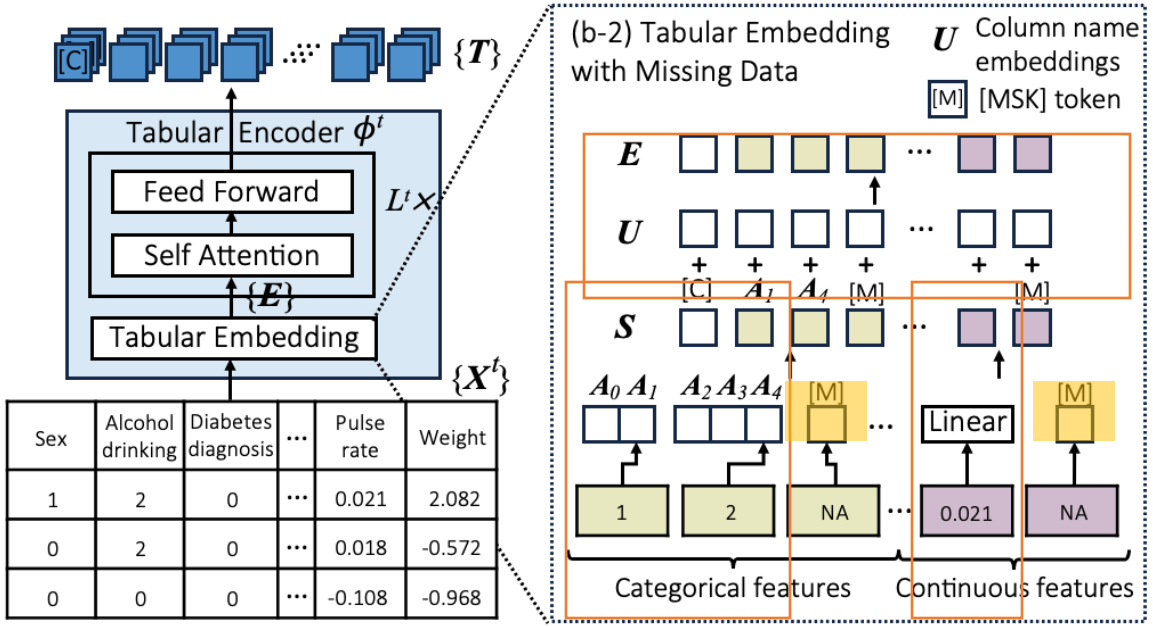


(a) Model Overview

Model Architecture

- Image Encoder: encodes image data
- Tabular Encoder: encodes heterogeneous and incomplete tabular data
- Multimodal Interaction: learns inter-modality relations and captures multimodal representations

2.2 Tabular Encoder



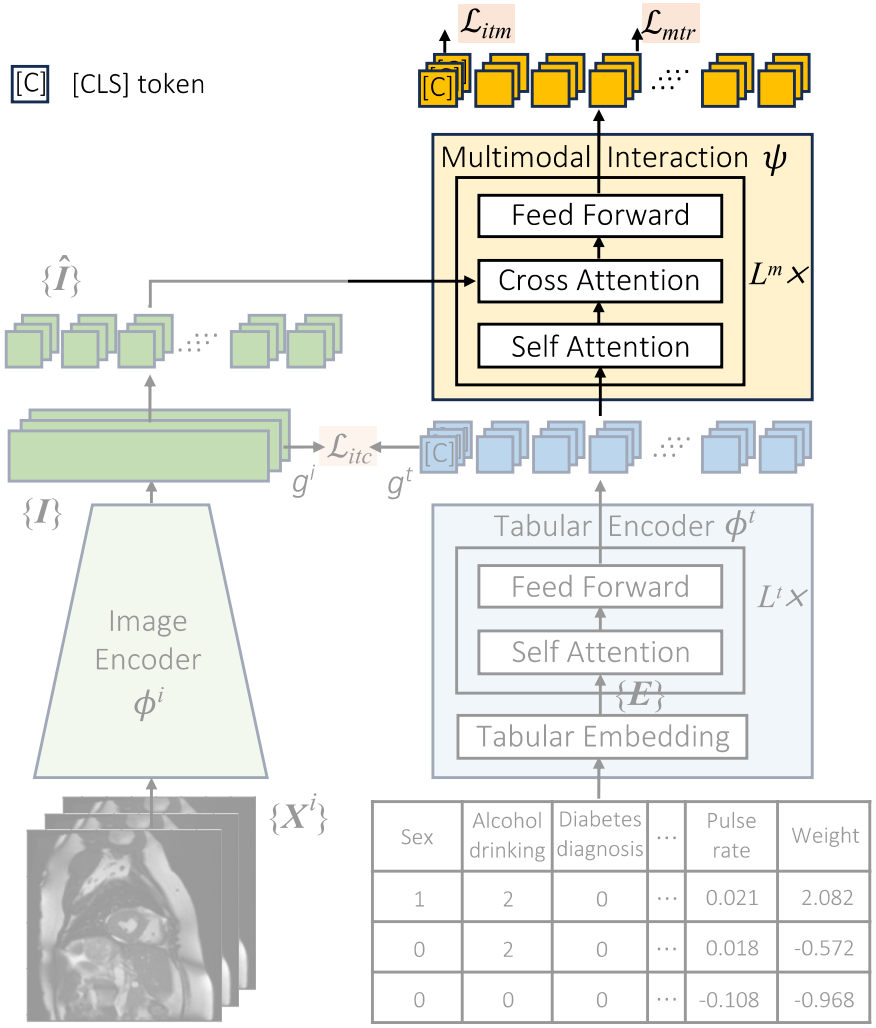
Tabular Embedding

- *Heterogeneous data processing*
  - Categorical data: learnable embeddings
  - Continuous data: a shared linear layer
- *Missing data processing*
  - A trainable [MSK] token for various types of missing data
- *Column diversity integration*
  - Learnable column name embeddings to dynamically capture inter-column relationships

High-dimensional Feature Extraction

- Transformer layers based on self-attention

2.2 Modality Interaction



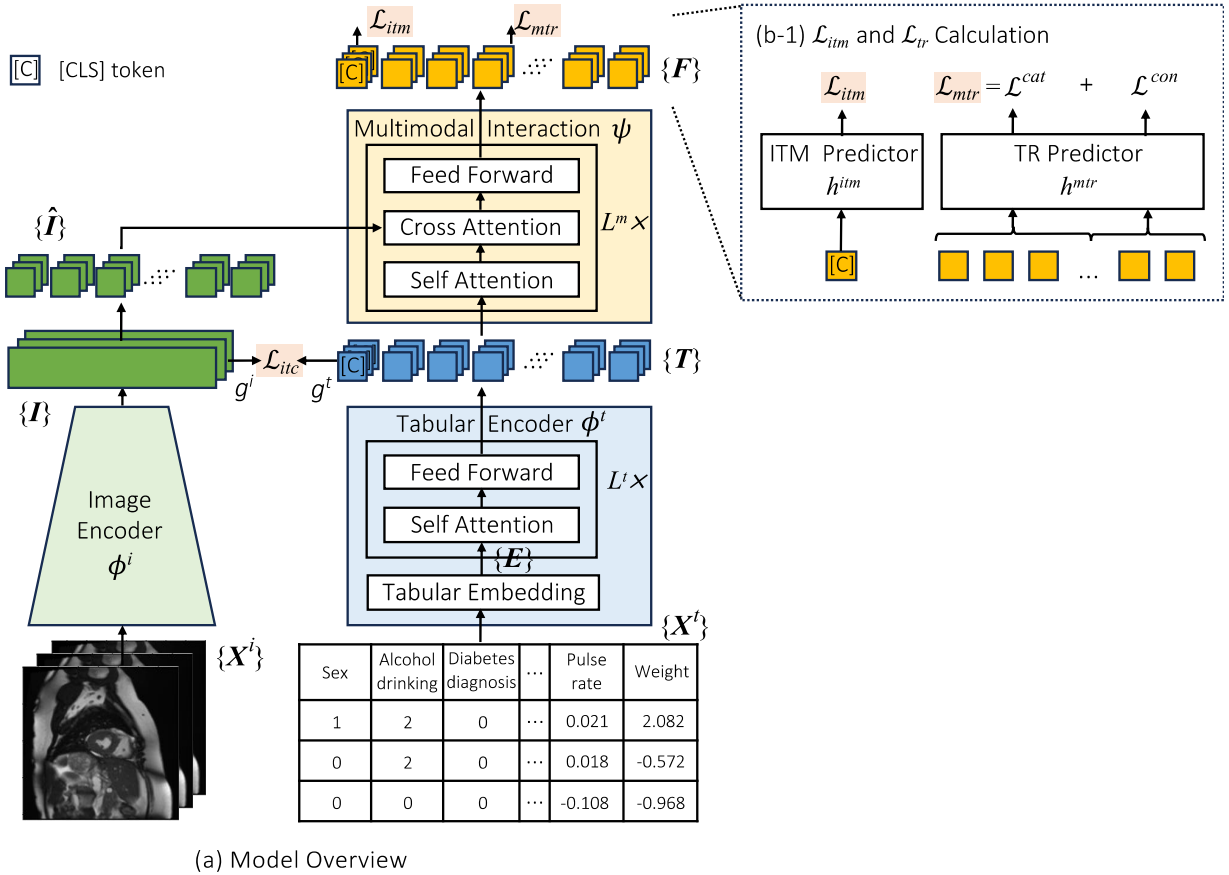
(a) Model Overview

- Transformer layers with cross-modal attention

$$CrossAttention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = softmax(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{d_k})\boldsymbol{V}$$

$$\boldsymbol{Q} = \boldsymbol{F}_{l-1}\boldsymbol{W}_l^Q,\ \boldsymbol{K} = \hat{\boldsymbol{I}}\boldsymbol{W}_l^K,\ \boldsymbol{V} = \hat{\boldsymbol{I}}\boldsymbol{W}_l^V,\ \text{and}\ \boldsymbol{F}_0 = \boldsymbol{T}$$

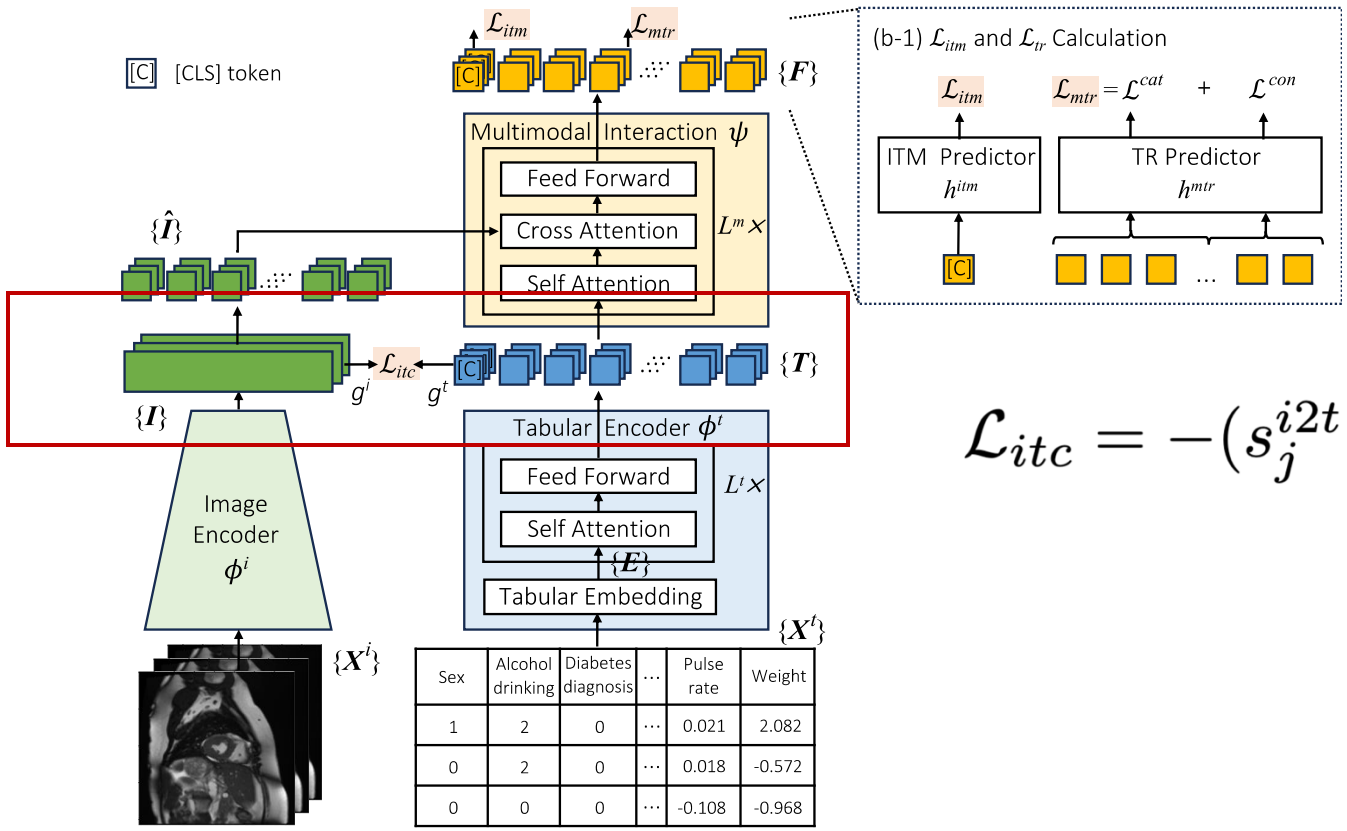
2.3 SSL Pre-training Strategy



(a) Model Overview

- Image-Tabular Contrastive Learning: captures better unimodal representations and aligns their feature spaces
- Image-Tabular Matching: captures inter-modality relations and generates a joint multimodal representation
- Masked Tabular Reconstruction: learns multimodal representations robust to missing tabular data

2.3 SSL Pre-training Strategy

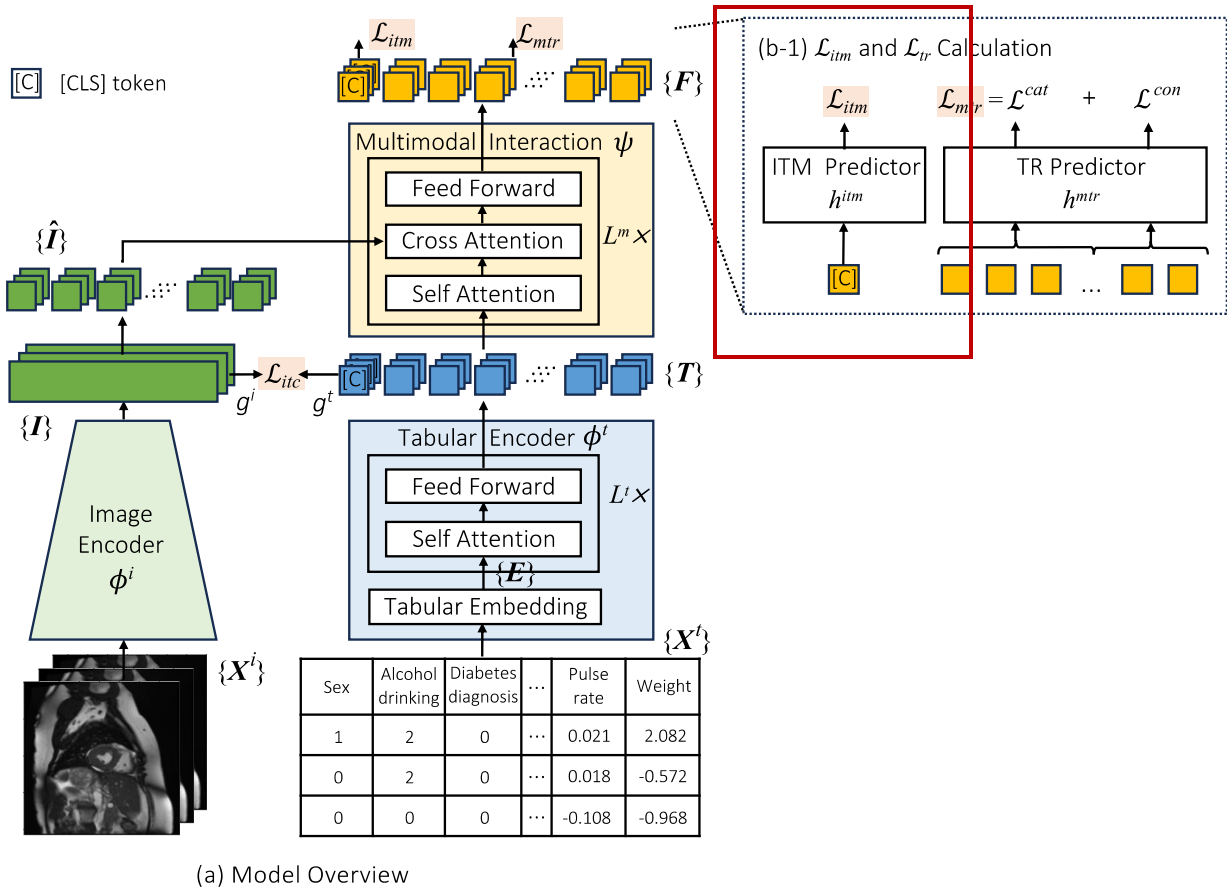


(a) Model Overview

$$\mathcal{L}_{itc} = -(s_j^{i2t} + s_j^{t2i})/2$$

- Image-Tabular Contrastive Learning: captures better unimodal representations and aligns their feature spaces
- Image-Tabular Matching: captures inter-modality relations and generates a joint multimodal representation
- Masked Tabular Reconstruction: learns multimodal representations robust to missing tabular data

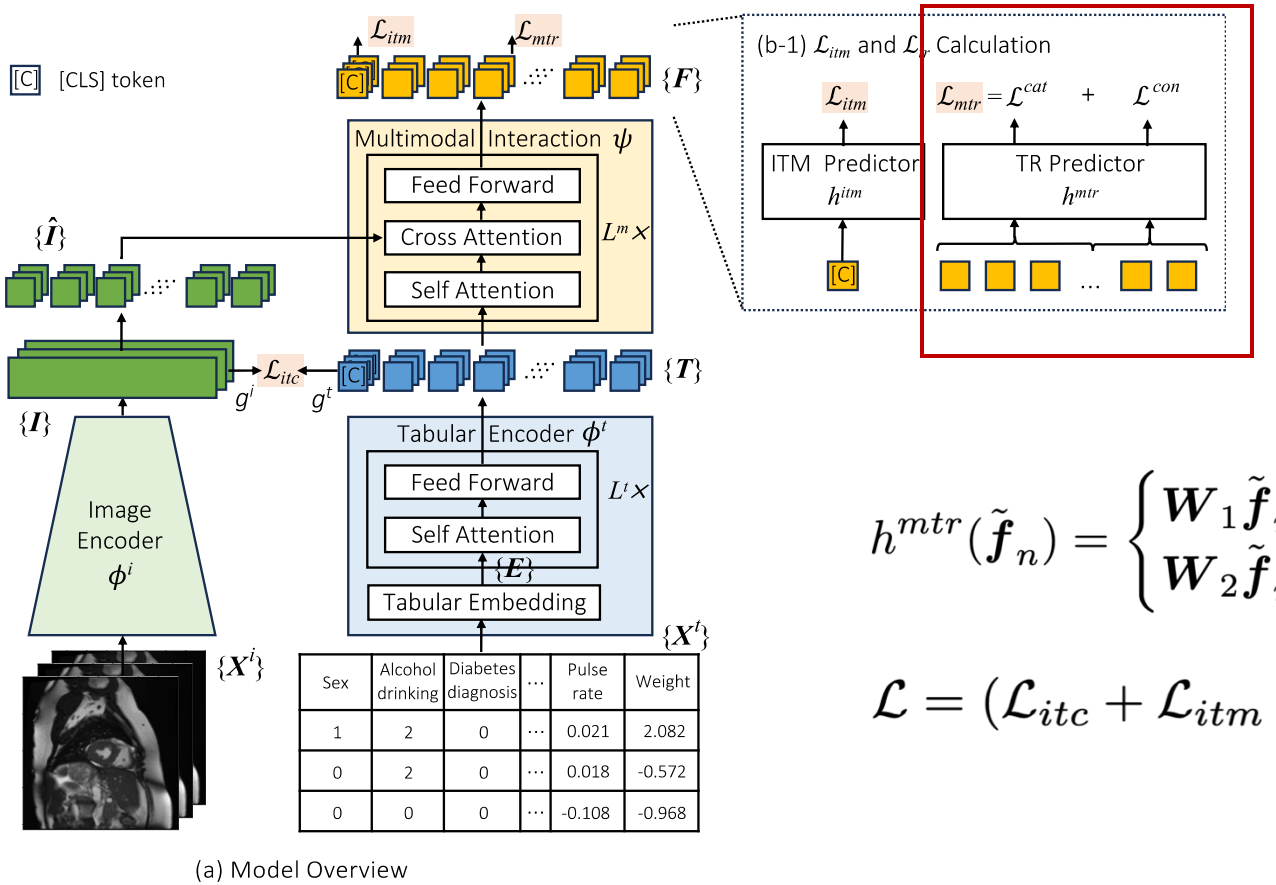
2.3 SSL Pre-training Strategy



(a) Model Overview

- Image-Tabular Contrastive Learning: captures better unimodal representations and aligns their feature spaces
- Image-Tabular Matching: captures inter-modality relations and generates a joint multimodal representation
- Masked Tabular Reconstruction: learns multimodal representations robust to missing tabular data

2.3 SSL Pre-training Strategy



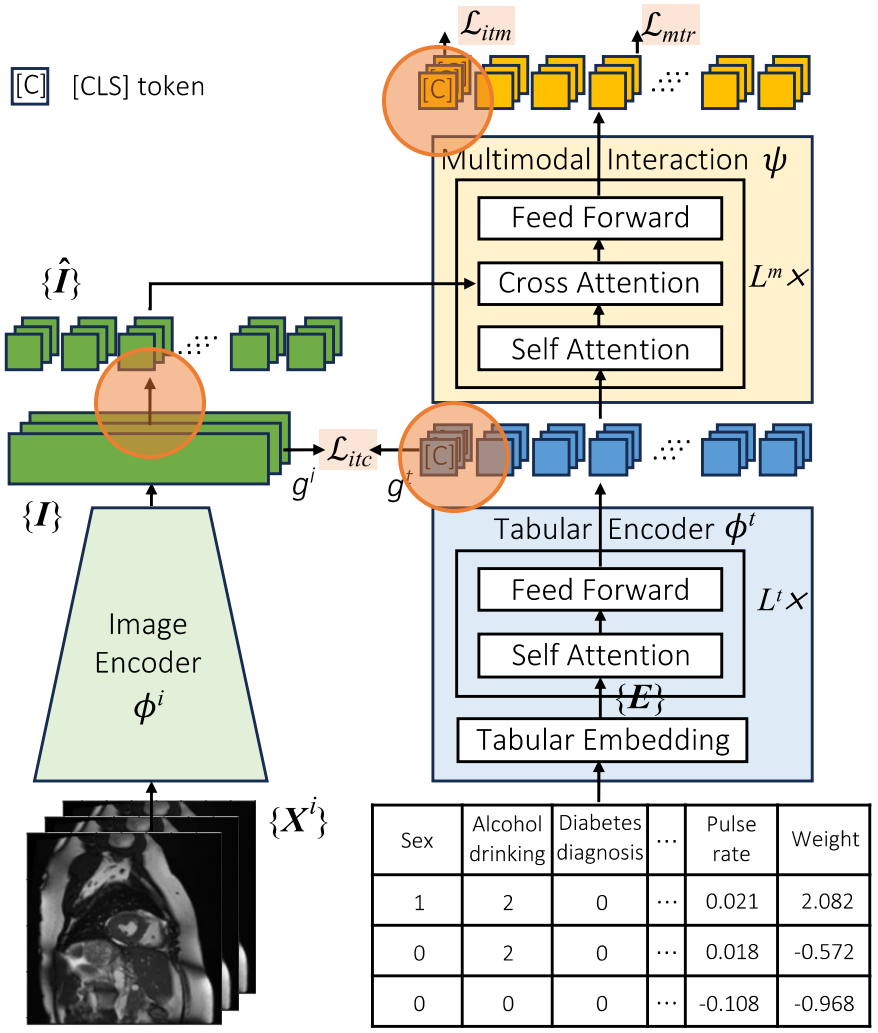
(a) Model Overview

$$h^{mtr}(\tilde{\boldsymbol{f}}_n) = \begin{cases} \boldsymbol{W}_1\tilde{\boldsymbol{f}}_n + \boldsymbol{b}_1, \boldsymbol{W}_1 \in \mathbb{R}^{\bar{N}_a \times D} & 0 < n \leq N_a, \\ \boldsymbol{W}_2\tilde{\boldsymbol{f}}_n + b_2, \boldsymbol{W}_2 \in \mathbb{R}^{1\times D} & N_a < n \leq N. \end{cases}$$

$$\mathcal{L} = (\mathcal{L}_{itc} + \mathcal{L}_{itm} + \mathcal{L}_{mtr})/3.$$

- Image-Tabular Contrastive Learning: captures better unimodal representations and aligns their feature spaces
- Image-Tabular Matching: captures inter-modality relations and generates a joint multimodal representation
- Masked Tabular Reconstruction: learns multimodal representations robust to missing tabular data

2.4 Ensemble Learning during Fine-tuning



(a) Model Overview

- Each feature extractor learns rich representations beneficial to downstream tasks
- Ensemble learning's ability to boost models' generalizability and results

3.1 Datasets and Metrics

- **UK Biobank (UKBB)**
  - Two binary classification tasks: myocardial infarction (Infarction) vs. health, coronary artery disease (CAD) vs. health
  - 2D MRI cardiac image and 76 tabular features (49 continuous, 26 categorical)
  - 36,167 samples (26,040 train, 6,510 val, 3,617 test)
  - Metric: AUC

- **Data Visual Marketing (DVM)**
  - Multi-class classification task: car model prediction
  - 2D RGB image and 17 tabular features (11 continuous, 4 categorical)
  - 176,414 samples (70,565 train, 17,642 val, 88,207 test)
  - Metric: Accuracy

3.2 Complete Downstream Task Data

Table 1: Results of DVM, CAD, and Infarction classification tasks comparing TIP with supervised/SSL image/multimodal techniques on complete data. ❄ denotes linear probing, *i.e.*, the feature extractors are frozen, and only the linear classifiers of the pre-trained models are tuned. 🔥 means fully fine-tuning, *i.e.*, all parameters are trainable. For supervised methods, all parameters are trainable in both ❄ and 🔥 columns.

Model	DVM Accuracy (%) ↑		CAD AUC (%) ↑		Infarction AUC (%) ↑	
	❄	🔥	❄	🔥	❄	🔥
(a) Supervised Image and Multimodal Methods						
ResNet-50 [32]	87.68	87.68	63.11	63.11	59.48	59.48
Concat Fuse (CF) [60]	94.60	94.60	85.76	85.76	**85.05**	85.04
Max Fuse (MF) [64]	94.39	94.39	85.31	85.31	84.75	84.75
Interact Fuse (IF) [23]	96.24	96.24	84.89	84.89	81.91	81.91
DAFT [67]	96.60	96.60	86.21	**86.21**	56.27	56.27
(b) SSL Image Pre-training Methods						
SimCLR [17]	61.06	87.65	68.42	72.58	68.86	75.07
BYOL [27]	56.26	88.64	65.67	69.18	66.63	70.12
SimSiam [18]	23.14	78.62	57.77	67.71	53.83	64.79
BarlowTwins [73]	53.60	88.36	55.64	61.68	50.01	60.14
(c) SSL Multimodal Pre-training Methods						
MMCL [28]	91.66	93.27	74.71	73.21	76.79	76.46
TIP	**99.72**	**99.56**	**86.43**	86.03	84.46	**85.58**

- Image only models 12.04% ↑
- Multimodal pre-training models 8.06% ↑
- Supervised multimodal models 3.12% ↑

3.2 Complete Downstream Task Data

- **Low-data regime**
  - 10% and 1% of the original dataset size

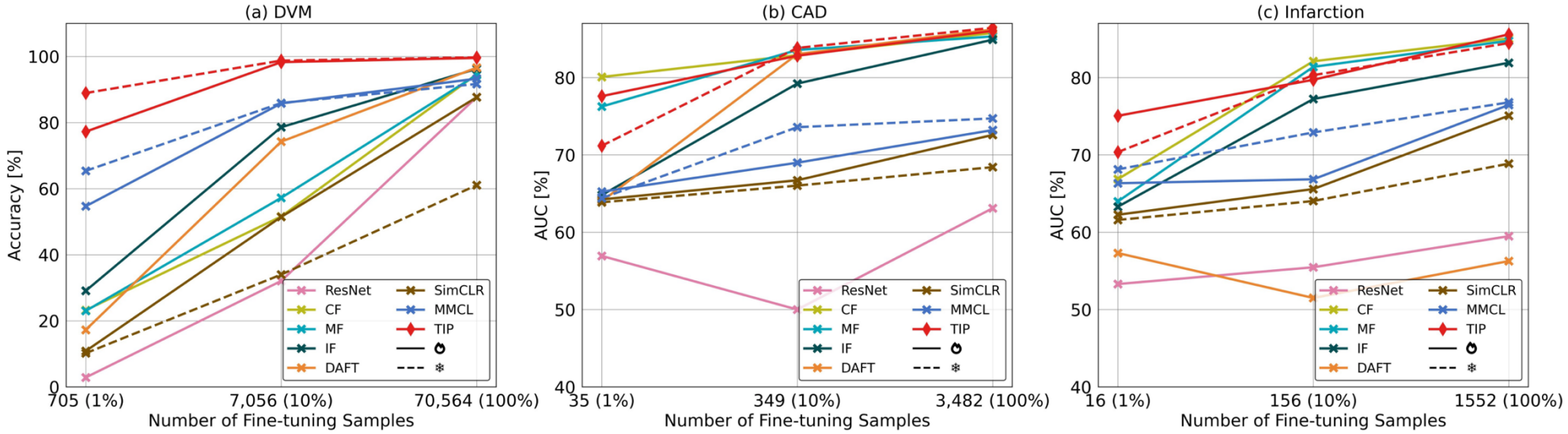
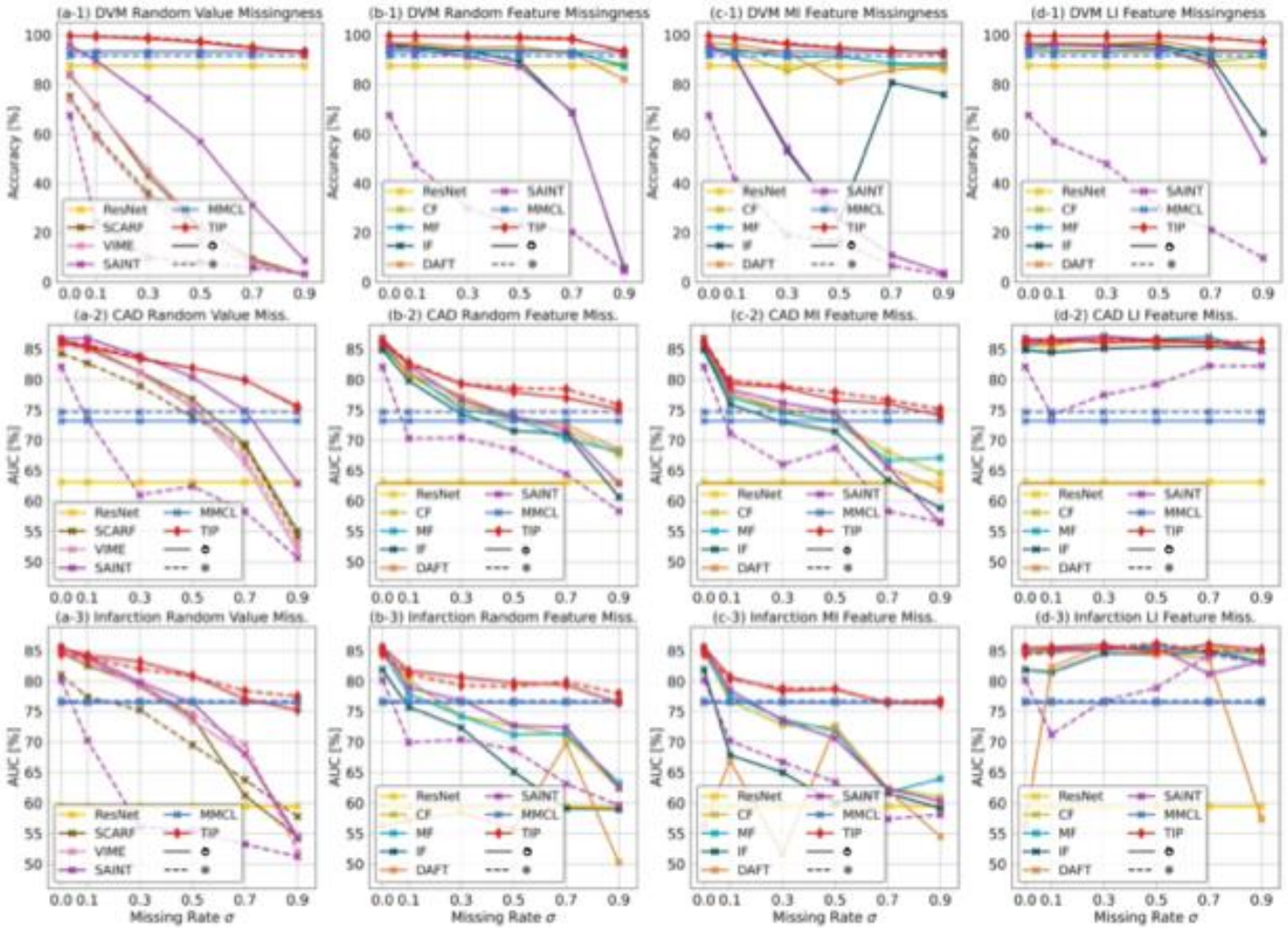


Fig. 3: Result comparison with supervised/SSL image/multimodal methods on various number of fine-tuning samples. 🔥 denotes fully fine-tuning, and ❄ means linear probing.

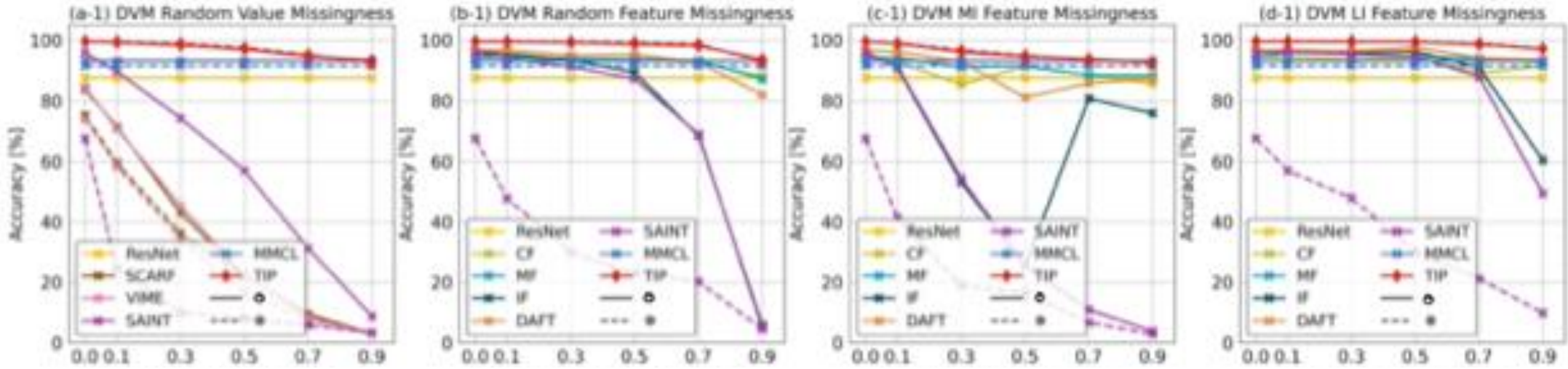
3.3 Incomplete Downstream Task Data

- Missing data scenarios
  - Random value
  - Random feature missingness
  - Most important feature
  - Least important feature

Sex	Alcohol drinking	Diabetes diagnosis	…	Pulse rate	Weight
NA	Never	No	…	NA	NA
Female	NA	No	…	66.0	NA
Female	Current	No	…	62.5	NA



3.3 Incomplete Downstream Task Data



- Most important feature Missingness (MIFM) is the most challenging scenario
- Incomplete data can still provide useful information, especially when MI features are not missing
- TIP achieves the best performance

3.3 Incomplete Downstream Task Data

- Effect of TIP's tabular encoder

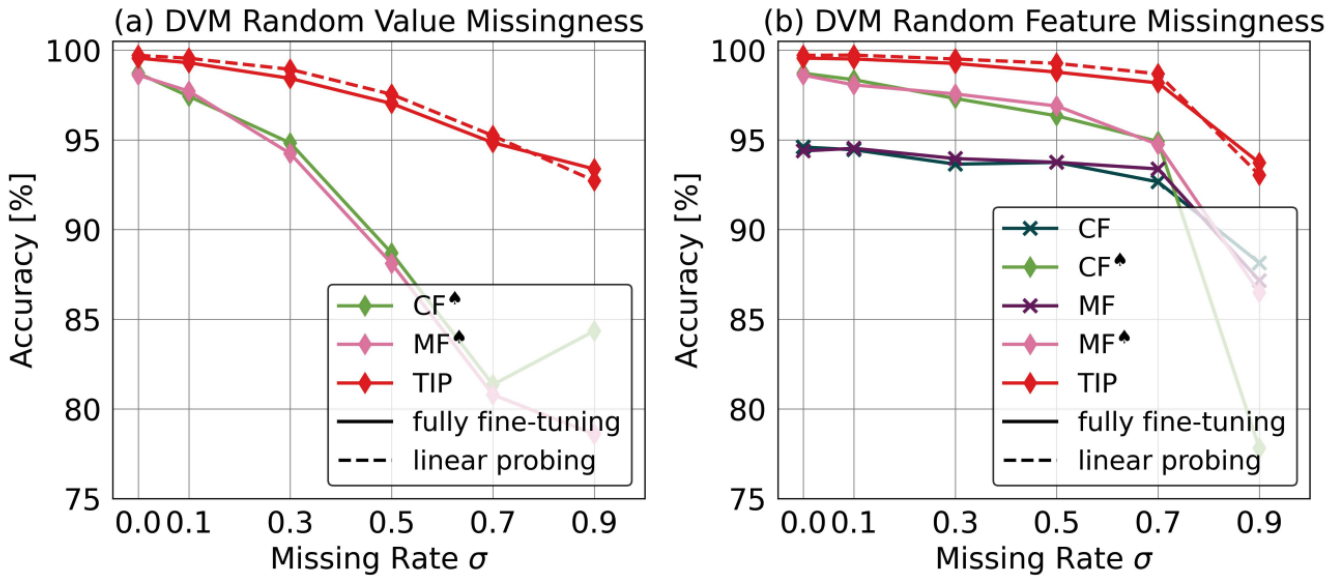


Fig. 9: Results comparing supervised multimodal methods and TIP on the DVM random value missingness (RVM) and random feature missingness (RFM) scenarios. ♠ means using TIP's tabular encoder.

3.3 Incomplete Downstream Task Data

➢ **Missing Value Reconstruction**

Table 2: Result comparison of TIP and data imputation methods for reconstructing missing continuous features across various missing rates on DVM and UKBB test sets.

Model	DVM RMSE ↓			UKBB RMSE ↓		
Missing rate σ	0.3	0.5	0.7	0.3	0.5	0.7
Mean [30]	0.9621	0.9783	0.9733	1.0162	1.0191	1.0070
MissForest [61]	0.6700	0.7653	0.8833	0.7516	0.7754	0.8177
GAIN [70]	1.0447	0.9428	2.9705	0.7920	2.0039	2.8130
MIWAE [49]	1.0105	1.0265	1.0218	1.0644	1.0680	1.0557
Hyperimpute [38]	0.6329	0.9428	0.9793	0.6803	0.7242	0.8060
TIP	**0.3899**	**0.4651**	**0.5055**	**0.6039**	**0.6460**	**0.7106**

- Outperform SOTA data imputation methods

3.4 Ablation Studies

- Applicable to different image encoder backbones
- Ablation study on key model components

Table 3: Experiments using different image encoder backbones and ablation study of TIP. ❄ means linear probing, and 🔥 represents fully fine-tuning.

Model	DVM Accuracy (%) ↑		CAD AUC (%) ↑		Infarction AUC (%) ↑	
	❄	🔥	❄	🔥	❄	🔥
(a) Applicability to Various Image Encoder Backbones						
TIP (ViT-S [22])	99.67	99.40	85.85	86.94	83.83	86.16
TIP (ViT-B [22])	99.40	99.28	84.90	86.93	83.15	85.76
(b) Ablation Study						
TIP w/o SSL pre-training	98.57	98.57	86.04	86.04	84.19	84.19
TIP w/o column name emb.	97.38	97.21	79.40	81.12	82.00	75.15
TIP w/o ensemble	99.63	99.35	86.00	**86.97**	84.43	84.00
TIP	**99.72**	**99.56**	**86.43**	86.03	**84.46**	**85.58**

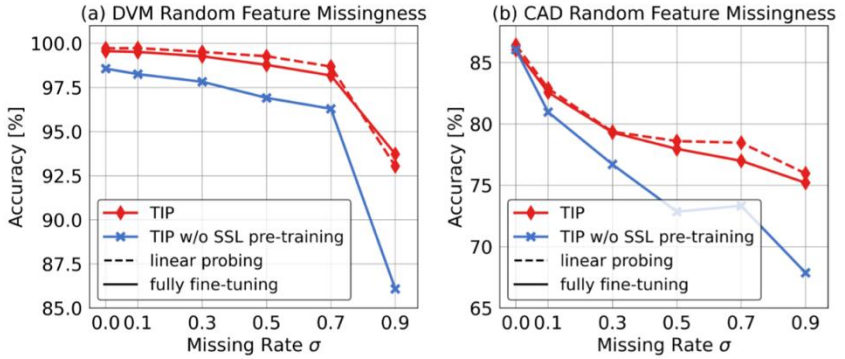


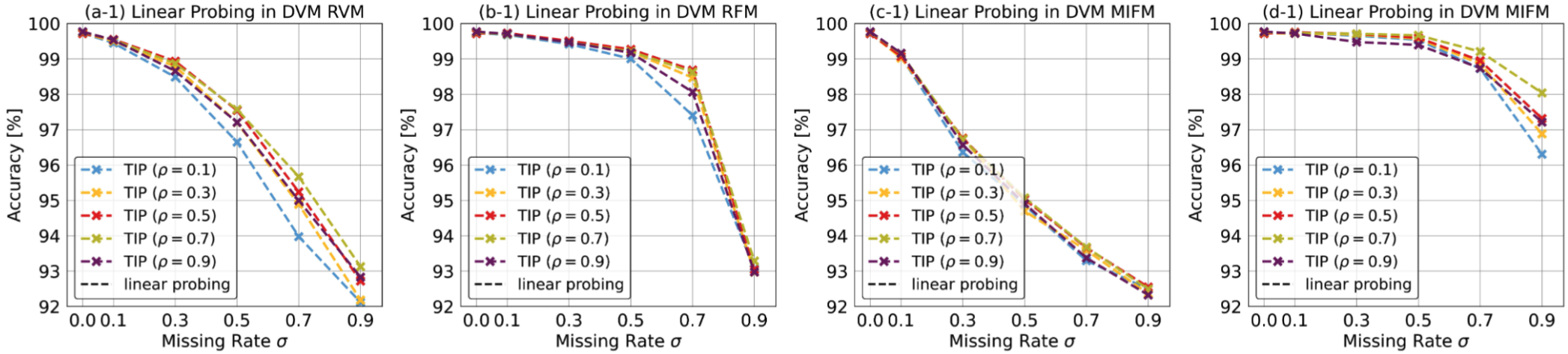
Fig. 5: Results comparing TIP with or without the proposed SSL pre-training strategy on the DVM and the CAD RFM scenario.

3.4 Ablation Studies

- Robust to different masking ratios

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



3.5 Visualization

- Visualization of TIP's tabular feature attention

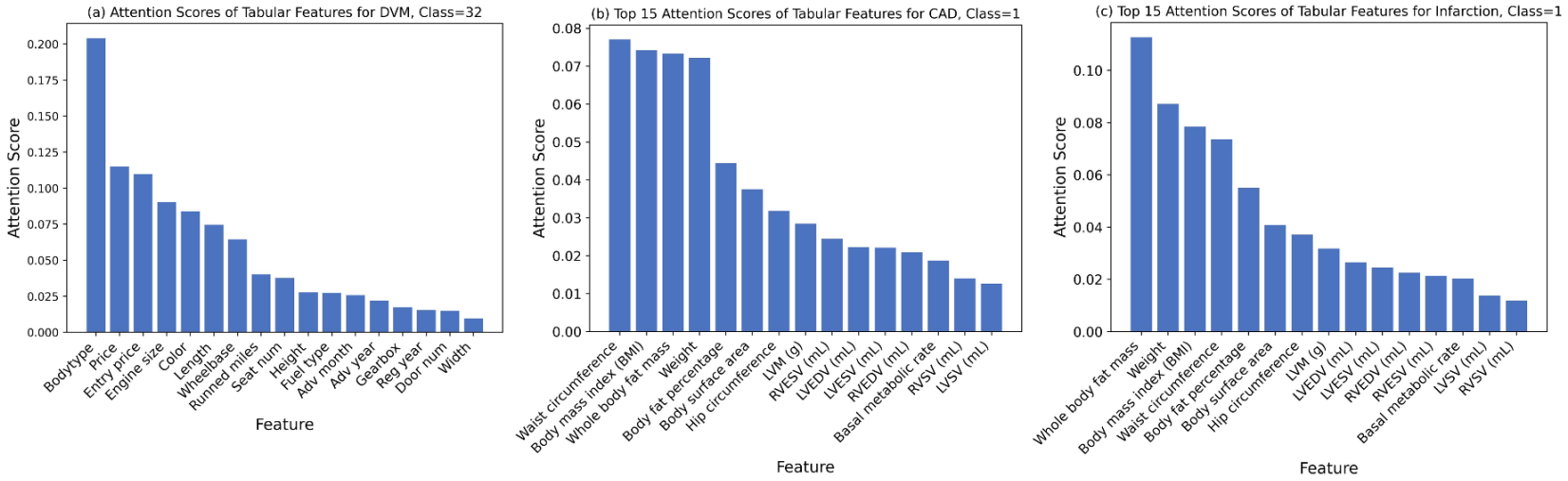
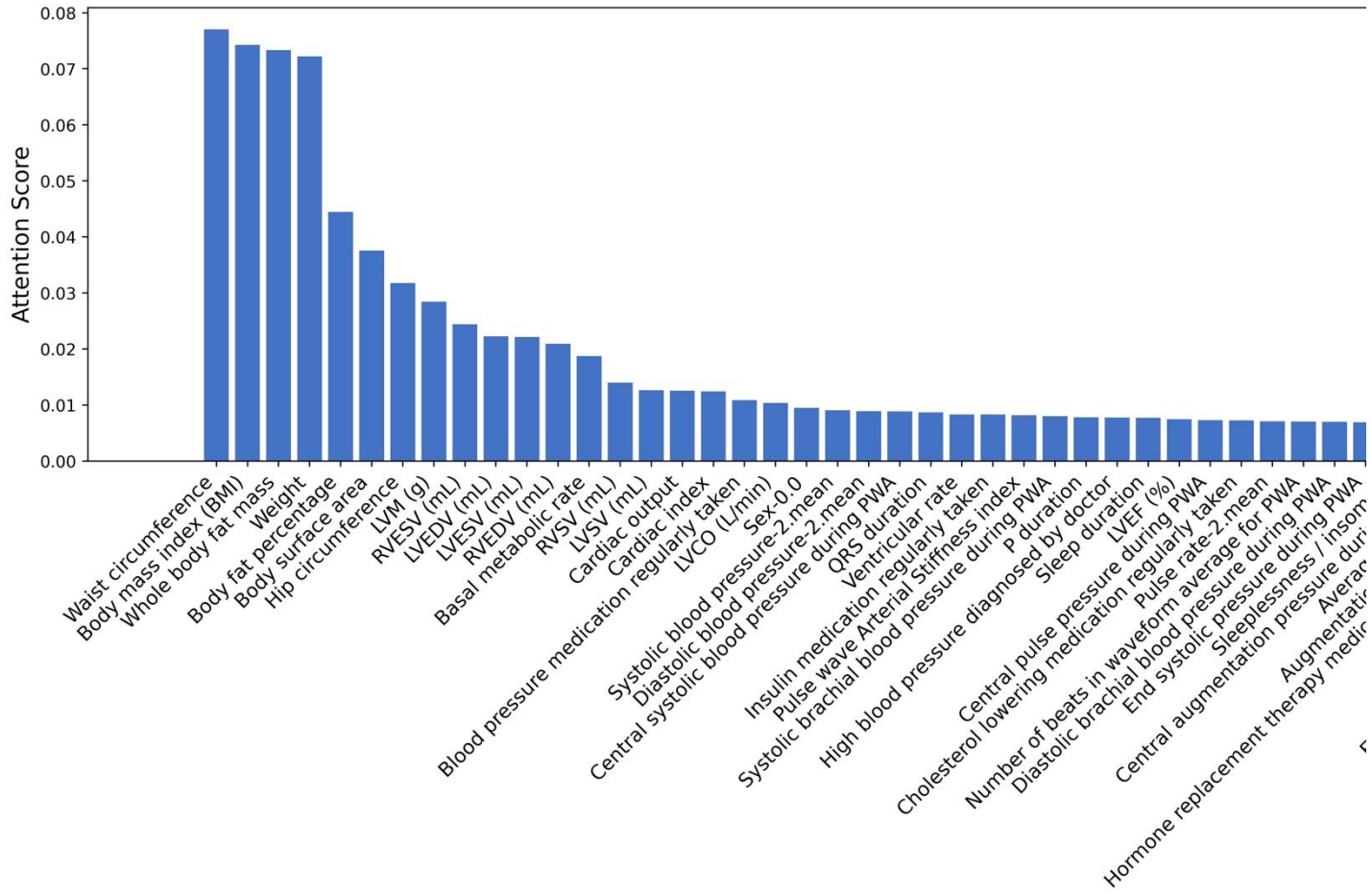


Fig. 6: The [CLS] token's attention scores to tabular features for a particular class from the last layer of TIP' tabular encoder.

- Attend to image-related features and the features that are not directly visible in the image

3.5 Visualization

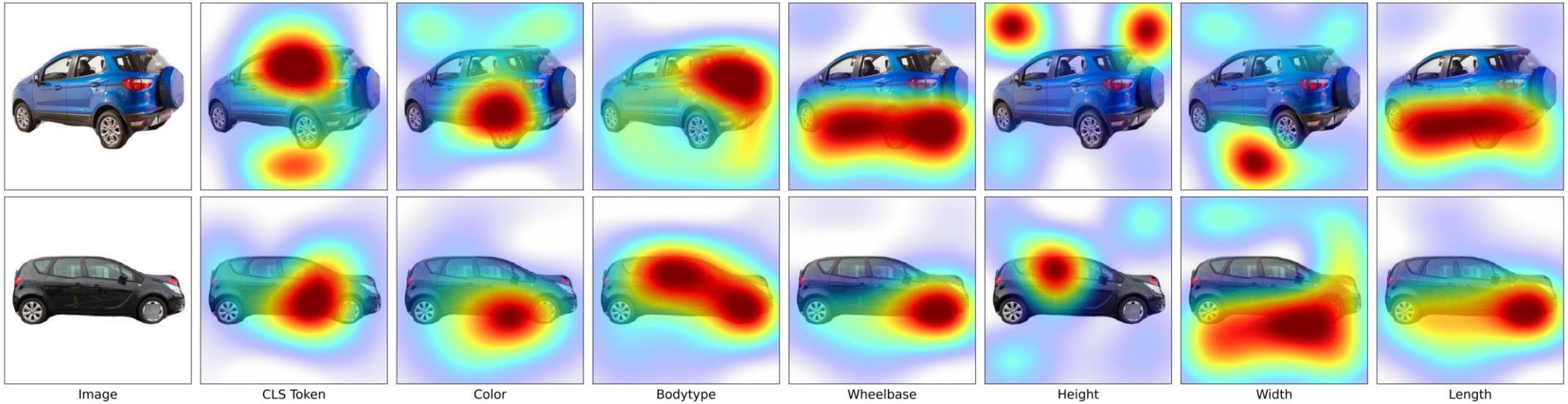
- Visualization of TIP's tabular feature attention



- Distinguish important imaging phenotypes
- Distinguish important non-imaging phenotypes

3.5 Visualization

- Visualization on the cross-attention map



- Identify the classification object
- Capture inter-modality relations

Conclusion

- We described the two challenges of image-tabular representation learning.
- We proposed TIP, a novel tabular-image pre-training framework for multimodal representation learning, featuring a transformer-based network and a self-supervised pre-training strategy.
- Experiments on natural and medical image datasets have demonstrated TIP's superior performance.
- We also observed that
  - Integrating tabular data can improve the image model performance
  - Incomplete tabular data still exists useful information

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

Paper is available at

Code is available at

Thank you!

Contact us: {s.du23, s.zheng22,y.wang23,w.bai,declan.oregan,c.qin15}@imperial.ac.uk

Backup

Data

Table 5: 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			

ID	Acro	Name
1	LVCO	
2	LVEDV	LV end-diastolic volume
3	LVEF	LV ejection fraction
4	LVESV	LV end-systolic volume
5	LVM	LV myocardial mass
6	LVSV	LV stroke volume
7	RVEDV	RV end-diastolic volume
8	RVEF	RV ejection fraction
9	RVESV	LV end-systolic volume
10	RVSV	RV stroke volume

Data

Table 6: 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			

Ablation Study

Table 8: 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

Ablation Study

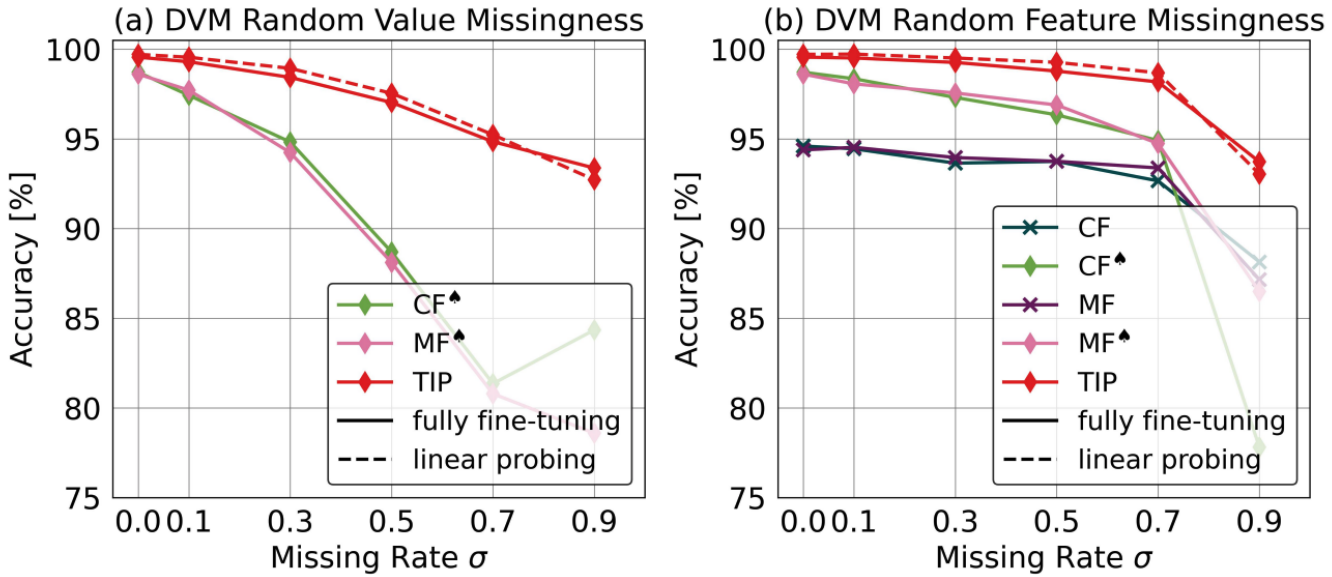


Fig. 9: Results comparing supervised multimodal methods and TIP on the DVM random value missingness (RVM) and random feature missingness (RFM) scenarios. ♠ means using TIP's tabular encoder.

Visualization

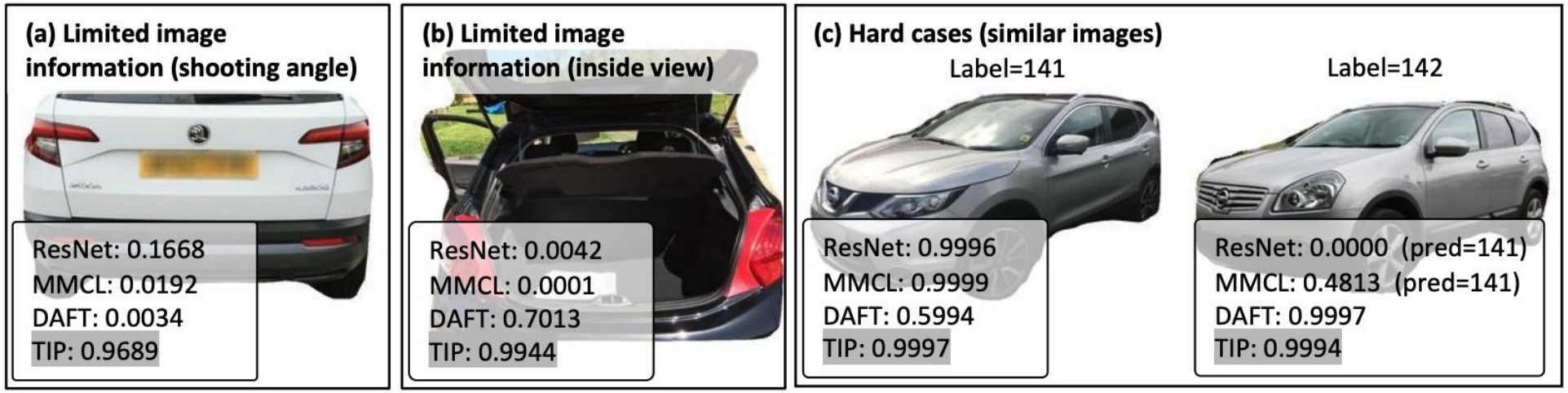
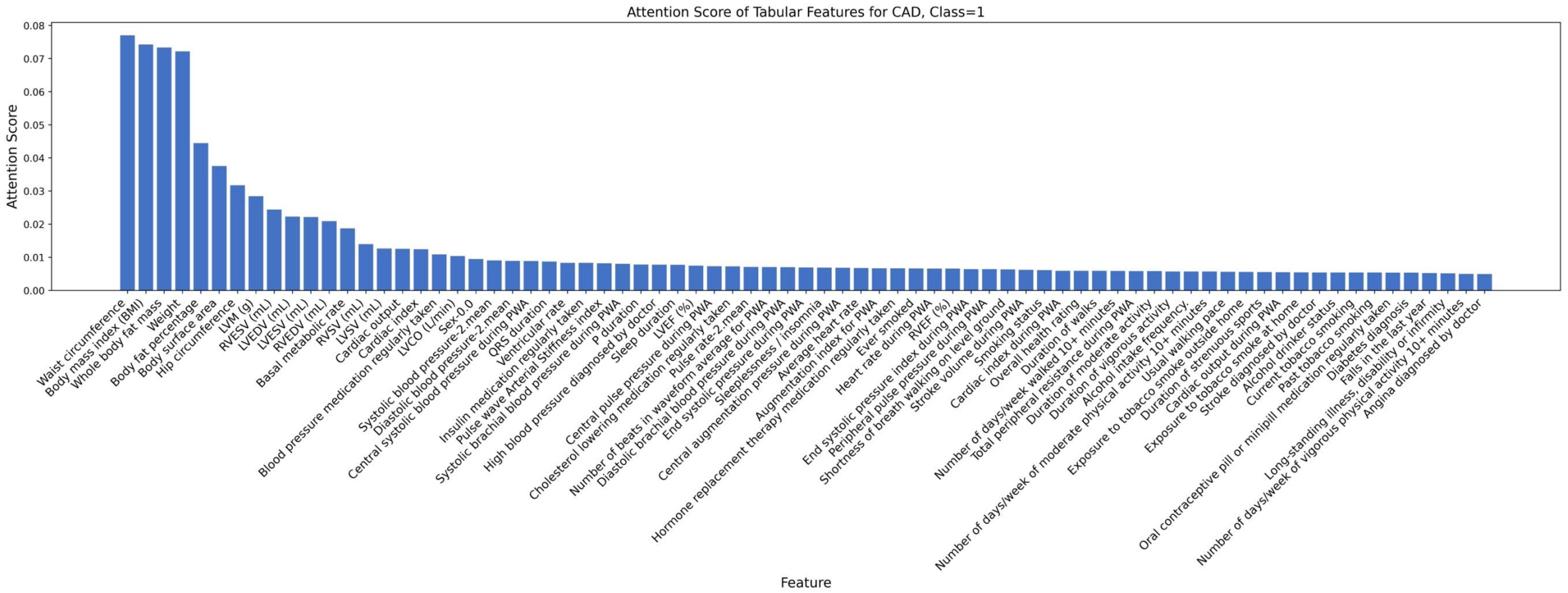


Fig. 13: DVM car visualization of samples and ground-truth class's predictions of TIP and other methods.

Visualization



- Distinguish important imaging phenotypes
- Distinguish important non-imaging phenotypes
- Focus more on physical measurements, which have high correlations with left ventricular function and structure