Multi-modal Medical Diagnosis via Large-small Model Collaboration

Wanyi Chen[1] Zihua Zhao[2] Jiangchao Yao[2,4]
Ya Zhang[3,4] Jiajun Bu[1] Haishuai Wang[1,✉]

[1]Zhejiang Key Laboratory of Accessible Perception and Intelligent Systems, College of Computer Science, Zhejiang University
[2]Cooperative Medianet Innovation Center, Shanghai Jiao Tong University
[3]School of Artificial Intelligence, Shanghai Jiao Tong University
[4]Shanghai Artificial Intelligence Laboratory

chenwanyi@zju.edu.cn, sjtuszzh@sjtu.edu.cn, Sunarker@sjtu.edu.cn
ya_zhang@sjtu.edu.cn, bjj@zju.edu.cn, haishuai.wang@zju.edu.cn

Abstract

Recent advances in medical AI have shown a clear trend towards large models in healthcare. However, developing large models for multi-modal medical diagnosis remains challenging due to a lack of sufficient modal-complete medical data. Most existing multi-modal diagnostic models are relatively small and struggle with limited feature extraction capabilities. To bridge this gap, we propose AdaCoMed*, an adaptive collaborative-learning framework that synergistically integrates the off-the-shelf medical single-modal large models with multi-modal small models. Our framework first employs a mixture-of-modality-experts (MoME) architecture to combine features extracted from multiple single-modal medical large models, and then introduces a novel adaptive co-learning mechanism to collaborate with a multi-modal small model. This co-learning mechanism, guided by an adaptive weighting strategy, dynamically balances the complementary strengths between the MoME-fused large model features and the cross-modal reasoning capabilities of the small model. Extensive experiments on two representative multi-modal medical datasets (MIMIC-IV-MM and MMIST ccRCC) across six modalities and four diagnostic tasks demonstrate consistent improvements over state-of-the-art baselines, making it a promising solution for real-world medical diagnosis applications. The code is available at* https://github.com/Zoew420/AdaCoMed.

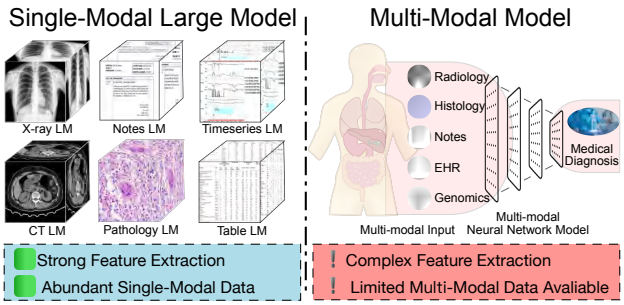

Figure 1. **Comparison of Single-Modal Large Models and Multi-Modal Models in Medical AI.** Left: Single-modal large models specialized in different medical data types, featuring strong feature extraction capabilities and abundant training data. Right: Multi-modal neural network model integrating multiple medical modalities for comprehensive medical diagnosis, but challenged by complex feature extraction and limited multi-modal data availability. The contrasting backgrounds (blue vs. red) highlight the fundamental trade-offs between these approaches.

1. Introduction

Recent years have witnessed significant advancements in artificial intelligence across healthcare, marked by the emergence of specialized large-scale models in different modalities. For instance, in natural language processing, MedPaLM [40, 41] has achieved remarkable performance in medical text analysis and knowledge retrieval. In computer vision, large models like MedSAM [34] have revolutionized medical imaging tasks, particularly excelling in anatomical structure segmentation. Other specialized models have advanced in processing electronic health records (EHR) and analyzing temporal medical data, establishing new benchmarks in their respective areas of AI-assisted healthcare.

Current large models in healthcare AI can be generally categorized based on their capability to process modalities.

The first two authors contributed equally.
✉Corresponding author: haishuai.wang@zju.edu.cn

Most of them are either single-modal, as shown in the left panel of Fig. 1, designed to handle one type of data (e.g., text-only [22, 26, 40, 51] or image-only [29, 34, 44, 49, 63] models), or bi-modal, yet limited to processing the pair-wise data, most commonly image-text pairs [10, 24, 28, 45]. When it comes to building large models that can simultaneously cope with three or more distinct types of medical data, including but not limited to medical images, clinical notes, temporal signals, and structured tabular data, it is still very challenging due to a lack of sufficient multi-modal paired data for training. As depicted in the right panel of Fig. 1, models that can handle such varied multi-modal analysis [8, 42, 47, 61, 62] are frequently "small models", namely models at a small scale. These models, however, have limited feature extraction capacity and usually underperform in clinical scenarios demanding complex transformation with various modalities. Consequently, there is a pressing need for a scalable multi-modal model that can balance the breadth of multi-modal integration with the depth of analysis required for advanced clinical applications.

Given the above limitation, a straightforward solution to multi-modal medical analysis would be directly combining multiple single-modal large models [16, 36]. However, our studies reveal that such naive fusion leads to *modality domination*, where pretrained model capabilities rather than clinical significance, determine the modality contribution (see Fig. 2). This might risk AI systems to deviate from established medical practices. Conversely, small task-specific models could potentially achieve more balanced feature integration and better align the inherent diagnostic value of each modality, although they struggle with limited feature extraction. How to leverage the merits of both large and small models seems to be a better choice for multi-modal medical diagnosis.

To achieve this goal, we present AdaCoMed, an adaptive co-learning framework that enables synergistic collaboration between single-modal medical large models and multi-modal small models. Our framework comprises two parallel pathways designed to process multi-modal medical data. The first pathway utilizes specialized large models to extract modality-specific embeddings and employs a Mixture-of-Modality-Experts Fusion module (MoME) for cross-modal fusion. The second pathway leverages a more efficient pretrained, task-driven small multi-modal model. After obtaining embeddings from both pathways, we employ contrastive learning to align the large models' collaborative fusion representations with the natural multi-modal patterns captured by small models. Specifically, the contrastive loss encourages the large models to learn the inherent cross-modal relationships that small models naturally encode. For the final prediction, we propose an adaptive weighting scheme to integrate logits from both pathways. We design four different adaptive weighting strategies to accommodate various tasks and data scales, allowing flexible selection based on specific application scenarios. Our key contributions are threefold:

- We propose a novel collaborative learning framework that synergistically combines single-modal large models with the small multi-modal models for multi-modal medical diagnosis. This framework effectively incorporates the merits of large and small models while circumventing the need for extensive paired multi-modal training data.
- We develop AdaCoMed with two critical components: a contrastive learning alignment between large and small models that is enhanced by shared co-training head, and an adaptive weighting mechanism with several instantiated schemes that automatically estimate the dynamics of both pathways to promote better collaborative prediction.
- We conduct extensive experiments across various multi-modal medical diagnostic tasks, proving that AdaCoMed consistently outperforms both standalone large models, their naive combination, and traditional multi-modal approaches, while maintaining the computational efficiency.

2. Related Works

Large Models in Healthcare. Large models has significantly impacted healthcare applications across three main streams: medical imaging foundation models, medical language models, and models for other healthcare data types. In medical imaging analysis, the notable advances including Med-SAM [34] for medical image segmentation, Rad-FM [49] for radiology analysis, and Prov-GigaPath [53] for pathology image understanding *etc.*, are achieved with impressive performance improvement. The medical language model domain evolved from encoder-based models like BioBert [23] to larger decoder-only architectures such as BioMistral [22] and Me-Llama [51]. To address the limitations in producing high-quality text embeddings, approaches like Llm2vec [2] have been further developed to transform decoder-only LLMs into encoder architectures. For specialized medical data types, models like LLM-Forest [15] for tabular data imputation, MedTsLLM [5] for time series analysis, and ECG-FM [35] for physiological signals emerged. Nevertheless, most existing models focus on single modalities, lacking the capability to effectively integrate information across different medical data types.

Multi-modal Medical Diagnosis. Traditional approaches for multi-modal medical diagnosis primarily focus on different fusion levels (data-level [54], feature-level [60], and decision-level [1]). At feature level, transformer-based [4, 43, 55, 59, 62] and graph-based [11–13, 61] architectures have gained significant attention due to their strong ability to model cross-modal interactions. For example, Meta-Transformer [59] proposed a general transformer encoder design for task finetuning, and HetMed [20] modeled relationships between different medical entities through graph structures. The landscape has then evolved with the

emergence of large-scale vision-language foundation models. Models like Med-VLP [9], Med-flamingo [37], Llava-med [24], pre-trained on vast amounts of paired medical images and clinical text data, demonstrate superior capability in learning generalizable cross-modal representations, though they are typically limited to bi-modal scenarios. Recently, general-purpose multi-modal foundation models like NExT-GPT [50], AnyGPT [56], and OneLLM [14] have emerged, capable of handling diverse modalities. However, advances in general multi-modal modeling have not yet been extensively adapted to medical domain, due to a lack of sufficient modal-complete medical datasets.

3. Method

3.1. Preliminary

Given a multimodal dataset $D = (x_1^i, x_2^i, ..., x_M^i, y^i)_{i=1}^N$, where x_m^i represents the input from the m-th modality ($m \in [1, M]$) for the i-th sample, y^i denotes the corresponding label, and N is the total number of samples in the dataset. Our task is to effectively integrate and utilize information from multiple modalities for robust prediction. Traditional single-modal large models address prediction tasks by independently projecting each modal input into high-dimensional representations e_{large}^m through pre-trained encoder, and then fusing representation through a Mixture-of-Modality-Experts Fusion module as e_{large}, followed by task-specific prediction heads f. While these models possess rich knowledge and powerful representation capabilities, they potentially cannot model sophisticated cross-modal patterns and relationships. In comparison, small multimodal models employ shallow encoders to extract features into a shared representation space e_{small}, subsequently fusing these representations before making predictions through a shared task head f. Although these models inherently capture cross-modal interactions, they suffer from limited model capacity and are constrained by the available training data. Our objective is to develop a unified framework that combines the extensive knowledge of large models with the natural cross-modal feature interaction capabilities of small multi-modal models.

3.2. Motivation

To leverage the powerful large models, a straightforward approach is directly concatenating single-modal large model representations. However, the empirical analysis in Fig. 2 (left) reveals the undesired problem about this naive fusion strategy. Specifically, for the direct combination, the fusion features (blue) exhibit heavy overlap with text embeddings (green). It counter-intuitively places the emphasis on medical data modalities based on encoder capabilities rather than clinical importance, since text dominance risks missing crucial visual patterns and physiological signals essential for accurate medical diagnosis. We term this phenomenon as *modality domination*. Given this potential problem, we target to explore a new way to leverage large models and the conventional small models for multi-modal medical diagnosis. A preliminary verification of our AdaCoMed has been shown in Fig. 2 (right), which learns better representation.

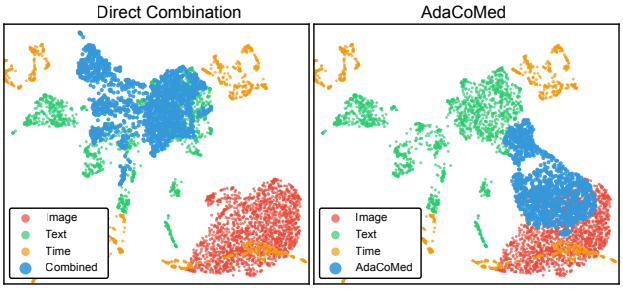

Figure 2. **UMAP visualization comparing feature distributions between direct combination (left) and AdaCoMed (right).** The direct combination shows problematic overlap between combined features (blue) and text features (green), indicating text modality domination. In contrast, AdaCoMed achieves better integration, as evidenced by bridging distribution of fused features across image (red), text (green), and temporal (yellow) modality clusters.

3.3. Large-small Model Collaboration

In this study, we propose a novel collaborative framework of large and small models (Fig. 3), AdaCoMed, which considers two major challenges in the multi-modal fusion: The first challenge is the significant representational gap between single-modal large models and multi-modal small models, arising from their differences in training data dependencies, feature dimensionality, and model capacities. The second challenge stems from the dynamic balance of large and small model contributions needed by the co-training task head as training progresses. To tackle these challenges, our framework comprises four synergistic components: (1) a Mixture-of-Modality-Experts inspired fusion module for large model representations, (2) a contrastive learning representation alignment between large and small models enhanced by the co-training head, (3) an adaptive weighting strategy to dynamically balance the contributions from both pathways, and (4) a multi-objective optimization that coordinates different learning objectives. In the following, we give a detailed description of each module.

Mixture-of-Modality-Experts Fusion. As aforementioned in Preliminary, we first obtain modality embeddings $\{e_{\text{large}}^i\}_{i=1}^M$ from their respective large models, where M is the modality number. Then, they are projected and fused through a Mixture-of-Modality-Experts Fusion module:

$$\mathbf{h} = \text{SelfAttn}([\text{Proj}_1(e_{\text{large}}^1); ...; \text{Proj}_M(e_{\text{large}}^M)]) \quad (1)$$

$$e_{\text{large}} = \sum_{i=1}^{C} \text{Router}_i(\mathbf{h})\text{Expert}_i(\mathbf{h}) \quad (2)$$

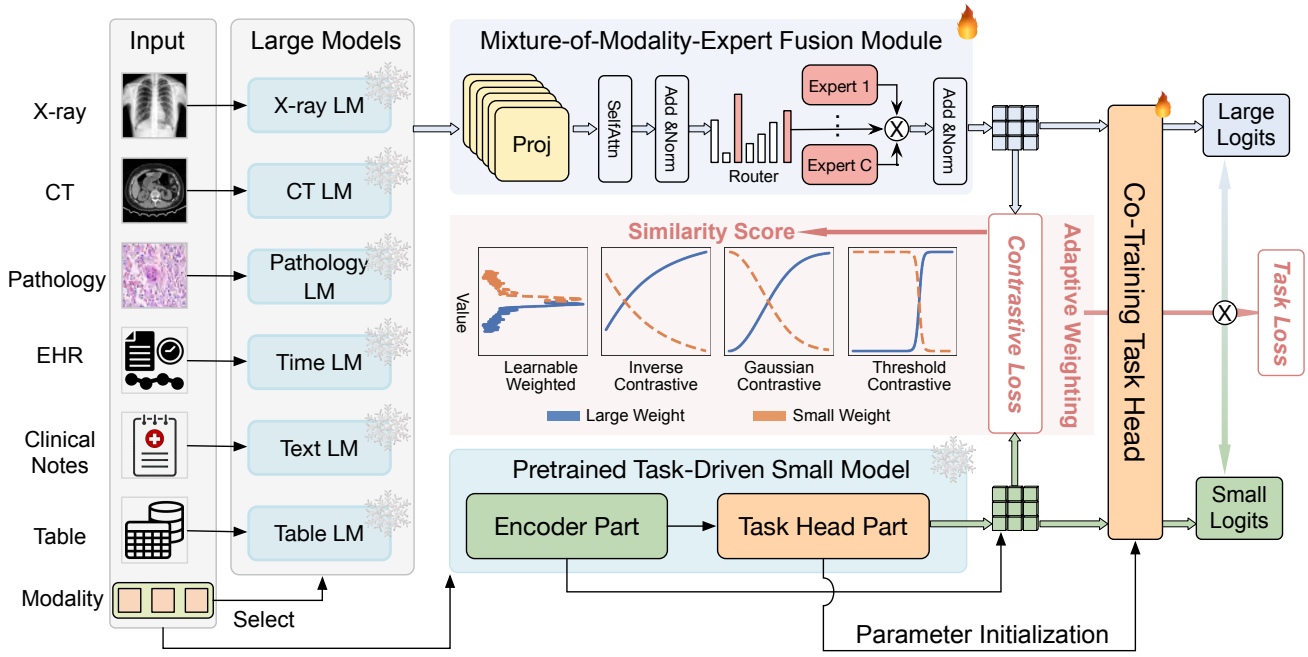

Figure 3. **AdaCoMed Workflow.** Our model simultaneously leverages features from large models (LMs) through a Mixture-of-Modality-Expert (MoME) fusion module and features from a pretrained multi-modal small model. The framework incorporates both contrastive learning and task-specific objectives. The task prediction combines logits from both pathways using an adaptive weighting strategy (shown in the red region), which implements four different strategies: Learnable Weighted, Inverse Contrastive, Gaussian Contrastive, and Threshold Contrastive, to dynamically balance the contributions from large and small models.

where Proj_i is the projection head implemented as MLP, [;] denotes concatenation operation, and $\text{C} = \sum_{k=1}^{M} \binom{M}{k}$ represents the total number of possible modality combinations, with each Expert_i specializing in processing a unique combination of input modalities to capture their interactions. For example, with M=3 modalities, C=7 experts, we would handle single modality, pair-wise, and three-way feature combinations respectively.

Contrastive Learning and Co-training Alignment. We propose a co-learning framework that aligns large and small models through two complementary mechanisms: contrastive learning and co-training. Our key insight is to leverage contrastive learning to guide large model representations towards the more balanced and naturally aligned feature of small models, while using a shared head co-training to ensure consistent task-specific predictions. This dual-alignment strategy helps maintain the rich semantic information from large models while benefiting from the well-structured representations of small models. Given the large model representation e_{large} and small model representation e_{small}, we first project them into a shared space with dimension d through learnable transformations:

$$z_{\text{large}} = p(e_{\text{large}}), \quad z_{\text{small}} = p(e_{\text{small}}) \tag{3}$$

where p is the projection layer implemented as MLP. We define the similarity score between large and small model representations as:

$$s(z_{\text{large}}, z_{\text{small}}) = \frac{\exp(sim(z_{\text{large}}, z_{\text{small}})/\tau)}{\sum_k \exp(sim(z_{\text{large}}, z^k_{\text{small}})/\tau)} \tag{4}$$

where $sim(\cdot,\cdot)$ denotes cosine similarity, τ is a temperature parameter, and z^k_{small} includes both the current sample and other samples in the mini-batch as negative examples. Then the contrastive loss can be written as:

$$\mathcal{L}_{\text{contrast}} = -\log(s(z_{\text{large}}, z_{\text{small}})) \tag{5}$$

By minimizing this contrastive loss, we encourage the large model to learn the natural cross-modal relationships captured by the small model while preserving its rich modality-specific features. While contrastive learning helps align the feature spaces, we further introduce parameter sharing in task heads to emphasize task-relevant representation knowledge transfer. Specifically, we apply the same task head to both large and small model representations:

$$\hat{y}_{\text{large}} = f(z_{\text{large}}), \quad \hat{y}_{\text{small}} = f(z_{\text{small}}) \tag{6}$$

where f is initialized from the pre-trained small model's task head. This parameter sharing strategy enables large models to benefit from the small model's well-trained prediction head during early training, leading to faster convergence while maintaining the rich semantics of large models.

Strategy	Formulation	Parameters	Key Characteristics
Learnable Weighted	$w_{\text{large}}, w_{\text{small}} = \text{softmax}(\theta)$	$\theta = (\theta_1, \theta_2)$: learnable parameters	Simply automatic optimization;
Inverse Contrastive	$w_{\text{large}}, w_{\text{small}} = \text{Norm}\left[\frac{1}{1+e^{-\mathbf{s}}}, \frac{1}{\alpha \mathbf{s}/T+1e-6}\right]$	T: temperature; α: scaling factor	Fine-grained control;
Gaussian Contrastive	$w_{\text{large}}, w_{\text{small}} = \left[1 - e^{-\frac{\mathbf{s}^2}{2\sigma^2}}, e^{-\frac{\mathbf{s}^2}{2\sigma^2}}\right]$	σ: standard deviation	Gradual transition;
Threshold Contrastive	$w_{\text{large}} = 0.5 + 0.5\tanh(k(\mathbf{s}-t))$ $w_{\text{small}} = 0.5 - 0.5\tanh(k(\mathbf{s}-t))$	t: threshold value k: steepness factor	Sharp phase change;

Table 1. **Summary of adaptive weighting strategies for large-small model collaboration.** Each strategy offers distinct advantages for different scenarios, where $\mathbf{s}$ represents the similarity score between large and small model embeddings.

Adaptive Weighting Strategy. Another key insight of our framework is that the relative importance of large and small models should vary throughout training. Initially, the small model's pre-trained task head provides reliable guidance for multimodal fusion, while the large model's representations become increasingly valuable as training progresses. To capture this dynamic, we propose adaptive weighting strategies based on similarity score $\mathbf{s} = s(z_{\text{large}}, z_{\text{small}})$ between large and small model embeddings. As shown in Tab. 1, we design four complementary strategies to accommodate different training scenarios: the Learnable Weighted offers simple optimization through backpropagation but requires careful parameter tuning; the Inverse Contrastive enables fine-grained control over the fusion process; the Gaussian Contrastive provides a smooth, continuous decay pattern through standard deviation parameter, suitable for scenarios requiring gradual transition between models; and the Threshold Contrastive implements a clear cutoff point with adjustable steepness, ideal when distinct phase transitions in model importance are desired. The choice among these strategies depends on specific requirements such as training stability, computational efficiency, and the desired transition pattern between model representations.

Multi-objective Optimization. The overall training objective combines two components: task loss and contrastive loss. Since all our tasks are classification problems, we employ cross-entropy loss for prediction:

$$\mathcal{L}_{task} = -y \log(w_{\text{large}}\hat{y}_{\text{large}} + w_{\text{small}}\hat{y}_{\text{small}}) \tag{7}$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{task} + \lambda \mathcal{L}_{\text{contrast}} \tag{8}$$

where y is the ground truth label, $\hat{y}_{\text{large}}$ and $\hat{y}_{\text{small}}$ are the predictions from the co-training task heads output. The parameter λ controls the weight of the contrastive loss. The training process of AdaCoMed is summarized in Algorithm 1.

Algorithm 1: AdaCoMed Training

Input: Dataset D, single-modal large models $\{E^i_{\text{large}}\}_{i=1}^{M}$, multi-modal small model E_{small}, epoch number T
Initialize task head f from pretrained small model E_{small}
for *each epoch* $e = 1, 2, \ldots, T$ **do**
  Extract large and small representations $e^i_{\text{large}}, e_{\text{small}}$
  Compute large models collaboration e_{large} by Eq. (2)
  Project embeddings $z_{\text{large}}, z_{\text{small}}$ by Eq. (3)
  Calculate similarity score $\mathbf{s}$ by Eq. (4)
  Calculate contrastive loss $\mathcal{L}_{\text{contrast}}$ by Eq. (5)
  Estimate $w_{\text{large}}, w_{\text{small}}$ by chosen strategy in Tab. 1
  Calculate predicted value $\hat{y}_{\text{large}}, \hat{y}_{\text{small}}$ by Eq. (6)
  Calculate fusion loss $\mathcal{L}_{\text{task}}$ by Eq. (7)
  Optimize the combination of two losses by Eq. (8)
Output: Adaptive collaborative-learning network

3.4. Discussion

In this study, our method leverages the complementary strengths of single-modal large models and multi-modal small models, achieving efficient model collaboration. Compared to single-modal large models or small multi-modal models, our approach effectively balances performance and data consumption. In contrast to fine-tuned large-scale multi-modal models, we show significant advantages in terms of computational efficiency and resource requirements. Additionally, our method extends beyond traditional model collaboration approaches, such as ensemble learning [17, 31], co-training [6, 27], and knowledge distillation [25, 30], by incorporating concepts from multi-agent frameworks [52, 57, 58] to strategically utilize large models. Applied in medical field, this innovative model collaboration achieves promising results in multi-modal medical tasks, contributing to advancements in medical AI.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our approach on two representative multi-modal medical datasets encompassing diverse clinical tasks. The first dataset is derived from the MIMIC-IV-MM [42] database, which integrates data from MIMIC-CXR-JPG [18], MIMIC-IV-Note, and MIMIC-IV [19]. This comprehensive dataset combines chest X-ray images (XRay), radiology reports (Note), and time-series EHR data containing 46 key clinical variables (Time). We use a 0.72-0.13-0.15 split ratio for training, validation, and testing respectively. The second, MMIST ccRCC [38], focuses on clear cell renal cell carcinoma and comprises CT scans,

Method Type	Method	Mortality						Longstay						Readmission					
		Acc	Auroc	Auprc	mAP	mAR	mF1	Acc	Auroc	Auprc	mAP	mAR	mF1	Acc	Auroc	Auprc	mAP	mAR	mF1
Finetuned Unified Multimodal Encoder	Meta-Transformer	0.562	0.717	0.309	0.561	0.641	0.489	0.615	0.662	0.736	0.605	0.607	0.605	0.482	0.582	0.075	0.517	**0.602**	0.372
	OneLLM	0.736	0.511	0.208	0.501	0.501	0.497	0.453	0.505	0.638	0.502	0.501	0..434	0.866	0.521	**0.118**	0.509	0.519	0.506
Baseline	XRay-Large	0.806	0.685	0.236	0.573	0.581	0.576	0.600	0.617	0.680	0.586	0.585	0.586	0.919	0.583	0.056	0.499	0.499	0.499
	Text-Large	0.829	0.765	0.341	0.622	0.637	0.628	0.618	0.654	0.727	0.602	0.598	0.599	0.915	0.544	0.052	0.503	0.504	0.503
	TimeSeries-Large	0.776	0.529	0.142	0.519	0.521	0.519	0.610	0.630	0.687	0.603	0.606	0.603	**0.955**	0.592	0.056	0.478	0.499	0.489
	Large-Ensemble	0.844	0.774	0.343	0.645	0.653	0.649	0.633	0.664	0.734	0.618	0.613	0.614	0.915	0.563	0.060	0.520	0.523	0.521
	MultimodalPred-Small	0.806	0.641	0.197	0.564	0.569	0.567	0.657	0.690	0.747	0.646	0.647	0.646	0.922	0.548	0.054	0.515	0.514	0.514
Ours	AdaCoMed	**0.862**	**0.805**	**0.366**	**0.673**	**0.654**	**0.663**	**0.674**	**0.732**	**0.782**	**0.667**	**0.671**	**0.668**	0.931	**0.619**	0.080	**0.533**	0.525	**0.528**

Table 2. **Performance comparison of different methods on MIMIC-IV-MM.** We evaluate three critical healthcare prediction tasks: mortality, longstay, and readmission. Models are grouped into: (1) Pre-trained foundation models, (2) Baselines including unimodal and traditional multimodal methods and (3) AdaCoMed. The best results in each column are shown in **bold** and second best in underline.

Method	Mortality						Longstay						Readmission					
	Acc	Auroc	Auprc	mAP	mAR	mF1	Acc	Auroc	Auprc	mAP	mAR	mF1	Acc	Auroc	Auprc	mAP	mAR	mF1
Ensemble	0.784	0.664	0.193	0.562	0.581	0.567	0.648	0.679	0.724	0.635	0.633	0.634	0.922	0.542	0.049	0.501	0.501	0.501
Leanable Weighted	0.840	0.786	0.338	0.640	0.651	0.645	0.669	0.718	0.775	0.658	0.657	0.658	0.930	0.628	0.064	0.507	0.505	0.505
Inverse Contrastive	**0.862**	**0.805**	0.366	**0.673**	0.654	**0.663**	0.662	0.718	0.770	0.649	0.644	0.645	0.927	**0.637**	**0.100**	0.521	0.517	0.518
Gaussian Contrastive	0.846	0.799	**0.380**	0.654	**0.673**	**0.663**	**0.674**	**0.732**	**0.782**	**0.667**	**0.671**	**0.668**	**0.931**	0.619	0.080	**0.533**	**0.525**	**0.528**
Threshold Contrastive	0.844	0.802	0.361	0.652	0.670	0.660	0.664	0.714	0.767	0.652	0.648	0.650	0.925	0.601	0.074	0.503	0.502	0.502

Table 3. **Performance comparison of different adaptive weighting schemes of our AdaCoMed model on MIMIC-IV-MM.** Our proposed strategies include learnable weighted fusion and contrastive-based approaches, where each method shows task-specific advantages. The ensemble baseline averages the logits from large and small models. The best results in each column are shown in **bold**.

whole slide images (WSI) from surgical pathology, and clinical data with 20 key clinical variables (Clinical). We use the same train and test split as the dataset mentioned.

Metrics. To evaluate the performance, we employ multiple metrics. The basic classification performance is measured using Accuracy (Acc) and Area Under the Receiver Operating Characteristic curve (Auroc). Considering potential class imbalance, we include the Area Under the Precision-Recall Curve (Auprc). We also report macro-averaged metrics across all classes: macro Average Precision (mAP), macro Average Recall (mAR), and macro F1-score (mF1).

Baselines. We compare our approach against broad representative baselines, which we categorize into three groups: (1) Finetuned Unified Multimodal Encoders, including Meta-Transformer and OneLLM that are fine-tuned on our medical tasks; (2) Modality-specific baselines, where "Modality name"-Large models (e.g., XRay-Large) refer to the methods that use corresponding large models as encoders followed by task-specific heads. We employ Dinov2 [39] for X-ray images, CLAM [33]-preprocessed and UNI [7]-extracted features for WSI, Merlin [3] for CT scans, MOIRAI [48] for time series data, and Me-Llama-13B [51] for textual information. Clinical data is converted to text format through prompting methods before encoding; (3) Fusion methods, including Large-Ensemble that concatenates features from modality encoders, and lighter-weight small model MultimodalPred-Small [46] that uses modality-specific networks followed by fusion networks.

Implementation. All experiments were conducted on NVIDIA A100 GPUs. For pre-training our task-driven small multi-modal model, we initialized the learning rate at 1e-4 with Cosine LR scheduler [32], and employed Adam [21] optimizer. Class weights were incorporated to address class imbalance in the training data. For the Large-small collaboration model, we maintained the same learning rate and optimizer settings, while setting the contrastive loss weight to 0.5 with a temperature of 0.07. We configured the inverse temperature to 1.0, gaussian standard deviation to 1.0, and threshold slope to 10 as default in adaptive weighting strategy. All training procedures were limited to 100 epochs with early stopping mechanisms implemented.

4.2. Comparison with the State-of-the-Arts

Results on MIMIC-IV-MM. In Tab. 2, the experimental results on MIMIC-IV-MM dataset demonstrate the superior performance of AdaCoMed across three critical healthcare prediction tasks. For mortality prediction, our AdaCoMed variant surpasses the Large-Ensemble baseline by 3.1% in Auroc, while achieving a notable 57.5% improvement in Auroc over OneLLM. In the length-of-stay prediction task, AdaCoMed outperforms the best baseline approach (MultimodalPred-Small) by a significant margin of 6.1% in Auroc and demonstrating a 3.4% improvement in mF1. For readmission prediction, while maintaining comparable accuracy with baseline methods, our AdaCoMed achieves substantial improvements in mAP metrics, particularly in Auroc and mF1. Furthermore, while some baselines achieve higher scores in specific metrics, these apparent advantages often come with significant trade-offs: Meta-Transformer's higher mAR is offset by lower stability metrics, OneLLM's better Auprc comes with poor accuracy suggesting overfitting, and TimeSeries-Large's high Acc likely stems from bias on imbalanced dataset. Our AdaCoMed maintains more robust performance across all cases.

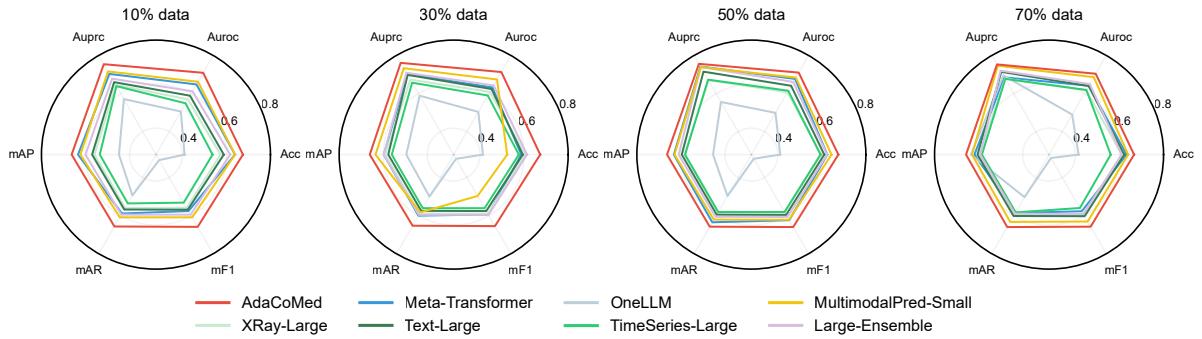

Figure 4. **Performance comparison of different methods on different training data ratios of MIMIC-IV-MM longstay task.** Each subplot represents the performance of various training data ratios (10%, 30%, 50%, 70%). The radial axes show the performance value for each metric, ranging from 0.28 to 0.8.

Method Type	Method	Vital-12					
		Acc	Auroc	Auprc	mAP	mAR	mF1
Finetuned Unified Multimodal Encoder	Meta-Transformer	0.896	0.694	0.972	0.464	0.481	0.473
	OneLLM	0.868	0.531	0.939	0.507	0.505	0.504
Baseline	WSI-Large	**0.931**	0.750	0.978	0.466	0.500	0.482
	CT-Large	**0.931**	0.759	0.981	0.466	0.500	0.482
	Clinical-Large	0.862	0.597	0.963	0.463	0.463	0.463
	Large-Ensemble	0.897	0.778	0.983	0.464	0.481	0.473
	MultimodalPred-Small	0.621	0.769	0.980	0.548	0.681	0.482
Ours	AdaCoMed	0.862	**0.898**	**0.992**	**0.605**	**0.694**	**0.628**

Table 4. **Performance comparison of different methods on MMIST ccRCC.** Methods are evaluated on downstream task: Vital 12. The best results in each column are shown in **bold** and second best in underline.

Method	Vital-12					
	Acc	Auroc	Auprc	mAP	mAR	mF1
Ensemble	0.810	0.822	0.986	0.525	0.551	0.524
Learnable Weighted	0.828	0.810	0.985	0.532	0.560	0.535
Inverse Contrastive	**0.862**	0.847	0.987	**0.640**	**0.810**	**0.675**
Gaussian Contrastive	**0.862**	0.856	0.989	0.554	0.579	0.562
Threshold Contrastive	**0.862**	**0.898**	**0.992**	0.605	0.694	0.628

Table 5. Performance comparison of adaptive weighting schemes on MMIST ccRCC. The best results in each column are shown in **bold**.

Results on MMIST. As shown in Tab. 4, we further conduct experiments on MMIST ccRCC for the task of 12-month survival prediction. In this setting, Large-MultimodalPred-based AdaCoMed significantly outperforms the best baseline approach (Large-Ensemble), showing an impressive improvement of 15.4% in Auroc and 32.8% in mF1. Notably, although WSI-Large and CT-Large achieved higher Acc scores, these results were primarily due to prediction collapse arising from label imbalance in the dataset, which led to biased predictions. This imbalance also negatively impacted the Auroc and mF1 of both methods. In contrast, AdaCoMed demonstrated balanced and robust performance across all evaluation metrics. Besides, AdaCoMed also surpasses pre-trained models like Meta-Transformer and OneLLM, which are often unable to achieve optimal performance when finetuning on small datasets like MMIST. This further emphasizes the significance of AdaCoMed proving its capability to excel when multi-modal data is scarce.

4.3. Further Analysis

Analysis on different adaptive weighting strategies. As shown in Tab. 3, we evaluate different adaptive weighting schemes across three clinical prediction tasks on MIMIC-IV-MM dataset. The ensemble baseline simply averages the logits from large and small models, serving as a basic fusion strategy. Our proposed methods demonstrate task-specific advantages: the Gaussian Contrastive approach excels in longstay and readmission prediction, and Inverse Contrastive achieves superior performance in mortality prediction. These collaborative inference strategies effectively leverage both models' strengths based on their confidence levels, demonstrating the flexibility of confidence-based model selection in diverse clinical scenarios. For the MMIST dataset, as recorded in Tab. 5, all adaptive weighting methods perform comparably, with the Threshold Contrastive approach achieving the best results. Notably, most of our adaptive weighting strategies outperform the best baselines and pre-trained unified multimodal encoders, demonstrating the stability of our proposed AdaCoMed framework.

Analysis on training data ratio. Since the MIMIC-IV-MM dataset is relatively large within multimodal datasets, we conducted experiments on longstay tasks using different data ratios (10%, 30%, 50%, 70%) to analyze both model performance and training data sensitivity in Fig. 4. Overall, AdaCoMed achieves the best performance across all metrics and data ratios. By leveraging single-modal large model architectures' characteristics, models like XRay-Large and our AdaCoMed demonstrate relatively stable performance across varying data ratios. In contrast, models based on traditional multi-modal fusion architectures, such as MultimodalPred-Small, and non-medical pretrained unified multi-modal encoders, such as Meta-Transformer and OneLLM, exhibit more fluctuations in performance, sug-

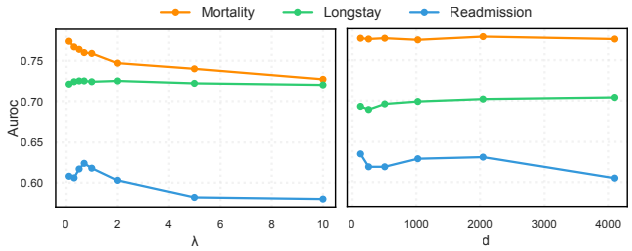

Figure 5. **Analysis of λ and d on three tasks of MIMIC-IV-MM dataset.** λ is the balancing parameter between $\mathcal{L}_{\text{task}}$ and $\mathcal{L}_{\text{contrast}}$ and d is the dimension of shared projection space.

Method	Parameter Num (M)	GPU Memory Per Batch (GB)	Training Time Per Epoch (min)
Meta-Transformer	302.8	19.9	41.89
OneLLM	383.0	55.6	74.53
MultimodalPred-Small	49.7	2.4	2.09
AdaCoMed	**21.3**	**1.0**	**0.02**

Table 6. **Comparison of training cost across different methods on MIMIC-IV-MM,** with parameter numbers counted in Millions (M), GPU memory per batch (1 for OneLLM, 64 for others) in Gigabyte (GB) and training time per epoch in Minutes (min). Least costs are shown in **bold**.

gesting greater sensitivity to the amount of training data.

Analysis on training parameter scales. Beyond its notable performance, AdaCoMed also exhibits remarkable efficiency in training costs. In Tab. 6, we present the training costs for various methods, where Meta-Transformer is fully finetuned while OneLLM is tuned with its LLM component frozen. Among these methods, AdaCoMed stands out by requiring the fewest trainable parameters. Notably, the parameter count needed to fine-tune a pre-trained model is an entire order of magnitude higher than that of AdaCoMed, and even training a smaller multimodal model demands significantly more resources than AdaCoMed. This comparison highlights AdaCoMed's advantages in parameter optimization and resource efficiency, achieving superior performance while minimizing computational requirements.

Impact of hyper-parameters λ and d. Our method employs two crucial hyper-parameters including λ that balances $\mathcal{L}_{\text{task}}$ and $\mathcal{L}_{\text{contrast}}$ and common projection dimension d. Analysis on MIMIC-IV-MM (Fig. 5) reveals task-dependent sensitivity to λ: Mortality and Readmission exhibit performance degradation at larger values, while Longstay remains stable. Empirically, $\lambda \in [0.1, 1]$ performs well across all tasks. About dimension d, it exhibits more stable behavior. Performance generally plateaus around $d = 2048$, with Mortality maintaining consistently high Auroc, while Longstay shows gradual improvement and Readmission remains relatively stable across different projection dimensions. This suggests our method is robust to projection dimension choice, with $d \in [512, 2048]$ providing optimal results.

Visualization of co-learning. Our 3D visualization (Fig. 6)

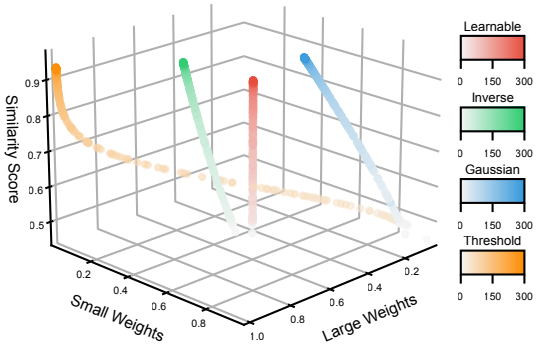

Figure 6. **Visualization of weight evolution and similarity trends for different weighting strategies on MIMMIC-IV-MM longstay task.** The 3D scatter plot shows the relationship between large weights (x-axis), small weights (y-axis), and similarity scores of large and small representations (z-axis) during training. Each color represents a different weighting strategy: Learnable (red), Inverse (green), Gaussian (blue), and Threshold (orange). The color gradients indicate the training progression from step 0 to step 300, with lighter shades representing earlier steps and darker shades representing later steps. All strategies demonstrate a general trend of similarity closer as training progresses, with distinct convergence patterns in the weight space.

illustrates distinct convergence patterns among four weighting strategies, mapping large weights (x-axis), small weights (y-axis), and similarity score (z-axis) over training steps 0–300. All strategies start with low similarity and converge to higher similarity, following unique paths. The Learnable Weighted strategy (red) maintains a stable, near-vertical path with minimal weight variation. Inverse Contrastive (green) shows a rapid descent, while Gaussian Contrastive (blue) converges slowly, and Threshold Contrastive (orange) shows sharp transitions between weight states.

5. Conclusion

In this paper, we introduce AdaCoMed, the first multimodal medical model designed with large-small model collaboration in diagnostic classification. The primary innovation lies in efficiently leveraging existing large single-modal models alongside traditional small multi-modal models for collaborative diagnostic classification. We propose a novel approach for feature fusion with large models and a collaborative classification framework involving both large and small models. AdaCoMed achieves superior performance across four medical classification tasks on two medical datasets. In the future, we will explore more multi-modal scenarios under such a collaboration paradigm.

Acknowledgement

This work is supported by the National Key R&D Program of China (2022ZD0160703), and the National Natural Science Foundation of China (62202422 and 62372408).

References

[1] Ayelet Akselrod-Ballin, Michal Chorev, Yoel Shoshan, Adam Spiro, Alon Hazan, Roie Melamed, Ella Barkan, Esma Herzel, Shaked Naor, Ehud Karavani, et al. Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology*, 292(2):331–342, 2019. 2

[2] Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*, 2024. 2

[3] Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, et al. Merlin: A vision language foundation model for 3d computed tomography. *arXiv preprint arXiv:2406.06512*, 2024. 6

[4] Yiming Cao, Lizhen Cui, Lei Zhang, Fuqiang Yu, Zhen Li, and Yonghui Xu. Mmtn: multi-modal memory transformer network for image-report consistent medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 277–285, 2023. 2

[5] Nimeesha Chan, Felix Parker, William Bennett, Tianyi Wu, Mung Yao Jia, James Fackler, and Kimia Ghobadi. Medtsllm: Leveraging llms for multimodal medical time series analysis. *arXiv preprint arXiv:2408.07773*, 2024. 2

[6] Mingcai Chen, Yuntao Du, Yi Zhang, Shuwei Qian, and Chongjun Wang. Semi-supervised learning with multi-head co-training. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6278–6286, 2022. 5

[7] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. 6

[8] Wanyi Chen, Jianjun Yang, Zhongquan Sun, Xiang Zhang, Guangyu Tao, Yuan Ding, Jingjun Gu, Jiajun Bu, and Haishuai Wang. Deepasd: a deep adversarial-regularized graph learning method for asd diagnosis with multimodal data. *Translational Psychiatry*, 14(1):375, 2024. 2

[9] Zhihong Chen, Shizhe Diao, Benyou Wang, Guanbin Li, and Xiang Wan. Towards unifying medical vision-and-language pre-training via soft prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23403–23413, 2023. 3

[10] Matthew Christensen, Milos Vukadinovic, Neal Yuan, and David Ouyang. Vision–language foundation model for echocardiogram interpretation. *Nature Medicine*, pages 1–8, 2024. 2

[11] Chaitanya Dwivedi, Shima Nofallah, Maryam Pouryahya, Janani Iyer, Kenneth Leidal, Chuhan Chung, Timothy Watkins, Andrew Billin, Robert Myers, John Abel, et al. Multi stain graph fusion for multimodal integration in pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1835–1845, 2022. 2

[12] Hui Fang, Haishuai Wang, Yang Gao, Yonggang Zhang, Jiajun Bu, Bo Han, and Hui Lin. Insgnn: Interpretable spatio–temporal graph neural networks via information bottleneck. *Information Fusion*, page 102997, 2025.

[13] Jianliang Gao, Tengfei Lyu, Fan Xiong, Jianxin Wang, Weimao Ke, and Zhao Li. Mgnn: A multimodal graph neural network for predicting the survival of cancer patients. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1697–1700, 2020. 2

[14] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26584–26595, 2024. 3

[15] Xinrui He, Yikun Ban, Jiaru Zou, Tianxin Wei, Curtiss B Cook, and Jingrui He. Llm-forest for health tabular data imputation. *arXiv preprint arXiv:2410.21520*, 2024. 2

[16] Yuting He, Fuxiang Huang, Xinrui Jiang, Yuxiang Nie, Minghao Wang, Jiguang Wang, and Hao Chen. Foundation model for advancing healthcare: challenges, opportunities and future directions. *IEEE Reviews in Biomedical Engineering*, 2024. 2

[17] Javon Hickmon. Multimodal ensembling for zero-shot image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 23747–23749, 2024. 5

[18] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019. 5

[19] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023. 5

[20] Sein Kim, Namkyeong Lee, Junseok Lee, Dongmin Hyun, and Chanyoung Park. Heterogeneous graph learning for multi-modal medical data analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5141–5150, 2023. 2

[21] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[22] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024. 2

[23] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. 2

[24] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon,

and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[25] Mingcheng Li, Dingkang Yang, Yuxuan Lei, Shunli Wang, Shuaibing Wang, Liuzhen Su, Kun Yang, Yuzheng Wang, Mingyang Sun, and Lihua Zhang. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10074–10082, 2024. 5

[26] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6), 2023. 2

[27] Xiwen Liang, Yangxin Wu, Jianhua Han, Hang Xu, Chunjing Xu, and Xiaodan Liang. Effective adaptation in multi-task co-training for unified autonomous driving. *Advances in Neural Information Processing Systems*, 35:19645–19658, 2022. 5

[28] Fenglin Liu, Tingting Zhu, Xian Wu, Bang Yang, Chenyu You, Chenyang Wang, Lei Lu, Zhangdaihong Liu, Yefeng Zheng, Xu Sun, et al. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine*, 6(1): 226, 2023. 2

[29] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A. Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023. 2

[30] Yucheng Liu, Ziyu Jia, and Haichao Wang. Emotionkd: a cross-modal knowledge distillation framework for emotion recognition based on physiological signals. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6122–6131, 2023. 5

[31] Zhicheng Liu, Ali Braytee, Ali Anaissi, Guifu Zhang, Lingyun Qin, and Junaid Akram. Ensemble pretrained models for multimodal sentiment analysis using textual and video data fusion. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1841–1848, 2024. 5

[32] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[33] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021. 6

[34] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 1, 2

[35] Kaden McKeen, Laura Oliva, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. Ecg-fm: An open electrocardiogram foundation model. *arXiv preprint arXiv:2408.05178*, 2024. 2

[36] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023. 2

[37] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023. 3

[38] Tiago Mota, M Rita Verdelho, Diogo J Araújo, Alceu Bissoto, Carlos Santiago, and Catarina Barata. Mmist-ccrcc: A real world medical dataset for the development of multimodal systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2395–2403, 2024. 5

[39] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6

[40] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620 (7972):172–180, 2023. 1, 2

[41] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023. 1

[42] Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, 5(1):149, 2022. 2, 5

[43] Wei Tang, Fazhi He, Yu Liu, and Yansong Duan. Matr: Multimodal medical image fusion via multiscale adaptive transformer. *IEEE Transactions on Image Processing*, 31:5134–5149, 2022. 2

[44] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kunz, Matthew C. H. Lee, Jan Bernhard, Ran A. Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Juan Retamero, William A. Moye, Razik Yousfi, Christopher Kanan, David Klimstra, Brandon Rothrock, and Thomas J. Fuchs. Virchow: A million-slide digital pathology foundation model, 2024. 2

[45] Zhongwei Wan, Che Liu, Mi Zhang, Jie Fu, Benyou Wang, Sibo Cheng, Lei Ma, César Quilodrán-Casas, and Rossella Arcucci. Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[46] Yuanlong Wang, Changchang Yin, and Ping Zhang. Multimodal risk prediction with physiological signals, medical images and clinical notes. *Heliyon*, 10(5), 2024. 6

[47] Zifeng Wang, Zichen Wang, Balasubramaniam Srinivasan, Vassilis N Ioannidis, Huzefa Rangwala, and Rishita Anubhai. Biobridge: Bridging biomedical foundation models via knowledge graph. *arXiv preprint arXiv:2310.03320*, 2023. 2

[48] Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Forty-first International Conference on Machine Learning*, 2024. 6

[49] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023. 2

[50] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023. 3

[51] Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, et al. Me llama: Foundation large language models for medical applications. *arXiv preprint arXiv:2402.12749*, 2024. 2, 6

[52] Canwen Xu, Yichong Xu, Shuohang Wang, Yang Liu, Chenguang Zhu, and Julian McAuley. Small models are valuable plug-ins for large language models, 2023. 5

[53] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, pages 1–8, 2024. 2

[54] Qinmei Xu, Xianghao Zhan, Zhen Zhou, Yiheng Li, Peiyi Xie, Shu Zhang, Xiuli Li, Yizhou Yu, Changsheng Zhou, Longjiang Zhang, et al. Ai-based analysis of ct images for rapid triage of covid-19 patients. *NPJ digital medicine*, 4(1): 75, 2021. 2

[55] Zhen Xu, David R So, and Andrew M Dai. Mufasa: Multimodal fusion architecture search for electronic health records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10532–10540, 2021. 2

[56] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024. 3

[57] Kaiyan Zhang, Jianyu Wang, Ning Ding, Biqing Qi, Ermo Hua, Xingtai Lv, and Bowen Zhou. Fast and slow generating: An empirical study on large and small language models collaborative decoding. *arXiv preprint arXiv:2406.12295*, 2024. 5

[58] Kaiyan Zhang, Jianyu Wang, Ermo Hua, Biqing Qi, Ning Ding, and Bowen Zhou. Cogenesis: A framework collaborating large and small language models for secure context-aware instruction following. *arXiv preprint arXiv:2403.03129*, 2024. 5

[59] Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023. 2

[60] Ziping Zhao, Tian Gao, Haishuai Wang, and Björn Schuller. Mfdr: Multiple-stage fusion and dynamically refined network for multimodal emotion recognition. In *Proc. Interspeech 2024*, pages 3719–3723, 2024. 2

[61] Shuai Zheng, Zhenfeng Zhu, Zhizhe Liu, Zhenyu Guo, Yang Liu, Yuchen Yang, and Yao Zhao. Multi-modal graph learning for disease prediction. *IEEE Transactions on Medical Imaging*, 41(9):2207–2216, 2022. 2

[62] Hong-Yu Zhou, Yizhou Yu, Chengdi Wang, Shu Zhang, Yuanxu Gao, Jia Pan, Jun Shao, Guangming Lu, Kang Zhang, and Weimin Li. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nature biomedical engineering*, 7 (6):743–755, 2023. 2

[63] Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023. 2