Decoupling Clinical and Class-Agnostic Features for Reliable Few-Shot Adaptation under Shift

Umaima Rahman[1], Raza Imam[1], Mohammad Yaqub[1], and Dwarikanath Mahapatra[2]

[1] Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[2] Khalifa University, Abu Dhabi, UAE
umaima.rahman@mbzuai.ac.ae

Abstract. Medical vision-language models (VLMs) offer promise for clinical decision support, yet their reliability under distribution shifts remains a major concern for safe deployment. These models often learn task-agnostic correlations due to variability in imaging protocols and free-text reports, limiting their generalizability and increasing the risk of failure in real-world settings. We propose *DRiFt*, a structured feature decoupling framework that explicitly separates clinically relevant signals from task-agnostic noise using parameter-efficient tuning (LoRA) and learnable prompt tokens. To enhance cross-modal alignment and reduce uncertainty, we curate high-quality, clinically grounded image-text pairs by generating captions for a diverse medical dataset. Our approach improves in-distribution performance by +11.4% Top-1 accuracy and +3.3% Macro-F1 over prior prompt-based methods, while maintaining strong robustness across unseen datasets. Ablation studies reveal that disentangling task-relevant features and careful alignment significantly enhance model generalization and reduce unpredictable behavior under domain shift. These insights contribute toward building safer, more trustworthy VLMs for clinical use. The code is available at https://github.com/rumaima/DRiFt.

Keywords: OOD Generalization · Medical VLMs · Distribution Shifts

1 Introduction

Recent developments in healthcare rely increasingly on automated medical imaging analysis to support critical diagnoses and treatment decisions [1]. However, real-world clinical scenarios often present substantial variability [2] ranging from differences in patient demographics and imaging protocols to heterogeneous textual descriptions from radiology reports [3]. These variations introduce *spurious correlations* that can confound traditional vision-language models and degrade their performance when encountering domain shifts [4]. Motivated by this, we propose a feature decoupling framework for vision-language models that targets both the visual and textual domains, explicitly separating clinically relevant, i.e., *invariant* features, from spurious artifacts. As illustrated in Fig. 1, non-clinical details such as the presence of clothing shadows or incidental notes about patient

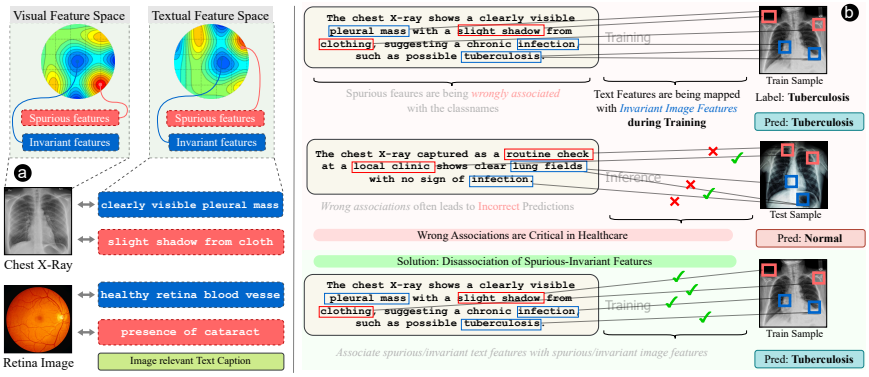

Fig. 1: Impact of spurious correlations in medical imaging. **(a)** Conceptual illustration of multi-modal decoupling. **(b)** Shows how spurious features can dominate embeddings, leading to misclassification, critical in healthcare contexts.

scheduling, can mislead a model if these spurious elements become incorrectly associated with diagnostic findings (e.g., lung infiltrates) [5].

Conventional fine-tuning strategies such as [6,7] that treat vision and text holistically risk conflating essential signals with irrelevant context, undermining spurious relationships which can be very sensitive in high-stakes medical settings. Building on [8], who align unpaired images and texts for domain robustness, we leverage caption-driven supervision to enhance clinical relevance while maintaining feature disentanglement. Inspired by the success of our prior work on disentangled prompt learning in natural images [9] and its relevance to medical imaging, we extend the approach to address clinical domain shift. We aim to enforce a structured decoupling of features in both modalities. Crucially, spurious image features must map to spurious text descriptors, while clinically pertinent cues in the image remain aligned with relevant textual annotations. Through parameter-efficient fine-tuning (via LoRA [10,11]) and systematic image-caption curation within MedIMeta [12], we show that this *decoupled approach* not only improves generalization under cross-dataset and domain generalization conditions but also reduces the computational burden of adapting large multi-modal models [13]. Our contributions can be summarized as follows:

1. **Structured Feature Decoupling Across Modalities:** We design a framework that separates image and text embeddings into invariant and spurious components, ensuring that clinically relevant signals remain unaffected by domain-specific artifacts. This minimizes erroneous cross-modal associations and enhances generalization in healthcare settings.
2. **Parameter-Efficient LoRA Fine-Tuning:** To address computational constraints, we employ LoRA-based fine-tuning, updating only low-rank adapter layers instead of the entire model. By optimizing contrastive and cross-modal alignment losses, our framework mitigates domain biases and improves robustness to out-of-distribution data.
3. **Enriching MedIMeta with Caption Generation for Few-Shot Training:** We leverage the MedIMeta dataset covering multiple imaging modal-

ities to generate clinically relevant captions. This ensures the model learns invariant medical patterns without being misled by domain-specific artifacts.

Related Works: As healthcare data diversifies, domain generalization strategies [14,15,16] become essential for robust performance across patient populations and imaging conditions. Modern vision-language models (VLMs) enable few-shot or zero-shot inference by jointly encoding text and images but risk carrying over spurious correlations [17,18,19]. While prompt engineering provides a parameter-efficient adaptation strategy [20], it rarely disentangles clinically relevant features from domain-specific noise, leading to overfitting [21]. Recent unified multi-modal frameworks [6] enhance image-text alignment but seldom separate spurious correlations from invariant clinical markers across both modalities, limiting reliability in real-world diagnostics. In contrast, our work emphasizes *explicit feature decoupling* to reduce reliance on non-clinical details, ensuring more robust and interpretable medical learning algorithms.

2 Methodology

Medical imaging data in *MedIMeta* [12] spans multiple modalities (e.g., chest X-ray, fundus, ultrasound), each subject to domain-specific artifacts and protocol variations. We introduce a decoupled multi-modal framework that separates invariant, clinically relevant features from spurious correlations in both vision and text modalities. This is achieved through caption generation and parameter-efficient adaptation using LoRA-based fine-tuning in a few-shot setting as illustrated in Fig. 2

2.1 Workflow

Caption Generation for MedIMeta: We generate captions for selected datasets within MedIMeta, such as skin-lesion and fundus images, ensuring a minimum image count threshold (e.g., 100+) for diversity. Captions are generated using InstructBLIP [13], prioritizing clinical relevance while preserving minor spurious details. Among these captions, we validated their clinical relevance, removing those that were irrelevant while retaining the meaningful contextual elements. A subset of these captions is selected for a few-shot setting, aiding the model in distinguishing clinically meaningful information from spurious correlations.

Decoupled Feature Representation: To enhance generalization, we decompose each image and text embedding into invariant and spurious components [9]. Given an input embedding $\mathbf{z}_{\mathbf{M}i}$, two projection functions are defined by $\mathbf{z}_{\mathbf{M}i,u} = \phi_M(\mathbf{z}_{\mathbf{M}i})$, where $M = \mathbf{v}, \mathbf{t}$, with $\mathbf{v}$ and $\mathbf{t}$ representing visual and textual embeddings, respectively. This decomposition ensures clinically meaningful features are preserved while mitigating domain-specific artifacts.

Multi-Modal Low-Rank Adaptation for Image-Caption Learning: We introduce *LoRA* adapters [10] to efficiently update the query (Q) and key (K) projections of the vision and text encoders. Instead of fine-tuning the entire

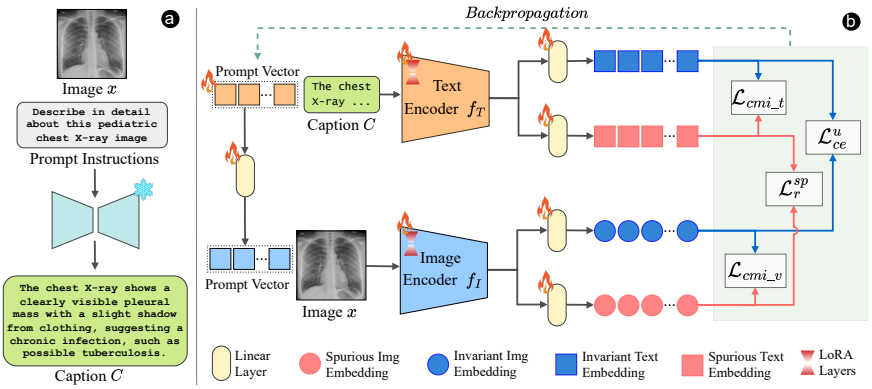

Fig. 2: DRiFt: Our proposed feature disentanglement framework. **(a)** Captions are generated for images from the MedIMeta dataset. **(b)** Multi-modal feature decoupling is applied using LoRA fine-tuning and learnable prompts to focus on invariant features.

Transformer layers, LoRA injects low-rank matrices into the original weight matrix $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, decomposing it as, $W_\theta = W_0 + \alpha AB$, where $A \in \mathbb{R}^{d_{\text{out}} \times r}$, $B \in \mathbb{R}^{r \times d_{\text{in}}}$, and $\theta = \{A, B\}$ denotes the trainable LoRA parameters. Alongside LoRA, we integrate *learnable prompt tokens* $\gamma = \{p_1, p_2, ..., p_m\}$, where $P_\gamma \in \mathbb{R}^{m \times d}$, inserted at the input layer to guide the model towards clinically meaningful patterns while suppressing spurious correlations [22]. The overall trainable parameter set is $\Theta = \theta \cup \gamma$.

2.2 Objectives

To ensure the model captures clinically relevant features, we define an objective function comprising:

1. Invariant Alignment: A contrastive loss $\mathcal{L}^u_{ce}$ aligns visual and textual invariant embeddings:

$$\mathcal{L}^u_{ce} = -\sum_{i=1}^{N} y_i \log p^u_\Theta(y_i \mid x_i), \quad p^u_\Theta(y_i \mid x_i) = \frac{\exp(\text{sim}(\mathbf{z_v}_{i,u}, \mathbf{z_t}_{i,u})/\tau)}{\sum_{j=1}^{C} \exp(\text{sim}(\mathbf{z_v}_{j,u}, \mathbf{z_t}_{j,u})/\tau)}.$$

2. Spurious Neutralization: A KL-divergence loss $\mathcal{L}^{sp}_r$ prevents spurious features across both modalities from influencing classification:

$$\mathcal{L}^{sp}_r = \sum_{i=1}^{N} \ell_{KL}(p^s_\Theta(y_i \mid x_i) \,||\, p_0). \tag{1}$$

3. Reduced Statistical Dependence: We enforce conditional independence between the invariant and textual features across vision and textual modalities using the Conditional Hilbert-Schmidt Independence Criterion:

$$\mathcal{L}_{con_v} = I(\mathbf{z_v}_{i,u}; \mathbf{z_v}_{i,s} \mid Y), \quad \mathcal{L}_{con_t} = I(\mathbf{z_t}_{i,u}; \mathbf{z_t}_{i,s} \mid Y). \tag{2}$$

The overall loss function integrates these components:

$$\mathcal{L} = \mathcal{L}^u_{ce} + \alpha \mathcal{L}^{sp}_r + \beta \mathcal{L}_{con}, \quad \text{where} \quad \mathcal{L}_{con} = \text{avg}(\mathcal{L}_{con_v}, \mathcal{L}_{con_t}). \tag{3}$$

Table 1: Comparison of Top-1 Accuracy, Macro-F1, and AUC on in-distribution data using our methodology - DRiFt with respect to other multi-modal prompt learning approaches adapted to an image-caption pair setting for fair comparison.

# Classes	Methods →	MaPLe-IC			CoOp-OOD-IC			DRiFt		
	MedIMeta Dataset [12]	Acc.	Macro-F1	AUC	Acc.	Macro-F1	AUC	Acc.	Macro-F1	AUC
3	bus	11.5	10.1	29.3	17.3	9.8	**50.5**	**42.3**	**21.7**	34.9
15	skinl-derm	1.3	0.2	**52.1**	4.1	2.1	48.0	**21.5**	**3.2**	49.5
7	derm	34.1	10.1	**46.8**	16.9	6.1	45.6	**57.3**	**11.6**	46.5
2	glaucoma	84.8	45.9	56.3	15.2	13.2	**60.7**	**82.1**	**52.2**	51.2
2	fundus	21.1	17.5	**68.7**	20.9	17.3	21.1	**22.2**	**19.0**	59.1
3	pneumonia	30.3	26.3	47.6	37.5	18.2	28.4	**37.5**	18.2	**52.4**
2	mammo_mass	**50.5**	**49.1**	**46.5**	38.6	27.9	42.1	50.3	48.3	45.9
2	mammo_calc	41.4	40.8	48.0	41.1	32.9	**60.4**	**52.8**	**52.0**	53.3
	average	34.4	25.0	**49.4**	24.0	15.9	44.6	**45.8**	**28.3**	49.1

Table 2: Cross-dataset evaluation comparing MaPLe-IC, CoOp-OOD-IC, DRiFt.

Method→	**MaPLe-IC**		**CoOp-OOD-IC**		**DRiFt**	
MedIMeta Dataset	Target	Accuracy	Target	Accuracy	Target	Accuracy
skinl_derm	avg(derm, isic)	6.4	avg(derm, isic)	6.5	avg(derm, isic)	10.4
derm	avg(skinl_derm, isic)	18.9	avg(skinl_derm, isic)	18.8	avg(skinl_derm, isic)	17.8
glaucoma	avg(fundus, idrid)	24.9	avg(fundus, idrid)	25.7	avg(fundus, idrid)	24.8
fundus	avg(glaucoma, idrid)	55.6	avg(glaucoma, idrid)	55.6	avg(glaucoma, idrid)	56.4
pneumonia	avg(s-tb, m-tb)	54.3	avg(s-tb, m-tb)	51.8	avg(s-tb, m-tb)	52.2

where α and β regulate spurious suppression and statistical independence respectively while $\mathcal{L}_{con}$ is the independence loss. By leveraging curated captions, decoupled embeddings, LoRA-enhanced prompts, and the above objective functions, DRiFt achieves comparatively better generalization under real-world distribution shifts in medical imaging as seen in Figure 4(left).

3 Experimentation and Results

3.1 Experimental Setup

Datasets: We assess DRiFt on 8 different MedIMeta datasets [12] belonging to 6 different medical modalities, such as chest-Xray, breast ultrasound, mammography, dermatoscopy, and fundus images. We also evaluate our models in a cross-dataset generalization setting for the datasets S-TB [23], M-TB [24], IDRID [25], and ISIC [26]. We choose *MedIMeta* dataset as it encompasses multi-classification tasks which provides a rigorous test bed for evaluating model resilience to distribution shifts to ensure better cross-domain performance in real-world clinical scenarios [27].

Baselines: We compare DRiFt against two other multi-modal foundation methods aimed at out-of-distribution (OOD) generalization in the vision domain. In particular, we adapt *MaPLe* [6] and *CoOp-OOD* [28] to accommodate image-caption pairs and integrate LoRA-based fine-tuning for parameter efficiency. MaPLe originally improved generalization by conditioning visual prompts

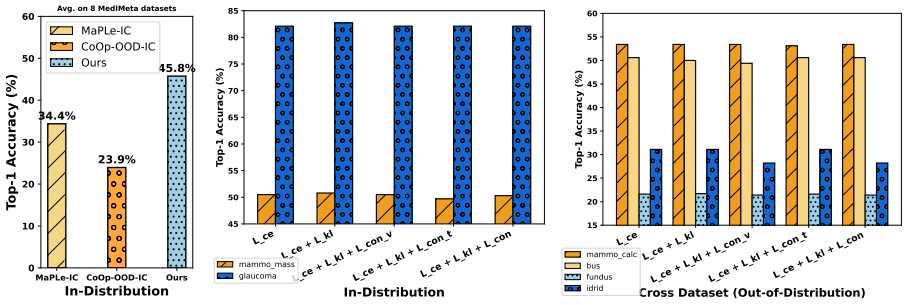

Fig. 3: **(left)** In-distribution performance comparison across MaPLe-IC, CoOp-OOD-IC, and DRiFt (Ours), with DRiFt achieving the highest average Top-1 accuracy (45.8%) across 8 MedIMeta tasks. **(middle)** Impact of different loss components on in-distribution performance, showing Top-1 accuracy variations for the mammo_mass and glaucoma tasks. **(right)** Cross-dataset evaluation of loss components, highlighting Top-1 accuracy changes for mammo_calc, bus, fundus, and IDRID.

on textual prompts, while CoOp-OOD enhanced robustness by decoupling visual features and aligning invariant representations with text. However, these methods do not explicitly separate spurious features within their textual counterparts. To address this, we introduce a unified decoupling framework that ensures clinically meaningful features are preserved while mitigating domain-specific artifacts.

Implementation Details: DRiFt adopts a few-shot learning setup by selecting 16 random samples per class for training. A pre-trained ViT-B/16-based vision-language model serves as the backbone, using dimension settings of $d_l = 512$ for language, $d_v = 768$ for vision, and $d_{vl} = 512$ for combined embeddings. In total, we employ $J = 3$ transformer layers of prompt tuning, each with two learnable prompt tokens for vision and language. We train on a single NVIDIA A100 GPU for 10 epochs using stochastic gradient descent with a batch size of 4 and an initial learning rate of 2.6e-6. For initialization, the first-layer text prompt vectors are taken from a pre-trained word embedding (a photo of a <class>) and concatenated with the caption corresponding to the image, while subsequent layers are randomly initialized from a normal distribution.

Evaluation Metrics: Table 1 presents the evaluation of different medical datasets using *Accuracy*, *Macro F1-score*, and *ROC-AUC score*. Accuracy is the percentage of correctly classified instances and provides an overall performance measure, howvere, it may be misleading for imbalanced datasets. Macro F1-score is the average of F1-scores across all classes. It treats all classes equally, making it more informative when class distributions are skewed. Furthermore, a higher ROC-AUC score indicates better classification performance and measures the model's ability to distinguish between classes by plotting the true positive rate (TPR) against the false positive rate (FPR) at different thresholds.

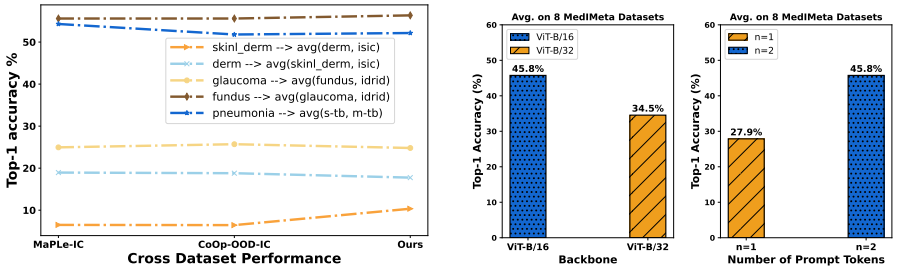

Fig. 4: **(left)** Comparison of cross-dataset generalization for MaPLe-IC, CoOp-OOD-IC, and DRiFt (Ours). **(middle)** Comparison of ViT-B/16 and ViT-B/32 backbones on MedIMeta datasets, showing that ViT-B/16 achieves higher average Top-1 accuracy. **(right)** Increasing the number of prompt tokens from 1 to 2 significantly improves the average Top-1 accuracy across MedIMeta datasets, from 27.9% to 45.8%.

3.2 Main Results

In-Distribution Evaluation: As shown in Table 1, DRiFt outperforms MaPLe and CoOpOOD across multiple datasets, achieving the highest accuracy in bus (42.3%), skinl_derm (21.5%), and mammo_calc (52.8%). It also achieves the best Macro-F1 in glaucoma (52.2%), mammo_calc (52%) and bus (21.7%), ensuring better class balance. While pneumonia accuracy improves (37.5%), its low Macro-F1 (18.2%) highlights challenges in minority class predictions. On average DRiFt surpasses MaPLe and CoOp-OOD with 45.75% accuracy and 28.275 Macro-F1, demonstrating superior generalization in medical classification.

Cross-Dataset Evaluation: The cross-dataset setting assesses model generalization by training on one dataset and testing on another with similar modalities but distinct characteristics. This evaluates how well the model adapts to variations in imaging protocols and anatomical structures. For skin lesion classification, DRiFt outperforms MaPLe and CoOp-OOD on skinl_derm (10.35% vs. 6.5% and 6.45%) but slightly underperforms on derm (17.75% vs. 18.95%) as tabulated in Table 2.

4 Ablation Study

a. Effect of Number of shots: Increasing the number of shots generally improves performance (Fig. 5(left)), particularly in skinl_derm (7.8% to 21.5%), derm (57.3% stable), and bus (50% to 42.3%), suggesting better class separability with more data. Glaucoma (83.9% to 82.1%) remains mostly unchanged, while pneumonia (37.5%) is entirely unaffected, indicating that some tasks do not benefit from additional samples. This suggest that while increased shots enhance fine-grained classification, its efficacy depends on dataset complexity and class distribution.

b. Effect of Number of prompt tokens: Using two prompt tokens ($n = 2$) consistently improves performance over a single token ($n = 1$), particularly in bus (17.3% to 42.3%), skinl_derm (8.1% to 21.5%), and derm (30.9% to 57.3%), suggesting that additional tokens enrich feature extraction. Similarly,

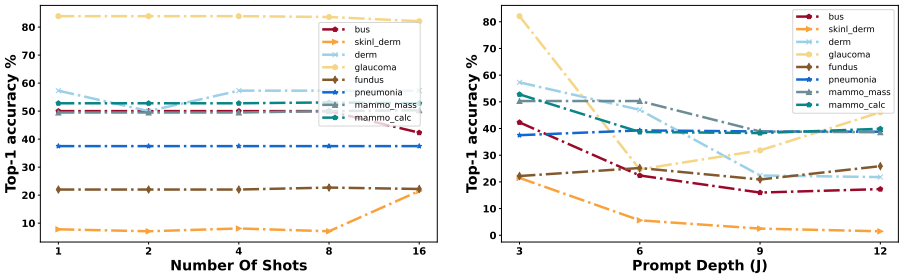

Fig. 5: **(left)** Influence of the number of training shots (1, 2, 4, 8, 16), showing that additional shots improve performance for some datasets while others remain stable. **(right)** Impact of varying prompt depth (J) on Top-1 accuracy across datasets, highlighting performance degradation with increasing depth.

mammo_mass (38.6% to 50.3%) and mammo_calc (41.1% to 52.8%) benefit from more tokens, aiding complex imaging tasks. However, fundus (33.6% to 22.2%) experiences a decline, likely due to redundant or non-discriminative information, while pneumonia (37.5%) remains unaffected. On average, increasing tokens enhances generalization, but tuning is essential based on the dataset.

c. Effect of Number of prompt depth: Increasing prompt depth generally degrades performance, with bus (42.3% to 16%), skinl_derm (21.5% to 1.5%), and derm (57.3% to 21.8%) showing steep declines. Glaucoma (82.1% to 24.7%) initially drops but partially recovers at $J = 12$ (46.1%), indicating instability. While some datasets, like fundus (22.2% to 25.9%), show marginal gains, deeper prompts tend to introduce redundancy, making careful selection crucial.

d. Effect of backbone: ViT-B/16 consistently outperforms ViT-B/32 in bus (42.3% vs. 9%), skinl_derm (21.5% vs. 4.8%), and derm (57.3% vs. 12.3%), demonstrating the advantage of higher spatial resolution in fine-grained tasks. However, glaucoma (82.1% vs. 85.1%) remains unaffected, and mammo_mass (50.3% vs. 60.8%) and mammo_calc (52.8% vs. 59.8%) favor ViT-B/32, suggesting that backbone choice should depend on task requirements, balancing local detail extraction and global contextual representation.

e. Effect of loss components: We evaluate the impact of different loss components on the mammo-mass and glaucoma datasets. Using only cross-entropy loss (L_{ce}) achieves a baseline accuracy of 50.5% for mammo-mass and 82.1% for glaucoma, while adding KL divergence regularization (L_{kl}) improves performance to 50.8% and 82.7%, respectively. Furthermore, to assess cross-dataset generalization, we train on one dataset and test on related datasets. For mammo-mass, transferring to mammo-calc maintains stable performance (53.4%) across all configurations. However, on bus, performance declines when adding contrastive loss, indicating that contrastive learning may not always enhance generalization.

5 Conclusion

We propose a structured decoupling framework for medical vision-language models that improves out-of-distribution generalization by isolating invariant clinical

features from task-agnostic correlations in both image and text. Using LoRA-based fine-tuning and learnable prompts, DRiFt enhances cross-modal alignment with minimal overhead. By disentangling features at both visual and textual levels, the model focuses on clinically relevant cues while suppressing noise, outperforming existing multi-modal prompt-tuning approaches. This work underscores the importance of targeted loss design and structured feature separation for robust generalization in medical AI.

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

References

1. Vitali Sintchenko and Enrico W Coiera. Which clinical decisions benefit from automation? a task complexity approach. *International journal of medical informatics*, 70(2-3):309–316, 2003.
2. Umaima Rahman, Guangyi Chen, and Kun Zhang. Mrishift: Disentangled representation learning for 3d mri lesion segmentation under distributional shifts. In *2024 12th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6. IEEE, 2024.
3. Samuel J White, Qi Sheng Phua, Lucy Lu, Kaspar L Yaxley, Matthew DF Mcinnes, and Minh-Son To. Heterogeneity in systematic reviews of medical imaging diagnostic test accuracy studies: a systematic review. *JAMA Network Open*, 7(2):e240649–e240649, 2024.
4. Khaled Saab, Sarah Hooper, Mayee Chen, Michael Zhang, Daniel Rubin, and Christopher Ré. Reducing reliance on spurious features in medical image classification with spatial specificity. In *Machine Learning for Healthcare Conference*, pages 760–784. PMLR, 2022.
5. Rhys Compton, Lily Zhang, Aahlad Puli, and Rajesh Ranganath. When more is less: Incorporating additional datasets can hurt performance by introducing spurious correlations. In *Machine Learning for Healthcare Conference*, pages 110–127. PMLR, 2023.
6. Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122, 2023.
7. Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.
8. Umaima Rahman, Raza Imam, Mohammad Yaqub, Boulbaba Ben Amor, and Dwarikanath Mahapatra. Can language-guided unsupervised adaptation improve medical image classification using unpaired images and texts? In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2025.
9. Umaima Rahman, Mohammad Yaqub, and Dwarikanath Mahapatra. Dimple–disentangled multi-modal prompt learning: Enhancing out-of-distribution alignment with invariant and spurious feature separation. *arXiv preprint arXiv:2506.21237*, 2025.
10. Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

11. Raza Imam, Hanan Gani, Muhammad Huzaifa, and Karthik Nandakumar. Test-time low rank adaptation via confidence maximization for zero-shot generalization of vision-language models. *arXiv preprint arXiv:2407.15913*, 2024.
12. Stefano Woerner, Arthur Jaques, and Christian F Baumgartner. A comprehensive and easy-to-use multi-domain multi-task medical imaging meta-dataset (medimeta). *arXiv preprint arXiv:2404.16000*, 2024.
13. Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.
14. Chenxin Li, Xin Lin, Yijin Mao, Wei Lin, Qi Qi, Xinghao Ding, Yue Huang, Dong Liang, and Yizhou Yu. Domain generalization on medical imaging classification using episodic training with task augmentation. *Computers in biology and medicine*, 141:105144, 2022.
15. Jee Seok Yoon, Kwanseok Oh, Yooseung Shin, Maciej A Mazurowski, and Heung-Il Suk. Domain generalization for medical image analysis: A review. *Proceedings of the IEEE*, 2024.
16. Lin Lawrence Guo, Stephen R Pfohl, Jason Fries, Alistair EW Johnson, Jose Posada, Catherine Aftandilian, Nigam Shah, and Lillian Sung. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Scientific reports*, 12(1):2726, 2022.
17. Fereshteh Shakeri, Yunshi Huang, Julio Silva-Rodríguez, Houda Bahig, An Tang, Jose Dolz, and Ismail Ben Ayed. Few-shot adaptation of medical vision-language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 553–563. Springer, 2024.
18. Zhengfeng Lai, Zhuoheng Li, Luca Cerny Oliveira, Joohi Chauhan, Brittany N Dugger, and Chen-Nee Chuah. Clipath: Fine-tune clip with visual feature fusion for pathology image analysis towards minimizing data collection efforts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2374–2380, 2023.
19. Raza Imam, Asif Hanif, Jian Zhang, Khaled Waleed Dawoud, Yova Kementchedjhieva, and Mohammad Yaqub. Noise is an efficient learner for zero-shot vision-language models. *arXiv preprint arXiv:2502.06019*, 2025.
20. Umaima Rahman, Raza Imam, Dwarikanath Mahapatra, and Boulbaba Ben Amor. Meduna: Language guided unsupervised adaptation of vision-language models for medical image classification. *arXiv preprint arXiv:2409.02729*, 2024.
21. Xiao Liu, Pedro Sanchez, Spyridon Thermos, Alison Q O'Neil, and Sotirios A Tsaftaris. Learning disentangled representations in the imaging domain. *Medical Image Analysis*, 80:102516, 2022.
22. Gita Sarafraz, Armin Behnamnia, Mehran Hosseinzadeh, Ali Balapour, Amin Meghrazi, and Hamid R Rabiee. Domain adaptation and generalization of functional medical data: A systematic survey of brain data. *ACM Computing Surveys*, 56(10):1–39, 2024.
23. Shenzhen Hospital. Shenzhen hospital chest x-ray (cxr) set. https://data.lhncbc.nlm.nih.gov/public/Tuberculosis-Chest-X-ray-Datasets/Shenzhen-Hospital-CXR-Set/index.html.
24. Montgomery County. Montgomery county chest x-ray (cxr) set. https://data.lhncbc.nlm.nih.gov/public/Tuberculosis-Chest-X-ray-Datasets/Montgomery-County-CXR-Set/MontgomerySet/index.html.
25. Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy

image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018.
26. International Skin Imaging Collaboration. Isic archive: International skin imaging collaboration dataset. https://www.isic-archive.com/.
27. Raza Imam, Rufael Marew, and Mohammad Yaqub. On the robustness of medical vision-language models: Are they truly generalizable? *arXiv preprint arXiv:2505.15425*, 2025.
28. Jie Zhang, Xiaosong Ma, Song Guo, Peng Li, Wenchao Xu, Xueyang Tang, and Zicong Hong. Amend to alignment: Decoupled prompt tuning for mitigating spurious correlation in vision-language models. In *Forty-first International Conference on Machine Learning*.