Unlocking the Power of Medical Tabular Data via Semantic-Aware Multimodal Pre-training

Yingsheng Liu[1,2], Haiming Li[2], Jingmin Zhu[1], Jiajun Sun[2], Victoria Mar[2], Monika Janda[3], H. Peter Soyer[3], Zongyuan Ge[2], and Zhen Yu[2](✉)

[1] Faculty of Information Technology, Monash University, Melbourne, Australia
[2] Monash University, Melbourne, Victoria, Australia
[3] The University of Queensland, Brisbane, Queensland, Australia
Zhen.Yu1@monash.edu

Abstract. While vision-language models dominate medical representation learning, unstructured text lacks the dense, quantitative diagnostic phenotypes inherent in structured clinical tables. However, existing multimodal pre-training methods underutilize this potential due to semantic-agnostic designs that treat tabular inputs as flat vectors and employ unstable continuous regression objectives. To overcome this, we propose a novel semantic-aware framework explicitly modeling the intrinsic two-dimensional structure of tabular data. First, addressing the inter-feature hierarchy of varying diagnostic importance, we introduce Importance-Aware Adaptive Masking to construct a label-free curriculum prioritizing salient features. Second, addressing the intra-feature continuity-discreteness duality, we propose a Soft-Label Discretized Module that replaces unstable numerical regression with stable distribution matching, thereby mathematically preserving ordinal relationships. Extensive experiments across large-scale dermatology (SLICE-3D, HOP) and ophthalmology (EyePACS) datasets establish a new state-of-the-art (SOTA), demonstrating exceptional robustness and cross-domain generalizability.
Code: https://github.com/Ethan-ysliu/AID

Keywords: Multimodal · Self-supervised Learning · Image-tabular.

1 Introduction

Clinical decision-making is a multimodal process synthesizing visual evidence and structured patient data [13]. For instance, a radiologist interpreting a chest X-ray benefits from patient smoking history and lab results, while a dermatologist evaluating a skin lesion considers demographic and morphological features, such as age and border irregularity. Although vision-language models have revolutionized representation learning[24, 17, 23, 25], unstructured text often lacks the dense, quantitative diagnostic phenotypes inherent in structured tabular records[12]. Large-scale medical databases present a significant opportunity to develop holistic multimodal AI systems utilizing this specific pairing. However, profound label scarcity limits this potential, as obtaining high-quality expert annotations remains an expensive bottleneck[7]. This reality creates a compelling

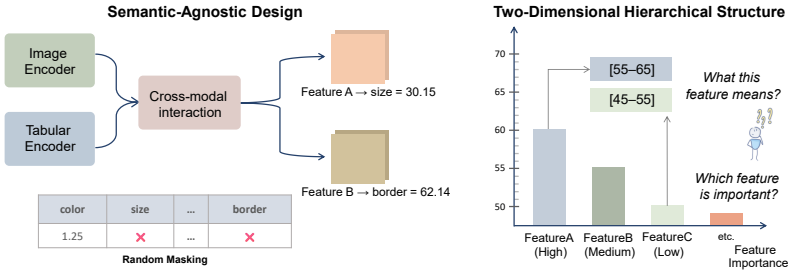

Fig. 1. Limitations of Current Semantic-Agnostic Designs and Motivation for the Two-Dimensional Hierarchical Structure.

need for self-supervised learning methods capable of harnessing unlabeled image-tabular data to acquire robust representations for downstream clinical tasks.

Despite this urgent need, existing multimodal pre-training methods are limited by a semantic-agnostic design[7, 11]. As in Fig. 1, SOTA approaches typically treat tabular inputs as flat, unstructured vectors and employ semantically naive pretext tasks. First, uniform masking strategies process all clinical features equally, ignoring inherent disparities in diagnostic importance. Second, and more critically, reconstructing masked values via direct continuous numerical regression (typically mean squared error) creates a severe optimization bottleneck. Specifically, forcing a model to regress exact continuous values from nearly identical visual patterns such as distinguishing visually indistinguishable asymmetry scores of 0.61 and 0.65 that conceptually map to the shared semantic category "moderate asymmetry" creates an ill-posed objective. Such a noisy learning formulation encourages overfitting and prevents the network from capturing robust, high-level semantic links between visual evidence and clinical concepts.

To unlock the full potential of multimodal medical representation learning, we must move beyond this semantic-agnostic paradigm and explicitly model the intrinsic properties of structured clinical data. We posit that medical tabular data possesses a unique two-dimensional hierarchical structure. At the inter-feature level, clinical attributes exhibit varying degrees of diagnostic importance; for example, lesion border asymmetry carries far more clinical weight than general demographic information. At the intra-feature level, values within a single clinical feature demonstrate a continuity-discreteness duality. While maintaining exact continuous precision is necessary for calculating global cross-modal alignment, local feature reconstruction reflects discrete medical semantic intervals. In clinical practice, precise numerical differences often map to shared diagnostic concepts, such as identifying a measurement simply as abnormally large.

To bridge this semantic gap, we propose Adaptive Importance-guided Discretized reconstruction (AID), a semantic-aware framework modeling these two structural dimensions. To address the inter-feature hierarchy, we introduce a label-free importance-aware adaptive masking strategy. Leveraging a frozen meta-learning prior to estimate data-driven significance without accessing downstream

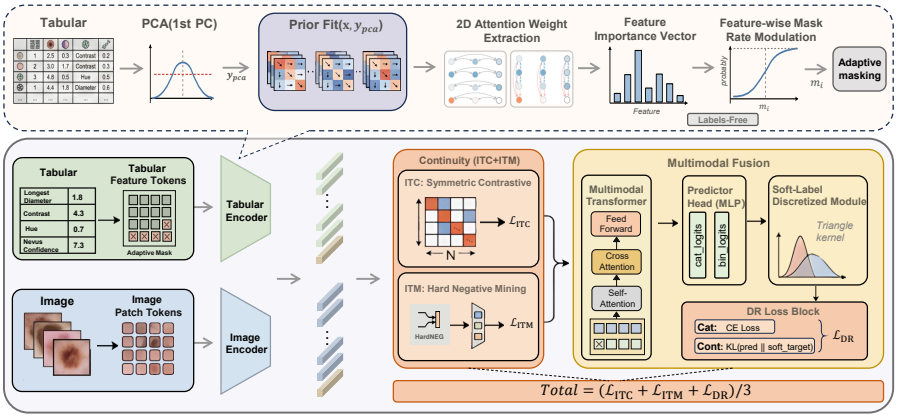

Fig. 2. Overall Architecture of the Proposed AID Framework.

targets, this module constructs a label leakage-free curriculum prioritizing diagnostically salient features. To address the intra-feature duality, an end-to-end soft-label discretized module replaces unstable numerical regression with a stable distribution matching objective. Crucially, by utilizing a triangle kernel to allocate probability mass across adjacent semantic intervals, the soft-labeling mechanism translates exact values into probabilities. This operation mathematically preserves the inherent ordinal relationships of clinical measurements, overcoming the limitations of both direct regression and naive hard-label classification.

While early methods established strong baselines for processing tabular data, integrating them into multimodal frameworks often relies on rudimentary fusion techniques. Recent multimodal self-supervised pre-training approaches achieve strong performance by combining contrastive learning with masked feature reconstruction[7, 5, 16]. However, these methods uniformly adopt semantic-agnostic view, treating all clinical features equally and relying on continuous value regression. Unlike TP-BERTa[22], which tokenizes numerical values for unimodal LMs, AID enforces ordinality at the reconstruction objective in image-tabular multimodal pre-training. To the best of our knowledge, no existing pre-training objective explicitly models diagnostic feature importance and continuity-discreteness duality of medical tabular data. Our work fills this critical void by establishing a new semantic-centric paradigm for image-tabular representation learning that translates raw clinical measurements into robust, generalizable representations.

2 Methodology

2.1 Overall Architecture and the Duality Principle

As illustrated in Fig. 2, the AID framework comprises a ViT-base [6] for image encoding, a hybrid Transformer for structured clinical features, and a cross-attention multimodal encoder. The AID framework fundamentally models the

continuity-discreteness duality of medical tabular data by processing tabular inputs conditionally based on the optimization objective. For cross-modal alignment and matching, the network utilizes exact continuous numerical values to preserve precise clinical measurement scales. Conversely, for discretized reconstruction, an end-to-end discretized module transforms these continuous values into soft probability distributions over discrete semantic intervals. This duality ensures precise global alignment while compelling the model to learn robust, interval-based clinical semantics during local reconstruction.

2.2 Feature Importance Extraction & Adaptive Masking

To construct a label-free importance-aware masking curriculum, AID extracts data-driven feature significance strictly offline prior to pre-training. Unlike tree-based models requiring downstream ground-truth labels, AID derives importance using an unsupervised approach coupled with a frozen meta-learning prior. Specifically, the framework applies principal component analysis to the standardized tabular feature matrix, utilizing the median of the first principal component to generate balanced binary pseudo-labels. These pseudo-labels adapt a frozen tabular foundation model[12]. Since TabPFN v2 is pre-trained on large-scale generative tabular distributions, it provides a universal meta-prior that naturally captures feature dependencies. Fitting this model with pseudo-labels enables its internal self-attention mechanisms to act as a robust inductive bias for dataset-specific feature interactions, mitigating the lack of theoretical guarantees for attention-based feature importance, all without accessing actual diagnostic targets. By registering forward hooks across the attention layers, the system extracts the query, key, and self-attention weight matrices. Aggregating these matrices across all layers, followed by L1 normalization and min-max scaling, produces a normalized global importance vector s, representing the relative significance of each clinical feature.

The inter-feature hierarchy of clinical data dictates that specific attributes possess substantially higher diagnostic significance. To embed this inductive bias, the pre-training pipeline utilizes the extracted importance vector s to govern feature corruption. For each continuous feature j, AID computes an adaptive masking rate r_j:

$$r_j = \min(r_{\text{base}} + \alpha s_j, r_{\max}) \tag{1}$$

where r_{base} denotes the base masking probability, α controls the sensitivity to the importance score, and $r_{\max}$ serves as a strict upper bound. This mechanism dynamically establishes a challenging learning curriculum. Highly relevant clinical features receive proportionally higher masking rates, compelling the multimodal encoder to heavily utilize cross-modal visual evidence for reconstructing the masked tabular tokens. To maintain data distribution stability, the masking operation exclusively replaces selected input tokens with a learnable special token embedding, thereby avoiding the representation noise typically introduced by random marginal replacement.

2.3 Soft-Label Discretized Module

The intra-feature duality of tabular data necessitates transforming continuous values into stable semantic intervals during reconstruction. Standard hard-label discretization assigns each continuous value to a single bin[2], treating semantic intervals independently and destroying the inherent ordinal relationships of clinical measurements. To overcome this limitation, AID introduces an end-to-end soft-label discretized module. Initially, this module establishes B discrete bins per feature using quantile boundaries strictly derived from the training distribution. During the forward pass, for a given feature value v falling into bin c with boundaries $[b_{c-1}, b_c)$, the module computes the intra-bin relative position $p = (v - b_{c-1})/(b_c - b_{c-1})$. A triangle kernel subsequently formulates a soft-label distribution q_b by allocating probability mass proportionally between the current bin c and its immediate adjacent neighbor ($c-1$ or $c+1$) based on the distance p, ensuring $\sum_{b=1}^{B} q_b = 1$. This soft-labeling process dynamically maps precise numerical measurements into smooth probability distributions across adjacent semantic intervals. Distributing mass based on boundary proximity mathematically preserves the ordinality of continuous variables, providing a stable, relation-aware target for cross-modal reconstruction.

2.4 Multimodal Pre-training

The comprehensive pre-training objective $\mathcal{L}_{\text{AID}} = (\mathcal{L}_{\text{ITC}} + \mathcal{L}_{\text{ITM}} + \mathcal{L}_{\text{DR}})/3$ minimizes a composite loss comprising Image-Tabular Contrastive (ITC), Image-Tabular Matching (ITM), and Discretized Reconstruction (DR) losses. The ITC and ITM objectives align global representations and perform fine-grained binary matching between genuine and mismatched image-tabular pairs using exact continuous values[11, 7, 9]. Specifically, the ITM objective employs an in-batch hard negative mining strategy, selecting mismatched pairs based on the multimodal similarity scores derived from the ITC computation. Crucially, the adaptive masking process does not degrade the quality of negative mining; the hard negatives are dynamically sampled from the unmasked global representations within the batch, ensuring semantically valid and structurally intact mismatched pairs.

Conversely, the core DR objective reconstructs the masked tabular features utilizing the dual-format output of the predictor head. The DR loss unifies two components: a standard Cross-Entropy loss for the D_{cat} categorical features, and a Kullback-Leibler (KL) divergence loss for the D_{con} continuous features. For a batch of N samples, DR evaluates the divergence between the predicted probabilities (e.g., pi, k for continuous bins, $\hat{y}i, j$ for categorical classes) and the corresponding targets (qi, k from the discretized module, $y_{i,j}$ for true classes) across all masked positions m:

$$\mathcal{L}_{\text{DR}} = \frac{1}{2}\left(\frac{\sum_{i=1}^{N}\sum_{j=1}^{D_{\text{cat}}} m_{i,j}\,\text{CE}(\hat{y}_{i,j},\, y_{i,j})}{\sum_{i,j} m_{i,j}} + \frac{\sum_{i=1}^{N}\sum_{k=1}^{D_{\text{con}}} m_{i,k}\,\text{KL}(p_{i,k} \parallel q_{i,k})}{\sum_{i,k} m_{i,k}} \right) \tag{2}$$

3 Experiments and Results

Table 1. Comparison with SOTA methods on SLICE-3D, HOP, and EyePACS datasets. Best results are in **bold**, second best are underlined (Max pAUC = 0.200).

Method	SLICE-3D (ID)		SLICE-3D (OOD)		HOP		EyePACS	
	AUC LP/FT	pAUC LP/FT	AUC LP/FT	pAUC LP/FT	AUC LP/FT	pAUC LP/FT	ACC LP/FT	QWK LP/FT
Supervised Methods								
ViT-B[6]	0.910	0.138	0.832	0.096	0.837	0.091	0.513	0.523
CatBoost[18]	0.913	0.139	0.843	0.098	0.851	0.094	0.496	0.501
FT-Trans.[10]	0.890	0.131	0.780	0.083	0.814	0.085	0.466	0.453
DAFT[21]	0.921	0.142	0.870	0.123	0.876	0.121	0.526	0.528
TabPFN v2[12]	0.892	0.127	0.870	0.119	0.881	0.129	0.518	0.520
SSL Pre-training Methods								
SimCLR[4]	0.926/0.929	0.143/0.152	0.843/0.855	0.107/0.112	0.847/0.858	0.116/0.123	0.523/0.527	0.538/0.540
SCARF[1]	0.910/0.920	0.137/0.141	0.839/0.861	0.111/0.109	0.829/0.845	0.107/0.118	0.520/0.524	0.530/0.535
SAINT[20]	0.930/0.941	0.137/0.147	0.846/0.869	0.115/0.124	0.850/0.865	0.127/0.129	0.567/0.583	0.593/0.631
Dino v3[19]	0.923	0.135	0.869	0.121	0.885	0.134	0.561	0.580
MMCL[11]	0.951/0.956	0.168/0.170	0.871/0.889	0.122/0.125	0.889/0.871	0.128/0.131	0.652/0.660	0.698/0.704
TIP[7]	0.969/0.971	0.178/0.184	0.909/0.911	0.133/0.141	0.904/0.912	0.131/0.132	0.703/0.709	0.723/0.725
CITab[9]	0.958/0.964	0.171/0.173	0.890/0.902	0.127/0.130	0.906/0.911	0.133/0.136	0.689/0.695	0.720/0.723
AID (Ours)	**0.984/0.986**	**0.192/0.193**	**0.942/0.944**	**0.161/0.162**	**0.919/0.926**	**0.143/0.147**	**0.736/0.740**	**0.753/0.758**

Table 2. Ablation study on SLICE-3D, HOP, and EyePACS datasets.

Ablation Study	SLICE-3D (ID)		SLICE-3D (OOD)		HOP		EyePACS	
	AUC LP/FT	pAUC LP/FT	AUC LP/FT	pAUC LP/FT	AUC LP/FT	pAUC LP/FT	ACC LP/FT	QWK LP/FT
w/o SSL pre-training	0.927/0.927	0.143/0.143	0.878/0.878	0.132/0.132	0.848/0.848	0.097/0.097	0.646/0.646	0.632/0.632
w/o DR (ITC+ITM only)	0.961/0.969	0.167/0.172	0.933/0.937	0.151/0.155	0.895/0.906	0.132/0.135	0.695/0.699	0.706/0.713
w/o Adaptive Masking	0.959/0.966	0.168/0.170	0.929/0.930	0.149/0.149	0.891/0.898	0.130/0.133	0.688/0.697	0.689/0.701
w/o Discretization	0.968/0.973	0.172/0.183	0.936/0.939	0.155/0.157	0.904/0.911	0.133/0.141	0.704/0.709	0.711/0.713
Equal-width Discretization	0.946/0.950	0.163/0.165	0.903/0.914	0.136/0.142	0.887/0.892	0.126/0.130	0.685/0.684	0.696/0.689
Gaussian kernel	0.976/0.978	0.187/0.186	0.938/0.940	0.158/0.159	0.908/0.917	0.137/0.141	0.710/0.721	0.728/0.736
Hard-Label Discretization	0.972/0.977	0.183/0.185	0.931/0.939	0.151/0.157	0.910/0.916	0.136/0.140	0.708/0.716	0.718/0.731
AID (Ours)	**0.984/0.986**	**0.192/0.193**	**0.942/0.944**	**0.161/0.162**	**0.919/0.926**	**0.143/0.147**	**0.736/0.740**	**0.753/0.758**

3.1 Experimental Setup

The primary dataset, SLICE-3D[14], comprises over 400,000 standardized lesion image tiles from 1,042 patients, positive-to-negative sample ratio 1‰. To ensure rigorous evaluation, 80,433 samples from 220 patients are geographically isolated as an Out-of-Domain (OOD) test set and strictly excluded from pre-training. The remaining 320,626 images from 822 patients are split into an 85/15 ratio for pre-training and validation. For downstream fine-tuning (FT) and linear probing (LP), 10,228 samples from the 822 patients are reserved as the In-Domain (ID) test set, while the remaining 310,398 samples undergo a 5-fold patient-stratified cross-validation. For downstream evaluation, unimodal and multimodal representations are processed by a three-head ensemble mechanism that averages distinct classification logits to generate final diagnostic predictions. Evaluations are reported on both the ID and OOD test sets. External validation utilizes HOP, a private dataset containing 208,540 images from 284 patients. Furthermore,

the EyePACS dataset[8], comprising 88,702 retinal fundus images processed via the AutoMorph pipeline[26], provides a rigorous cross-specialty ophthalmology benchmark. While the UK Biobank dataset[3] is inaccessible due to restricted paywalled access, the EyePACS dataset is open-source, and we will publicly release its corresponding tabular data. During downstream evaluation, class imbalance is mitigated through weighted sampling. Evaluation metrics include AUROC, Accuracy, Quadratic Weighted Kappa (QWK)[8], and partial AUC at 80% sensitivity (pAUC@80%)[15], prioritizing clinical specificity. The pre-training hyperparameter configuration utilizes a base masking rate $r_{\text{base}} = 0.1$, an importance scaling factor and maximum rate $\alpha = r_{\max} = 0.7$, and 50 quantization bins. All hyperparameter settings were empirically established through extensive validation experiments.

3.2 Main Results

As presented in Table 1, all baseline experiments were rigorously reproduced under identical experimental settings. The empirical results demonstrate that the proposed framework establishes a new SOTA. On the critical OOD geographic holdout, the model achieves a full fine-tuning AUC of 0.944 and a pAUC of 0.162, significantly outperforming the 0.911 AUC of the strongest semantic-agnostic baseline, TIP. This substantial margin confirms that the semantic-centric design generates representations highly resilient to geographical distribution shifts. Furthermore, on the ID SLICE-3D evaluation, the linear probe performance reaches an AUC of 0.984, which surpasses the 0.971 full fine-tuning AUC of the TIP baseline. This specific outcome indicates that the multimodal feature space produced by the proposed pre-training strategy possesses exceptional linear separability without requiring deep downstream optimization.

Evaluation on the private HOP dataset confirms that these performance gains remain consistent when tested on completely unseen patient populations, achieving a fine-tuning AUC of 0.926. Applying the framework to the EyePACS dataset further validates the universality of the two-dimensional structural hypothesis. In this challenging ophthalmic domain, the model achieves an accuracy of 0.740 and a QWK of 0.758. This cross-specialty success succinctly demonstrates that modeling feature importance and intra-feature duality provides fundamental benefits for multimodal clinical data integration across diverse medical disciplines.

Under our 5-fold full fine-tuning protocol, AID exhibits per-fold standard deviation approximately half of TIP's: on SLICE-3D OOD, AUC 0.944 ± 0.006 versus 0.911 ± 0.013; on EyePACS, QWK 0.758 ± 0.007 versus 0.725 ± 0.016. The AID–TIP gap exceeds fold-to-fold variation by multiple standard deviations, supporting that gains stem from semantic-aware pretext design rather than favorable fold splits.

3.3 Ablation Studies & Qualitative Analysis

The ablation study (Table 2) isolates the contribution of each proposed module. Removing adaptive masking decreases the out-of-domain (OOD) AUC from

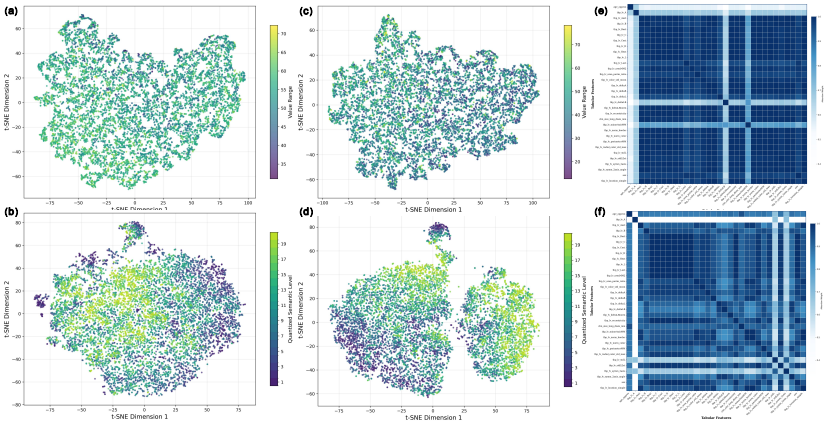

Fig. 3. Qualitative comparison between the SOTA baseline (top row) and the proposed framework (bottom row). (a) and (b) visualize t-SNE projections for color feature (tbp_lv_H) reconstructions, while (c) and (d) show t-SNE projections for shape feature (tbp_lv_L) reconstructions. (e) and (f) present self-attention weight heatmaps from the tabular pathway of the multimodal encoder.

0.944 to 0.930, confirming that the label-free curriculum guides the model to prioritize diagnostically salient features. Evaluating intra-feature structural modeling provides critical insights. Replacing the soft-label objective with standard continuous regression yields an OOD AUC of 0.939. Applying equal-width discretization degrades performance to 0.914, demonstrating that naive binning creates heavily skewed semantic targets. Hard-label quantile discretization recovers the AUC to 0.939. Notably, substituting the proposed triangle kernel with a computationally heavier Gaussian kernel for soft-label generation achieves an AUC of 0.940. However, the proposed soft-label binning module achieves the peak OOD performance of 0.944, demonstrating that the additional computational overhead of the Gaussian kernel does not translate into diagnostic gains over the simpler triangle kernel. This numerical progression provides evidence that dynamically distributing probability mass across adjacent bins via the simple triangle kernel mathematically preserves the inherent ordinal relationships of clinical measurements.

As shown in Fig. 3, visualizations of the learned representations and attention mechanisms corroborate the quantitative improvements. The t-SNE projections contrast the semantic-agnostic baseline against the proposed framework during tabular feature reconstruction. The baseline model maps continuous values into a chaotic, unorganized latent space. Conversely, the soft-label objective successfully transforms this continuous regression space into highly ordered, distinct medical semantic clusters. Furthermore, self-attention heatmaps from the tabular pathway reveal distinct behavioral differences. While the semantic-agnostic baseline exhibits a diffuse, disorganized pan-attention pattern, the proposed model learns a highly structured and importance-aware attention alloca-

tion mechanism. This refined attention distribution confirms that the framework efficiently assigns focus based on clinical relevance, faithfully reflecting the hierarchical logic utilized in human diagnostic reasoning.

4 Conclusion

We introduce AID, a semantic-aware multimodal pre-training framework that models the inherent two-dimensional structure of medical tabular data. By integrating importance-aware adaptive masking and soft-label discretization, the framework effectively overcomes traditional semantic-agnostic limitations. Evaluations confirm that this framework achieves robust cross-domain generalization and state-of-the-art performance across diverse medical disciplines.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bahri, D., Jiang, H., Tay, Y., Metzler, D.: Scarf: Self-supervised contrastive learning using random feature corruption. arXiv preprint arXiv:2106.15147 (2021)
2. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4009–4018 (2021)
3. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al.: The uk biobank resource with deep phenotyping and genomic data. Nature **562**(7726), 203–209 (2018)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
6. 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)
7. Du, S., Zheng, S., Wang, Y., Bai, W., O'Regan, D.P., Qin, C.: Tip: Tabular-image pre-training for multimodal classification with incomplete data. In: European Conference on Computer Vision. pp. 478–496. Springer (2024)
8. Dugas, E., Jared, Jorge, Cukierski, W.: Diabetic retinopathy detection. https://kaggle.com/competitions/diabetic-retinopathy-detection (2015), kaggle
9. Fu, Y., Zhao, Y., Zeng, Z., Chen, C., Jin, Y.: Unleashing the power of image-tabular self-supervised learning via breaking cross-tabular barriers. arXiv preprint arXiv:2512.14026 (2025)

10. Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A.: Revisiting deep learning models for tabular data. Advances in neural information processing systems **34**, 18932–18943 (2021)
11. Hager, P., Menten, M., Rueckert, D.: Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23924–23935 (2023)
12. Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S.B., Schirrmeister, R.T., Hutter, F.: Accurate predictions on small data with a tabular foundation model. Nature **637**(8045), 319–326 (2025)
13. Huang, S.C., Pareek, A., Seyyedi, S., Banerjee, I., Lungren, M.P.: Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ digital medicine **3**(1), 136 (2020)
14. Kurtansky, N.R., D'Alessandro, B.M., Gillis, M.C., Betz-Stablein, B., Cerminara, S.E., Garcia, R., Girundi, M.A., Goessinger, E.V., Gottfrois, P., Guitera, P., et al.: The slice-3d dataset: 400,000 skin lesion image crops extracted from 3d tbp for skin cancer detection. Scientific Data **11**(1), 884 (2024)
15. Kurtansky, N.R., Gillis, M.C., Codella, N.C., D'Alessandro, B.M., Ge, Z., Guitera, P., Halpern, A.C., Kittler, H., Malvehy, J., Liopyris, K., et al.: Automated triage of cancer-suspicious skin lesions with 3d total-body photography. npj Digital Medicine **8**(1), 708 (2025)
16. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
17. Li, X., Yan, S., Liu, Y., Soyer, H.P., Janda, M., Mar, V., Ge, Z.: Multi-aspect knowledge-enhanced medical vision-language pretraining with multi-agent data generation. arXiv preprint arXiv:2512.03445 (2025)
18. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: Catboost: unbiased boosting with categorical features. Advances in neural information processing systems **31** (2018)
19. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
20. Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C.B., Goldstein, T.: Saint: Improved neural networks for tabular data via row attention and contrastive pretraining. arXiv preprint arXiv:2106.01342 (2021)
21. Wolf, T.N., Pölsterl, S., Wachinger, C., Initiative, A.D.N., et al.: Daft: A universal module to interweave tabular data and 3d images in cnns. NeuroImage **260**, 119505 (2022)
22. Yan, J., Zheng, B., Xu, H., Zhu, Y., Chen, D., Sun, J., Wu, J., Chen, J.: Making pretrained language models great on tabular prediction. In: International Conference on Learning Representations. vol. 2024, pp. 23570–23593 (2024)
23. Yan, S., Hu, M., Jiang, Y., Li, X., Fei, H., Tschandl, P., Kittler, H., Ge, Z.: Derm1m: A million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12681–12690 (2025)
24. Yan, S., Li, X., Hu, M., Jiang, Y., Yu, Z., Ge, Z.: Make: Multi-aspect knowledge-enhanced vision-language pretraining for zero-shot dermatological assessment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 369–379. Springer (2025)

25. Yan, S., Li, X., Mo, D., Tschandl, P., Jiang, Y., Wang, Z., Hu, M., Ju, L., Alonso, C., Zheng, Y., et al.: A vision-language foundation model for zero-shot clinical collaboration and automated concept discovery in dermatology (2026)
26. Zhou, Y., Wagner, S.K., Chia, M.A., Zhao, A., Xu, M., Struyven, R., Alexander, D.C., Keane, P.A., et al.: Automorph: automated retinal vascular morphology quantification via a deep learning pipeline. Translational vision science & technology **11**(7), 12–12 (2022)