# Towards Class-Aware Semantic Decoupling in Multi-Modal Disease Diagnosis with Missing Modalities

Feixiang Zhou[1], Jianyang Xie[1], Zhuangzhi Gao[1], Qinkai Yu[2], Jing Li[3], Long Chen[4], Zheheng Jiang[5], Yitian Zhao[6], Yanda Meng[7], He Zhao[1], Gregory Y.H. Lip[8], and Yalin Zheng[1](✉)

[1] Department of Eye and Vision Sciences, University of Liverpool, UK
yalin.zheng@liverpool.ac.uk
[2] School of Computer Science, University of Exeter, UK
[3] School of Computer Science and Engineering, South China University of Technology, China
[4] Faculty of Medicine, Imperial College London, UK
[5] School of Computing and Mathematical Sciences, University of Leicester, UK
[6] Ningbo Institute of Materials Technology and Engineering, CAS, China
[7] BESE, King Abdullah University of Science and Technology, Saudi Arabia
[8] Department of Cardiovascular and Metabolic Medicine, University of Liverpool, UK

**Abstract.** Multi-modal disease diagnosis benefits from integrating complementary information across heterogeneous modalities such as medical images, clinical texts, and tabular records. However, in real-world clinical settings, multi-modal data are often incomplete, leading to missing modalities that severely degrade the performance of existing methods. Most prior approaches learn decoupled representations in a global feature space, which may overlook class-level semantic structures that directly govern the downstream task, resulting in suboptimal cross-modal consistency and discriminative capability. In this paper, we propose Class-aware Semantic Decoupling (CSD), a novel paradigm that elevates global feature disentanglement to class-level semantic decoupling, enabling explicit decoupling, alignment, and fusion of class-related shared and modality-specific representations within a unified framework. Specifically, we propose Semantic-guided Decoupling Learning (SDL) to extract both shared and modality-specific class-related representations via a class-query mechanism, and jointly perform class-level shared–specific decoupling and semantic relevance-guided cross-modal alignment. We further propose Class-aware Shared–Specific Fusion (CSSF) to adaptively integrate shared and specific class-related representations, yielding more robust and discriminative predictions. Extensive experiments on image–text and image–tabular multi-label disease diagnosis benchmarks demonstrate that the proposed method significantly outperforms state-of-the-art methods. Code will be made publicly available.

**Keywords:** Missing modality · Multi-modal disease diagnosis · Multi-label classification

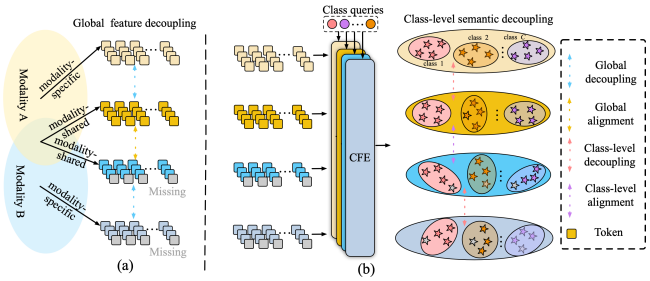


**Fig. 1.** Comparison of decoupling paradigms for multi-modal learning with missing modalities. (a) Global feature decoupling operates on globally pooled features to separate shared and modality-specific representations. (b) Our CSD projects features into a class-related semantic space via a class-related feature extractor (CFE), enabling class-level decoupling and alignment for more discriminative and robust representations under missing-modality scenarios.

## 1 Introduction

Deep learning has significantly advanced disease diagnosis by enabling automated analysis of complex medical data [1,8,28]. In particular, multi-modal learning [17,5,13] integrates complementary information from heterogeneous sources such as medical images, electronic health records (EHR), and clinical reports. By jointly modeling these modalities, multi-modal approaches provide a more comprehensive understanding of disease characteristics and consistently outperform single-modality models in various diagnostic tasks. Despite the success of multi-modal learning, most existing methods assume that all modalities are available during both training and inference. In real-world clinical practice, however, data are often incomplete due to missing examinations, irregular follow-ups, or equipment constraints [25]. When one or more modalities are absent, conventional multi-modal models may fail to effectively exploit complementary information, leading to unstable predictions and limited generalization capability. To address the missing modality problem, several strategies have been explored. Data imputation-based methods [26,15] attempt to reconstruct missing modalities from available inputs using generative models. However, the imputed data may introduce noise or fail to preserve discriminative disease cues, limiting their reliability for downstream diagnosis. Knowledge distillation-based methods [22,23] transfer knowledge from a complete-modality teacher to an incomplete-modality student, but require fully observed multi-modal data to train the teacher, which is often unavailable in practice. More recently, prompt-based methods [11,19] combine prompt learning with fine-tuning to harness the rich prior knowledge embedded in large pre-trained models. Nevertheless, they are predominantly built upon vision–language architectures and struggle to generalize to heterogeneous modalities beyond image–text pairs, such as tabular clinical data. Disentangled representation learning [27,3] has shown promise

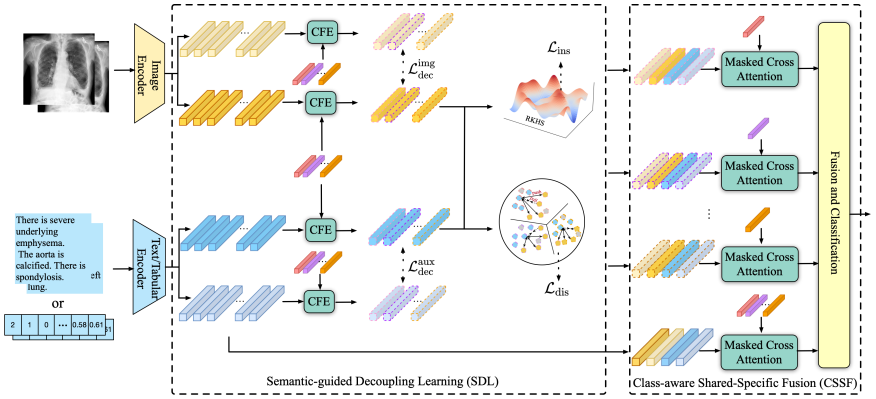


**Fig. 2.** Overview of the proposed CSD. SDL first extracts class-related shared-specific representations via a class-query mechanism. It then performs class-level shared–specific decoupling and semantic relevance-guided instance- and distribution-level alignment across modalities. Finally, CSSF adaptively fuses class-related shared and modality-specific representations for robust diagnosis.

by decomposing multi-modal features into shared and modality-specific components. Although this strategy enhances robustness under missing modalities, most methods perform decoupling in a global feature space, as shown in Fig. 1(a), emphasizing overall modality alignment while overlooking class-level semantic structures that directly influence diagnostic decisions.

In this paper, we propose a novel CSD framework that shifts multi-modal representation learning from global feature disentanglement to class-level semantic decoupling, allowing to better preserve disease-specific semantic structures under missing modalities. Instead of performing shared–specific decomposition in a holistic feature space, our approach explicitly organizes representations around disease-specific semantics, as shown in Fig. 1(b). Specifically, following [27], we first obtain shared and modality-specific token-level features, upon which SDL first employs a class-query mechanism to extract class-related shared and modality-specific representations. Then, SDL performs class-level shared–specific decoupling, while promoting instance-level semantic consistency and distribution-level alignment within shared class-related semantic space. In particular, these alignment processes are weighted by class semantic relevance obtained from class query-based attention, thereby suppressing semantically irrelevant samples and enhancing disease-relevant feature alignment. Finally, CSSF adaptively integrates shared and modality-specific class-related representations, yielding more robust and discriminative predictions under incomplete modality inputs.

The main contributions are summarized as follows: **1)** We propose a novel CSD framework for multi-modal disease diagnosis under missing modalities, which elevates global feature disentanglement to class-level semantic decoupling.

**2)** We design SDL to explicitly extract class-related shared-specific representations and perform class-level shared-specific decoupling and semantic relevance-guided cross-modal alignment. **3)** We introduce CSSF to adaptively integrate shared and modality-specific class-related representations, achieving superior robustness and performance under various missing-modality settings.

## 2 Methodology

An overview of the proposed method is illustrated in Fig. 2. Given a dataset of $N$ samples $\mathcal{D} = (x_i^{\text{img}}, x_i^{\text{aux}}, y_i, m_i^{\text{img}}, m_i^{\text{aux}})_{i=1}^{N}$, where $x_i^{\text{img}}$ denotes the image modality, $x_i^{\text{aux}}$ denotes an auxiliary modality (tabular or text), $y_i \in \{0,1\}^C$ is a multi-label disease annotation over $C$ disease categories, and $m_i^{\text{img}}, m_i^{\text{aux}} \in \{0,1\}$ are modality availability masks, our goal is to learn a diagnostic model that predicts disease probabilities $\hat{y}_i \in [0,1]^C$ using any subset of available modalities.

### 2.1 Semantic-guided Decoupling Learning

Following [27], we first extract shared and modality-specific token-level features for each modality using two encoders with partially shared parameters. For modality $k \in \{\text{img}, \text{aux}\}$ and shared (or specific) branch $b \in \{\text{sh}, \text{sp}\}$, the token-level features are $\mathbf{T}^{b,k} \in \mathbb{R}^{B \times T_k \times D}$, where $B$ is the batch size, $T_k$ the number of tokens, and $D$ the embedding dimension. Unlike prior methods that directly perform global feature decoupling, we project token features into a class-related semantic space by CFE. Specifically, we define a set of learnable class queries $\mathbf{Q} = \{\mathbf{q}_c\}_{c=1}^{C}$, where $\mathbf{q}_c \in \mathbb{R}^D$, which serves as a semantic anchor representing the $c$-th disease concept. For each modality $k$ and branch $b$, class queries interact with token-level features through cross-attention, which is formulated as:

$$\mathbf{Z}^{b,k} = \mathbf{A}^{b,k}(\mathbf{T}^{b,k} W_V), \mathbf{A}^{b,k} = \text{Softmax}\left(\frac{(\mathbf{Q}W_Q)(\mathbf{T}^{b,k}W_K)^\top}{\sqrt{D}}\right) \tag{1}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{D \times D}$ denote learnable projection matrices. $\mathbf{A}^{b,k} \in \mathbb{R}^{B \times C \times T_k}$ is the attention map. For sample $i$, $\mathbf{A}_i^{b,k}[c,t]$ indicates how strongly the query of class $c$ attends to token $t$. After obtaining the class-related representation $\mathbf{Z}^{b,k} \in \mathbb{R}^{B \times C \times D}$, we aim to perform shared-specific decoupling and cross-modal alignment in the class-related semantic space.

Although both shared and specific representations are class-discriminative, they encode complementary information. Enforcing strict orthogonality may suppress valid class-related cues. Therefore, we adopt a redundancy-reduction objective to penalize only excessive overlap. For each class $c$, we have:

$$\mathcal{L}_{\text{dec}}^{c} = \sum_{k \in \{\text{img},\text{aux}\}} \frac{1}{|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} \max\Big(0,\ \Big|\cos\Big(\mathbf{Z}_{i,c}^{\text{sh},k}, \mathbf{Z}_{i,c}^{\text{sp},k}\Big)\Big| - \delta\Big) \tag{2}$$

where $\mathcal{B}_k = \{i \mid m_i^k = 1\}$ and $|\mathcal{B}_k|$ is the number of available samples of modality $k$ in a mini-batch. $\cos(\cdot)$ denotes cosine similarity. $\delta$ is a margin that allows limited semantic overlap while penalizing excessive redundancy.

Unlike conventional methods [27] that treat all samples equally in cross-modal alignment, we weight alignment by class-related semantic strength to avoid introducing noise from weakly supported samples. To measure semantic relevance, we compute a class-level score via max pooling over the class-query attention map, which is defined as $w_{i,c}^{b,k} = \max_{t\in\{1,\dots,T_k\}} \mathbf{A}_i^{b,k}[c,t]$. Building upon the semantic relevance, we further introduce an instance-level contrastive consistency loss based on InfoNCE [14] to enforce consistency between shared representations. For each class $c$, the instance-level alignment loss is defined as:

$$\mathcal{L}_{\text{ins}}^{c} = -\frac{1}{\sum_i \alpha_{i,c}^{\text{ali}}} \sum_i \alpha_{i,c}^{\text{ali}} \log \frac{\exp\Big(\langle \hat{\mathbf{Z}}_{i,c}^{\text{sh,img}}, \hat{\mathbf{Z}}_{i,c}^{\text{sh,aux}} \rangle / \tau\Big)}{\sum_j \exp\Big(\langle \hat{\mathbf{Z}}_{i,c}^{\text{sh,img}}, \hat{\mathbf{Z}}_{j,c}^{\text{sh,aux}} \rangle / \tau\Big)} \tag{3}$$

where $\hat{\mathbf{Z}}$ denotes $\ell_2$normalized features, and $\tau$ is a temperature parameter. $\langle\cdot,\cdot\rangle$ denotes the inner product. $\alpha_{i,c}^{\text{ali}} = (w_{i,c}^{\text{sh,img}} + w_{i,c}^{\text{sh,aux}})/2$ measures the joint semantic relevance of both modalities for class $c$ in sample $i$, and is defined only when both modalities are available. This weighting strategy ensures that instance-level alignment emphasizes samples with strong cross-modal semantic support, while down-weighting ambiguous or weakly activated classes.

While the instance-level contrastive loss enforces pair-level consistency between corresponding samples, it does not explicitly guarantee alignment of the overall feature distributions within each class-related semantic space. To further reduce modality gaps at the distribution level, we thus introduce a weighted Maximum Mean Discrepancy (MMD) to align class-level shared representations. For each class $c$, the distribution-level alignment loss is defined as:

$$\begin{aligned} \mathcal{L}_{\text{dis}}^{c} = \frac{1}{(\sum_i \alpha_{i,c}^{\text{ali}})^2} \sum_{i,j} \alpha_{i,c}^{\text{ali}} \alpha_{j,c}^{\text{ali}} \Big[ & k(\mathbf{Z}_{i,c}^{\text{sh,img}}, \mathbf{Z}_{j,c}^{\text{sh,img}}) \\ & + k(\mathbf{Z}_{i,c}^{\text{sh,aux}}, \mathbf{Z}_{j,c}^{\text{sh,aux}}) - 2k(\mathbf{Z}_{i,c}^{\text{sh,img}}, \mathbf{Z}_{j,c}^{\text{sh,aux}}) \Big] \end{aligned} \tag{4}$$

where $k(\cdot,\cdot)$ is a Gaussian RBF kernel [6]. This objective complements $\mathcal{L}_{\text{ins}}^{c}$ by aligning global class-related distributions rather than individual instances. Finally, SDL combines class-level decoupling and alignment losses:

$$\mathcal{L}_{\text{sdl}} = \sum_{c=1}^{C} \Big( \mathcal{L}_{\text{dec}}^{c} + \mathcal{L}_{\text{ins}}^{c} + \mathcal{L}_{\text{dis}}^{c} \Big) \tag{5}$$

### 2.2 Class-aware Shared-Specific Fusion

After obtaining the decoupled class-related representations $\mathbf{Z}^{b,k}$, we further design the CSSF module to integrate shared and modality-specific evidence for each disease category, enabling each class query to perform fusion within its own decoupled semantic space rather than a shared global space, resulting in more targeted and robust fusion. Since the shared branch aims to encode modality-invariant semantics, we first aggregate the shared representations across modalities by $\mathbf{Z}^{\text{sh}} = \big(\mathbf{Z}^{\text{sh,img}} + \mathbf{Z}^{\text{sh,aux}}\big)/2$. For missing modalities, $\mathbf{Z}^{\text{sh}}$ is set to the

shared representation of the available modality. Then, we concatenate it with the modality-specific representations to obtain $\mathbf{Z} \in \mathbb{R}^{B\times 3\times C\times D}$. To adaptively integrate shared and specific evidence for each class, we perform class-wise fusion conditioned on the class query. Specifically, for each class $c$, we use its corresponding query $\mathbf{q}_c$ to attend over the three semantic components $\mathbf{Z}_c$:

$$\overline{\mathbf{F}}_c = \text{Softmax}\left(\frac{(\mathbf{q}_c W_f)(\mathbf{Z}_c W_g)^\top}{\sqrt{D}} + \mathbf{m}\right)\mathbf{Z}_c W_v \tag{6}$$

where $\overline{\mathbf{F}}_c$ is the fused $c$-related representation in $\overline{\mathbf{F}} \in \mathbb{R}^{B\times C\times D}$. $W_f, W_g, W_v \in \mathbb{R}^{D\times D}$ are learnable parameters, and $\mathbf{m} \in \mathbb{R}^{B\times 1\times 3}$ denotes the modality mask. When modality $k$ is missing, the corresponding modality-specific mask is set to $-\infty$; otherwise, it is set to 0.

To incorporate global contextual information, we further compute globally pooled token features and apply the same masked class-query fusion mechanism to obtain global class-aware embeddings $\hat{\mathbf{F}} \in \mathbb{R}^{B\times C\times D}$. Finally, we obtain the fused class-related representations, i.e., $\mathbf{F} = \overline{\mathbf{F}} + \hat{\mathbf{F}}$, from both class semantic space and global feature space. This complementary integration enables the model to capture both discriminative class-related evidence and global contextual information, improving robustness under missing-modality scenarios. Subsequently, $\mathbf{F}$ is used for classification using binary cross-entropy loss $\mathcal{L}_{\text{cls}}$. The overall loss is defined as $\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda\mathcal{L}_{\text{sdl}}$, where $\lambda$ is a trade-off hyperparameter.

## 3 Experiments

### 3.1 Experimental Setup

**Datasets and Evaluation Metrics.** IU-Xray [4] consists of 7,470 chest X-ray images and 3,955 radiology reports. Each report is associated with a pair of images corresponding to frontal and lateral views. Following [18], we employ a pretrained BERT-based model, CheXbert [20], to annotate each report with one or more of the 14 chest disease categories. In our experiments, we only use the frontal-view images. After preprocessing, 2,327 samples are used for training and 582 samples for testing. MIMIC [9,10] comprises time-series tabular data (EHR) and paired chest radiographs for multi-label disease prediction across 25 diagnostic categories. Following the preprocessing in [27], the dataset is split into 7,637 training, 857 validation, and 2,136 testing samples. Model performance is evaluated using PRAUC (Macro).

**Implementation Details.** All experiments are implemented in PyTorch with two NVIDIA A100 GPUs. We adopt the Adam optimizer with an initial learning rate of 1e-4 and set the batch size to 16. Following [27], ResNet-50 [7] and a transformer-based network are employed as the image and tabular encoders, respectively. For text modeling, we use the pretrained BioClinical BERT [2] as the text encoder. $\delta$ and $\lambda$ are set to 0.02 and 0.1, respectively. A missing rate of $\eta\%$ indicates that $\frac{\eta}{2}\%$ of the samples contain only image modality and $\frac{\eta}{2}\%$ contain only text (or tabular) modality, while the remaining $(100-\eta)\%$ samples have complete multi-modal inputs.

**Table 1.** Performance comparison on IU-Xray and MIMIC under different missing rates applied to both training and testing phases. The best results are shown in **bold** and the second best are underlined. Higher PRAUC (%) indicates better performance.

| | IU-Xray (Image-Text) | | | | MIMIC (Image-Tabular) | | | |
|---|---|---|---|---|---|---|---|---|
| **Missing Rate** | **30%** | **50%** | **70%** | **90%** | **30%** | **50%** | **70%** | **90%** |
| **Method** | PRAUC | PRAUC | PRAUC | PRAUC | PRAUC | PRAUC | PRAUC | PRAUC |
| ShaSpec [21] | 78.01 | 73.12 | 68.54 | 55.31 | 40.14 | 37.13 | 35.64 | 34.21 |
| DrFuse [27] | 79.12 | 75.30 | 71.35 | 56.67 | 40.34 | 38.64 | 36.20 | 34.56 |
| DMRNet [24] | 76.11 | 71.68 | 68.78 | 53.87 | 39.91 | 36.90 | 34.56 | 32.26 |
| Mora [19] | 79.67 | 74.78 | 72.03 | 57.45 | - | - | - | - |
| SimMLM [12] | 77.32 | 72.56 | 68.98 | 54.61 | 40.01 | 36.97 | 35.70 | 34.23 |
| IM-Fuse [16] | 75.32 | 71.24 | 67.05 | 53.10 | 39.39 | 36.98 | 34.23 | 31.97 |
| Ours | **82.27** | **78.83** | **75.78** | **61.08** | **42.06** | **40.62** | **38.70** | **37.84** |

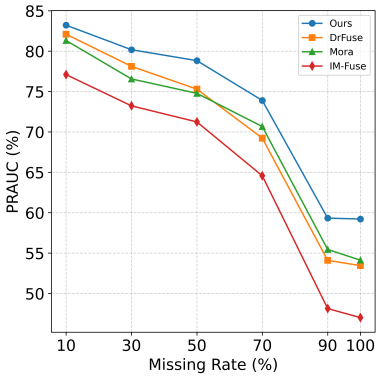


**Fig. 3.** Robustness analysis on IU-Xray.

**Table 2.** Ablation study on different components under 70% missing setting.

| Module Setting | IU-Xray | MIMIC |
|---|---|---|
| Baseline | 70.43 | 35.66 |
| +SDL | 74.12 | 37.96 |
| +CSSF | **75.78** | **38.70** |
| w/o $\mathcal{L}_{\text{ins}}$ | 73.38 | 37.23 |
| w/o $\mathcal{L}_{\text{dis}}$ | 74.45 | 37.98 |
| w/o SR | 73.67 | 36.90 |
| w/o $\overline{\mathbf{F}}$ | 73.01 | 37.20 |
| w/o $\hat{\mathbf{F}}$ | 74.40 | 38.02 |

### 3.2 Experimental Results

**Comparison with State-of-the-art Methods.** Table 1 reports the performance comparison with state-of-the-art incomplete multi-modal learning methods on IU-Xray and MIMIC under different missing rates. On IU-Xray, our method consistently achieves the best performance across all missing ratios. Under the moderate 50% missing setting, our approach attains 78.83% PRAUC, outperforming the strongest competitor, DrFuse by 3.53%. As the missing rate increases to 70% and 90%, the performance gap becomes more pronounced. Specifically, at 90% missing, our method achieves 61.08%, surpassing Mora by 3.63%. This demonstrates that our class-aware modeling is particularly robust under severe modality incompleteness. On MIMIC, which involves heterogeneous image–tabular modalities, our method also achieves competitive results across all missing rates. At 30% missing, we achieve 42.06%, improving over DrFuse by 1.72%. Under the challenging 90% missing setting, our method obtains 37.84%, exceeding the second-best performance by 3.28%. The stable im-

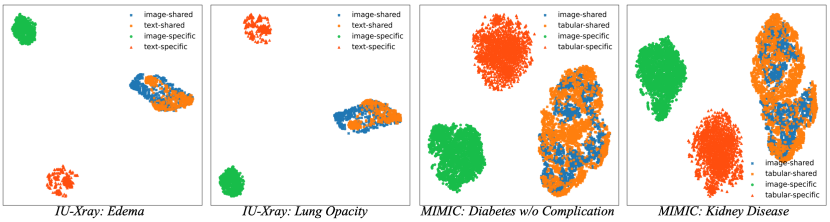


**Fig. 4.** Class-wise t-SNE visualization of class-related shared and modality-specific representations.

provement across increasing missing ratios further validates the effectiveness of performing shared–specific decoupling and cross-modal alignment in the class-related semantic space. We further evaluate the robustness of the performance gains on the IU-Xray dataset by computing 95% confidence intervals. Mora achieved a PRAUC of 57.45% (56.2–59.1), whereas our method achieved 61.08% (59.8–62.3), demonstrating a clear and consistent improvement.

To evaluate robustness under distribution shift of modality incompleteness, we fix the training missing rate at 50% and vary the testing missing ratio from 10% to 100%. As shown in Fig. 3, although all methods degrade as the missing rate increases, our method consistently achieves the best performance across all settings. The performance gap widens under high missing ratios, demonstrating stronger robustness and generalization to severe modality incompleteness.

**Visualization results.** To further analyze the effectiveness of the proposed class-aware semantic decoupling, we conduct class-wise t-SNE visualization of the learned class-related shared and modality-specific representations, as shown in Fig. 4. We observe that the shared representations from different modalities are well aligned within each class, forming compact and overlapping clusters. This indicates that our semantic relevance-guided alignment successfully reduces cross-modal discrepancy in the class-related semantic space. In contrast, modality-specific representations are clearly separated from the shared components and also remain distinguishable across modalities, demonstrating effective shared–specific disentanglement.

**Ablation Study.** As shown in Table 2, introducing SDL significantly improves performance over the baseline, demonstrating the effectiveness of performing class-level shared–specific decoupling and cross-modal alignment. Further incorporating CSSF yields the best results, indicating that adaptively integrating shared and modality-specific class-related semantics enhances discriminative capability under high missing rates. We further analyze the optimization objectives in SDL. Removing the instance-level alignment loss $\mathcal{L}_{\text{ins}}$ or the distribution-level loss $\mathcal{L}_{\text{dis}}$ consistently degrades performance, verifying that they are complementary for stabilizing the class-related shared space. Discarding semantic relevance (SR) weighting also leads to notable drops. This is because semantically weak or less relevant features are equally involved in the alignment process, introducing

noise and weakening class-discriminative learning. Finally, removing either of the two types of class-related representations results in a performance decline, suggesting that they capture complementary class-related information.

## 4 Conclusions

In this paper, we introduced Class-aware Semantic Decoupling (CSD), a novel framework for multi-modal disease diagnosis under missing-modality settings. By explicitly modeling class-level semantic structures through Semantic-guided Decoupling Learning (SDL) and Class-aware Shared–Specific Fusion (CSSF), CSD effectively enhances robustness to incomplete data while preserving discriminative class semantics. Extensive experiments across heterogeneous image–text and image–tabular multi-label diagnosis benchmarks demonstrate consistent improvements over state-of-the-art methods. Notably, the performance gains are achieved with only a modest increase in computational cost, since the proposed method mainly introduces lightweight class-query interactions. Future work will explore extending this framework to broader multi-modal medical applications and further investigate per-class performance.

**Acknowledgments.** This study was funded by TARGET, a project funded by the EU HORIZON EUROPE framework programme for research and innovation under the Grant Agreement No. 101136244.

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

## References

1. Aggarwal, R., Sounderajah, V., Martin, G., Ting, D.S., Karthikesalingam, A., King, D., Ashrafian, H., Darzi, A.: Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ digital medicine **4**(1), 65 (2021)
2. Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jindi, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. In: Proceedings of the 2nd clinical natural language processing workshop. pp. 72–78 (2019)
3. Chen, Y., Pan, Y., Xia, Y., Yuan, Y.: Disentangle first, then distill: a unified framework for missing modality imputation and alzheimer's disease diagnosis. IEEE Transactions on Medical Imaging **42**(12), 3566–3578 (2023)
4. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
5. Feng, J., Xu, H., Cai, J., Chang, Y., Zhang, D., Du, S., Wang, J.: Cross-modal brain graph transformer via function-structure connectivity network for brain disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 247–256. Springer (2025)

6. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. The journal of machine learning research **13**(1), 723–773 (2012)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
8. Holste, G., Zhou, Y., Wang, S., Jaiswal, A., Lin, M., Zhuge, S., Yang, Y., Kim, D., Nguyen-Mau, T.H., Tran, M.T., et al.: Towards long-tailed, multi-label disease classification from chest x-ray: Overview of the cxr-lt challenge. Medical Image Analysis **97**, 103224 (2024)
9. Johnson, A.E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., et al.: Mimic-iv, a freely accessible electronic health record dataset. Scientific data **10**(1), 1 (2023)
10. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
11. Lee, Y.L., Tsai, Y.H., Chiu, W.C., Lee, C.Y.: Multimodal prompting with missing modalities for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14943–14952 (2023)
12. Li, S., Chen, C., Han, J.: Simmlm: A simple framework for multi-modal learning with missing modality. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24068–24077 (2025)
13. Monajatipoor, M., Rouhsedaghat, M., Li, L.H., Jay Kuo, C.C., Chien, A., Chang, K.W.: Berthop: An effective vision-and-language model for chest x-ray disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 725–734. Springer (2022)
14. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
15. Pan, Y., Liu, M., Xia, Y., Shen, D.: Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multi-modality data. IEEE transactions on pattern analysis and machine intelligence **44**(10), 6839–6853 (2021)
16. Pipoli, V., Saporita, A., Marchesini, K., Grana, C., Ficarra, E., Bolelli, F., et al.: Im-fuse: A mamba-based fusion block for brain tumor segmentation with incomplete modalities. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2025 (2025)
17. Pölsterl, S., Wolf, T.N., Wachinger, C.: Combining 3d image and tabular data via the dynamic affine feature map transform. In: International conference on medical image computing and computer-assisted intervention. pp. 688–698. Springer (2021)
18. Sadman, N., Zulkernine, F., Kwan, B.: Interpreting biomedical vlms on high-imbalance out-of-distributions: An insight into biomedclip on radiology. arXiv preprint arXiv:2506.14136 (2025)
19. Shi, Z., Kim, J., Li, W., Li, Y., Pfister, H.: Mora: Lora guided multi-modal disease diagnosis with missing modality. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 273–282. Springer (2024)
20. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A., Lungren, M.: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1500–1519. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.117, https://aclanthology.org/2020.emnlp-main.117/

21. Wang, H., Chen, Y., Ma, C., Avery, J., Hull, L., Carneiro, G.: Multi-modal learning with missing modality via shared-specific feature modelling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15878–15887 (2023)
22. Wang, H., Ma, C., Zhang, J., Zhang, Y., Avery, J., Hull, L., Carneiro, G.: Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 216–226. Springer (2023)
23. Wang, X., Wang, Y., Liang, S., Tang, F., Liu, C., Hu, M., Hu, C., He, J., Ge, Z., Razzak, I.: Robust multimodal learning for ophthalmic disease grading via disentangled representation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 447–456. Springer (2025)
24. Wei, S., Luo, Y., Wang, Y., Luo, C.: Robust multimodal learning via representation decoupling. In: European Conference on Computer Vision. pp. 38–54. Springer (2024)
25. Wu, R., Wang, H., Chen, H.T., Carneiro, G.: Deep multimodal learning with missing modality: A survey. arXiv preprint arXiv:2409.07825 (2024)
26. Wu, Z., Dadu, A., Tustison, N., Avants, B., Nalls, M., Sun, J., Faghri, F.: Multimodal patient representation learning with missing modalities and labels. In: The Twelfth International Conference on Learning Representations (2024)
27. Yao, W., Yin, K., Cheung, W.K., Liu, J., Qin, J.: Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 16416–16424 (2024)
28. Zhou, F., Gao, Z., Zhao, H., Xie, J., Meng, Y., Zhao, Y., Lip, G.Y., Zheng, Y.: Glcp: Global-to-local connectivity preservation for tubular structure segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 237–247. Springer (2025)