# FLEXIBLE MODALITY INTEGRATION FOR REAL-WORLD MEDICAL AI: HANDLING STRUCTURAL DISTINCTION, HETEROGENEITY, AND ASYNCHRONICITY IN MULTIMODAL HEALTHCARE DATA

FENG YIDAN

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

School of Nursing

# Flexible Modality Integration for Real-World Medical AI: Handling Structural Distinction, Heterogeneity, and Asynchronicity in Multimodal Healthcare Data

**FENG Yidan**

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

June 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

FENG Yidan
_____ (Name of student)

# Abstract

Medical artificial intelligence demands robust multimodal fusion to integrate diverse data streams, including anatomical/functional imaging, heterogeneous clinical variables, and irregular longitudinal measurements, for comprehensive clinical insights. However, existing multimodal fusion methods typically assume fixed modality availability, severely limiting their real-world applicability in dynamic clinical environments characterized by institutional resource disparities, patient-specific contraindications, evolving diagnostic workflows, and temporal irregularities in data acquisition. Consequently, flexible modality integration is indispensable for clinical translation, yet significant technical challenges persist: 1) reliance on complete modality sets during training severely limits data utilization and generalization to partial inputs; 2) existing inter-modal alignment strategies inadequately preserve task-specific unique semantics while adapting to dynamically changing inputs; 3) architectural inflexibility hinders scalable integration of novel modalities; and 4) effective modeling of asynchronous temporal-modality dependencies remains critically underexplored. This thesis addresses these core challenges by developing a set of solutions for clinically adaptive multimodal learning, enabling robust integration of arbitrary modality subsets across diverse medical scenarios. Three clinically representative applications were selected to validate

our approach across the multimodal integration spectrum:

1. Multimodal MRI synthesis: A typical dense prediction task where complementary sequences are fundamental for soft-tissue characterization yet frequently compromised by variable acquisition success in clinical practice. A unified method is proposed to reconcile the artificial fragmentation between cross-modality synthesis (CMS) and multi-contrast super-resolution (MCSR) through fine-grained difference learning. Spatial misalignments inherent in clinical scans are resolved via multi-scale deformable convolutions, while modality-specific structures distinction is recovered through a synergistic mechanism comprising: a difference projection discriminator, distinction-aware feature regularization, and incremental feature modulation. This approach achieves consistent high-fidelity reconstruction across extreme degradation levels (2–16$\times$ undersampling), significantly outperforming task-specific alternatives.

2. Alzheimer's diagnosis with heterogeneous modalities: A prevalent clinical condition requiring diagnostic synthesis of diverse and inherently imbalanced multimodal data. The proposed AnyMod architecture addresses combinatorial missing-modality complexity and semantic heterogeneity by enabling training and inference on arbitrary combinations of imaging and non-imaging data. Its core innovations include representation-task decoupled alignment—preserving modality-unique semantics while mapping heterogeneous inputs to class-invariant prototypes, along with modality-agnostic Transformer projectors that eliminate dedicated encoders, and dynamic token clustering ensuring computational scalability across modality combi-

nations. Validation demonstrates increasing performance advantages over combination-specific models as modality count grows, with seamless extensibility to unseen modalities.

3. Dynamic (Acute Respiratory Distress Syndrome) ARDS risk monitoring with asynchronous modalities: A critical adverse event in ICU demanding continuous risk assessment from inherently asynchronous data streams (sparse CXRs, high-frequency vitals, intermittent labs). Effective integration of these irregularly sampled modalities is achieved through modality-wise encoding with adaptive positional encodings that preserve temporal-semantic relationships. The framework incorporates a Staged Temporal-Modal Fusion module decoupling cross-modal interaction from temporal processing, complemented by Progressive Context Memory enabling computationally efficient long-range dependency modeling. The framework provides hourly risk stratification with time-to-onset quantification (AUROC 0.94 <6h pre-onset), revealing 20-fold ARDS incidence in high-risk cohorts.

All methods are validated on publicly-available datasets, demonstrating performance gains over state-of-the-art techniques. By systematically addressing clinical and technical barriers, including data inefficiency, semantic heterogeneity, architectural rigidity, and temporal irregularities, this work advances multimodal learning toward clinically adaptive, data-efficient, and equitable AI-driven healthcare.

**Keywords:** Multimodal Learning, Heterogeneous Data, Medical Image Synthesis, Disease Diagnosis, Alzheimer's Disease, Risk Prediction, Intensive Care

Unit (ICU), Dynamic Monitoring, Deep Learning

# List of Publications

1. **Yidan Feng**, Sen Deng, Jun Lyu, Jing Cai, Mingqiang Wei, Jing Qin (2024). Bridging MRI Cross-Modality Synthesis and Multi-Contrast Super-Resolution by Fine-Grained Difference Learning. *IEEE Transactions on Medical Imaging* (TMI), vol. 44, no. 1, pp. 373-383 (IF:8.9)

2. **Yidan Feng**, Bingchen Gao, Sen Deng, Anqi Qiu, Jing Qin (2024). Unified Multi-modal Learning for Any Modality Combinations in Alzheimer's Disease Diagnosis. *International Conference on Medical Image Computing and Computer-Assisted Interventions* (MICCAI), pp. 487–497

3. **Yidan Feng**, Bohan Zhang, Sen Deng, Jing Qin (2025). Asynchronous Multi-Modal Learning for Dynamic Risk Monitoring of Acute Respiratory Distress Syndrome in Intensive Care Units. *International Conference on Medical Image Computing and Computer-Assisted Interventions* (MICCAI), accepted

4. Sen Deng, **Yidan Feng**, Haoneng Lin, Yiting Fan, Alex Pui-Wai Lee, Xiaowei Hu, Jing Qin (2024). Semi-supervised TEE segmentation via interacting with SAM equipped with noise-resilient prompting. *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI), pp. 11757-11765

5. Haoneng Lin, Jing Zou, Kang Wang, **Yidan Feng**, Cheng Xu, Jun Lyu, Jing Qin (2024). Dual-space high-frequency learning for transformer-based MRI super-resolution. *Computer Methods and Programs in Biomedicine*, p.108165

# Acknowledgements

First and foremost, I express the deepest thanks to my chief supervisor, Professor Jing Qin, for his constant support and invaluable guidance throughout my PhD journey. Beyond imparting research skills, he demonstrated how to conduct scientific work with compassion, humor, and balance. He taught me that taking breaks is part of good work, that responsibility matters, and that collaboration grows from mutual understanding. Under his mentorship, I not only advanced academically but evolved personally, cultivating independence, adaptability, and collaborative spirit. It was a great honor to work and study under his guidance.

I am deeply grateful to my colleagues at the Centre for Smart Health, whose companionship greatly enriched my doctoral experience. Our collective efforts, from lively brainstorming sessions to weekend hikes, cultivated an environment where optimism and mutual support flourished. Your constructive feedback significantly enhanced my work, while your friendship underscored the importance of community in scientific endeavors.

To my family and my partner Sen Deng: your unconditional love provided the foundation that sustained me through every challenge. When doubts clouded my research path, you were the light that reminded me of the value of perseverance. This milestone is as much yours as mine.

Lastly, I would like to acknowledge the administrative staff of the Hong Kong Polytechnic University School of Nursing. Your efficient management of logistics, resources, and encouragement created an ideal ecosystem where creativity and scholarship intersect.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Motivation

Advancements in medical technology have enabled the acquisition of diverse, complementary data modalities, from anatomical imaging (CT, MRI) to functional scans (PET) and heterogeneous clinical data (EHRs, genomic profiles), facilitating a holistic assessment of patient health. The integration of multiple data modalities offers significant potential for improving diagnostic precision, especially in complex diseases where reliance on a single data source is inadequate. Consequently, deep learning-based multimodal fusion has emerged as a critical tool for developing objective, data-driven computer-aided diagnostics. Unlike extensive multimodal studies on vision, language, or audio domains [1, 2, 3, 4], medical multimodal learning confronts distinctive challenges: 1) inherent richness of medical imaging modalities, such as X-ray attenuation (CT), nuclear magnetic resonance (MRI), and metabolic activity (PET), which capture spatially correlated anatomical structures yet exhibit modality-specific physical principles and structural rep-

resentations 2) integration of semantically misaligned, non-imaging data (e.g., lab tests, electronic health records), and 3) temporal irregularities in longitudinally sampled modalities. These characteristics demand specialized fusion strategies for clinically robust decision-making.

Despite the proven superiority of medical multimodal AI over unimodal baselines under fixed modality combinations [5, 6, 7, 8], these methods face four fundamental limitations impeding clinical adoption. First, existing research remains artificially fragmented into compartmentalized tasks constrained by static modality inputs, which forces redundant development. For example, in MRI synthesis, closely related objectives, including cross-modality translation, missing modality imputation and multimodal MRI reconstruction, all address target modality synthesis yet evolve as distinct research directions. In Alzhermer's Disease Diagnosis, different papers focused on different modality combinations, including T1-MRI & FDG-PET [9, 10], T1-MRI & specific tabular data [5, 11, 12], Amyloid-PET, T1-MRI & FDG-PET [13, 14], although they share the same task on improving diagnostic accuracy. Second, this artificial division ignores clinical reality that optimal modality availability is dynamic, determined by 1) institutional resource disparities between well-equipped tertiary hospitals (offering advanced modalities like PET/MRI) and resource-constrained clinics (limited to basic technologies such as X-ray or ultrasound), 2) patient-specific contraindications including metal implants prohibiting MRI, radiation risks during pregnancy, or contrast agent allergies, 3) time-critical emergencies where rapid decisions are paramount (e.g., stroke triage requiring immediate CT-based decisions), and 4) progressive diagnostic workflows involving iterative evidence accumulation from initial laboratory testing to subsequent advanced imaging. Third, methods assuming fixed modality

combinations require complete multimodal training samples, discarding valuable partial-modality data abundant in real-world settings. This constraint forces models to train on only a small subset of available data, often representing ideal cases rather than clinical reality, leading to overfitting on narrow data distributions and severely limiting generalization to diverse modality combinations encountered in practice. Lastly, static fusion models lack adaptability to evolving medical knowledge. Emerging technologies like portable MRI scanners or spatial transcriptomics platforms, alongside protocol updates, render inflexible systems obsolete, necessitating costly retraining cycles.

To bridge these gaps, this thesis studies flexible modality integration, where models robustly adapt to arbitrary modality subsets or sequences, aiming to provide unified methodological frameworks that consolidates fragmented research trajectories for medical multimodal AI applications. To summarize, the motivation for flexible multimodal integration is: 1) Ensuring clinical resilience across resource-constrained environments and dynamic patient scenarios to promote equitable healthcare access. 2) Maximizing data efficiency through training without full-modality requirements, enhancing generalization to both target tasks and unseen modality combinations. 3) Enabling adaptable architectures with potential capacity for emerging modalities, facilitating sustainable AI evolution alongside medical innovation.

## 1.2 Challenges

Despite growing efforts to enhance robustness against missing modalities in medical multimodal AI, significant challenges persist. First, current approaches, partic-

ularly imputation-based techniques, improve inference-stage robustness through modality completion but remain fundamentally limited by demanding complete multimodal availability during training. This requirement severely constrains data utilization, as real-world medical datasets inherently contain abundant partial-modality samples due to clinical constraints and data heterogeneity. Consequently, models fail to leverage the full spectrum of available training data, compromising their ability to generalize across variable modality combinations. Second, effective modeling of inter-modal relationships presents unresolved complexities. While existing alignment methods predominantly focus on cross-modal similarity, they inadequately address two critical needs: 1) preserving and leveraging task-specific modality-unique information during similarity modeling, and 2) maintaining robust inter-modal relationship modeling under dynamically changing input configurations. Third, architectural extensibility for emerging modalities is underexplored. Most frameworks lack mechanisms for seamless integration of novel data types without costly retraining or structural overhauls, hindering sustainable adaptation to evolving clinical technologies. Lastly, temporal modeling represents a critical frontier. Few studies address the integration of irregularly sampled longitudinal modalities, where time-series data from diverse sources exhibit complex temporal-modality inter-dependencies. Moreover, the application of such temporal-modality modeling frameworks to dynamic clinical prediction tasks is underexplored, despite its critical potential for generating actionable insights into disease progression trajectories and individualized risk stratification.

This thesis investigates three typical medical applications spanning two essential medical AI paradigms: dense prediction tasks applied to multimodal medical imaging, requiring pixel/voxel-level precision with emphasis on anatomical

fidelity and structural preservation; and decision-centric tasks integrating heterogeneous or time-dependent non-imaging modalities for clinical decision support. Subsequent sections will examine the distinct technical challenges of each application.

### 1.2.1 Multimodal MRI Synthesis: Integration of Multi-Modal Images

Compared to decision-centric tasks, dense prediction tasks for multimodal medical imaging exhibit distinct characteristics. These tasks typically involve fewer modalities, with explicit anatomical alignment expectations across imaging sources. However, acquired multimodal scans rarely achieve perfect spatial registration in practice. Furthermore, while multimodal generation research has proliferated across specialized branches, including missing modality imputation, multimodal reconstruction, and cross-modality translation, these efforts remain fragmented despite sharing the core objective of target modality synthesis. It remains unexplored on the fundamental connections between these related tasks.

Multimodal MRI sequences constitute a well-established research domain [15], as clinical protocols often require complementary sequences for comprehensive soft-tissue characterization despite varying acquisition difficulties. Using flexible integration of dual MRI modalities for target synthesis as our methodological entry point, three specific challenges are identified:

- Current approaches remain fragmented by modality availability constraints. Simultaneous high-resolution acquisition across multiple MRI sequences remains challenging due to widely varying scan times [16]. This has bifur-

cated research into two disconnected paradigms: cross-modality synthesis (CMS) when target sequences are entirely absent, and multi-contrast super-resolution (MCSR) when only undersampled targets exist [17, 18]. Though both fundamentally aim to reconstruct high-resolution target-modality images within the same anatomical space, no unified framework adapts to this modality availability spectrum.

- Spatial alignment assumptions fail to reflect clinical realities. Most methodologies presume perfectly coregistered inputs through affine registration, yet residual non-linear deformations inevitably persist due to patient motion, tissue deformation, and imaging artifacts [19]. These misalignments substantially degrade reconstruction fidelity [20, 21] and diagnostic utility [22], particularly affecting fine anatomical structures.

- Recovering modality-specific structures presents unresolved difficulties. Even assuming perfect alignment, clinically critical features appearing exclusively in target modalities, or severely degraded in available inputs, challenge current approaches. While existing work predominantly exploits structural similarity across modalities, it largely ignores these diagnostically valuable yet spatially limited distinct structures [23]. This oversight manifests in two failure modes: alignment-based methods generate spurious signals from interfering reference structures, while global similarity approaches become unreliable when processing heavily degraded target regions. Consequently, reconstruction of distinct structures consistently suffers from vanishing details or pathological blurring, compromising diagnostic value.

### 1.2.2 Alzheimer's Disease Diagnosis: Integration of Image and Non-Image Data

Alzheimer's Disease (AD) diagnosis exemplifies the critical need for integrating heterogeneous non-imaging modalities with neuroimaging data. Unlike spatially aligned multimodal images, these data streams lack inherent structural correspondence, instead collectively reflecting complex disease pathophysiology through task-specific relationships. This application presents three fundamental challenges for flexible multimodal integration:

- The assumption of fixed modality sets contradicts clinical reality. Current methods require perfectly matched multimodal training data, yet real-world AD datasets exhibit severe modality imbalance. As the assumed modality count increases, the available complete samples diminish exponentially, creating data scarcity for model training. Simultaneously, the combinatorial space of possible modality subsets expands beyond what any fixed-model can accommodate during inference. Critically, valuable partial-modality samples remain unused, while novel biomarkers emerging outside predefined modalities cannot be incorporated.

- Architectural inflexibility limits practical deployment. Existing approaches, whether parallel processing via shared attention layers [2, 24] or serial cross-attention [25], fail to address combinatorial scalability. Shared self-attention incurs quadratic computational growth with added modalities, while cross-attention designs lack permutation invariance to input order. More fundamentally, dependency on modality-specific components (dedicated backbones [2], transformers [25], or FFNs [24]) introduces parameter ineffi-

ciency that exacerbates overfitting in data-scarce medical contexts. This precludes lightweight adaptation to arbitrary modality combinations encountered in clinical workflows.

- Inappropriate alignment strategies obscure diagnostically relevant information. Prevailing cross-modal alignment techniques force feature uniformity across modalities under the assumption of underlying semantic equivalence [2, 24]. While enhancing robustness to missing data in vision/language domains, this strategy falls short when applied to medical contexts. In contrast to typical modalities such as vision, language, and audio, medical modalities lack inherent semantic connections. Instead, their relationships are task-dependent and frequently associated with specific diseases of interest. Consequently, enforcing direct cross-modal alignment may impede the discovery of unique yet complementary information from various modalities, which is crucial for differential diagnosis.

### 1.2.3 ARDS Risk Monitoring: Integration of Asynchronous Modalities

In the third application, asynchronous multimodal learning for dynamic Acute Respiratory Distress Syndrome (ARDS) risk prediction is investigated, which naturally requires flexible integration of modalities. As in each prediction interval, the modality could be missing or repeatedly appears at different time points. The challenges in this application are:

- Existing ARDS risk models operate under static paradigms, relying either on fixed time windows (e.g., first 24h of ICU admission) or isolated imaging

assessments, which critically misalign with clinical realities during rapid deterioration. While chest X-rays (CXRs) offer irreplaceable diagnostic specificity for pulmonary infiltrates, their sparse acquisition (typically 1–2/day) and delayed manifestation of pathology complement rather than replace high-frequency physiological trends from vital signs (VS) and laboratory results (LAB). This asynchronicity creates a temporal disconnect: VS/LAB streams capture minute-to-minute deterioration, whereas CXRs provide definitive but lagging structural confirmation. It remains unexplored to construct a system that dynamically integrates these asynchronous modalities to deliver the iterative risk reassessments clinicians need during actionable windows.

- Existing approaches fail to reconcile the temporal and semantic heterogeneities inherent in ICU data. Late fusion strategies neglect cross-modal dependencies between sparse imaging findings and continuous physiological trends, relationships critical for early warning. Meanwhile, deep time series models assumes temporal synchrony [26], which force arbitrary resampling to process irregularly sampled data, corrupting latent biological signatures. It remains challenging on how to encode these asynchronous, irregular multimodal time series into a unified latent space while preserving observational timestamps and signal provenance, while efficiently modeling temporal and semantic dependencies.

- Continuous risk monitoring imposes computational burdens absent in one-pass prediction. Simultaneously, severe class imbalance is amplified during sequential inference, drowning subtle risk signals in majority-class noise.

These constraints directly conflict with clinical needs: real-time predictions must complete within clinician workflow cycles while maintaining sensitivity to rare events. Current architectures either sacrifice temporal resolution or degrade into heuristic approximations, compromising prognostic fidelity.

## 1.3 Contributions

### 1.3.1 Multimodal MRI Synthesis: Integration of Multi-Modal Images

- A unified framework bridges the fragmentation between cross-modality synthesis (CMS) and multi-contrast super-resolution (MCSR) through fine-grained difference learning. This approach establishes shared anatomical coordinate systems for both tasks, eliminating modality-specific processing pipelines. By formulating CMS as a generalized case of MCSR with extreme undersampling, the framework enables consistent handling of diverse modality availability scenarios.

- Spatial misalignment challenges are addressed through multi-scale deformable convolution modules integrated directly into the super-resolution pathway. These learnable warping operations perform feature-level alignment of degraded target inputs to reference modalities, compensating for non-linear anatomical shifts without separate registration networks or affine transformation assumptions.

- Modality-specific structural recovery is enhanced via synergistic compo-

nents: (a) A difference projection discriminator explicitly isolates modality-exclusive signatures; (b) Distinction-aware feature regularization preserves clinically critical patterns during reconstruction; (c) Incremental feature modulation dynamically leverages available target inputs at varying degradation levels. This triad mitigates blurring artifacts in diagnostically decisive regions.

- Comprehensive validation demonstrates superiority in perceptual quality across unprecedented downsampling ranges (2×–16×). The framework exhibits consistent performance where existing methods fail, particularly for fine anatomical structures at high degradation levels.

### 1.3.2 Alzheimer's Disease Diagnosis: Integration of Image and Non-Image Data

- The AnyMod architecture enables training and inference on arbitrary modality combinations through representation-task decoupled alignment. Representation-level alignment establishes a unified metric space where N modalities discover complementary features, while task-level alignment maps heterogeneous combinations to class-invariant prototypes. This dual mechanism preserves modality-specific semantics while ensuring combination robustness.

- Modality-agnostic Transformer projectors replace dedicated encoders, reducing parameters while permitting seamless adaptation to new modalities. These tunable projectors map raw inputs to the unified space using shared self-attention mechanisms, eliminating modality-specific components.

- Dynamic token clustering bounds computational complexity for variable-length inputs. By projecting multimodal tokens onto fixed-dimensional task factors prior to fusion, the framework maintains stable inference costs regardless of combination cardinality. This design supports real-time operation while processing exponentially growing modality subsets.

- The experimental outcomes demonstrate that, in this novel framework for learning combinations of modalities, the proposed approach allows a single unified model to gain increasing advantages over models trained individually on each combination as the number of modalities grows. Additionally, the model adapted to new modalities can effectively manage previously unseen combinations without requiring additional training, highlighting the strong scaling potential of the presented model.

### 1.3.3 ARDS Risk Monitoring: Integration of Asynchronous Modalities

- A modality-wise encoding strategy is introduced to handle asynchronous data streams. For each modality, a dedicated Transformer layer processes its constituent variables, incorporating attention masking to accommodate arbitrary missing measurements without imputation. Temporal alignment is achieved through a sliding-window algorithm that identifies shared timestamps across co-sampled variables, avoiding distortion from resampling. Furthermore, adaptive positional encodings, conditioned jointly on modality type and relative observation time, are proposed to preserve temporal-semantic relationships, demonstrating significant improvements over stan-

dard encoding schemes in asynchronous learning. Complementing this, the Staged Temporal-Modal (STM) fusion module addresses architectural limitations through a hierarchical transformer design that decouples cross-modal interaction from temporal processing.

- Progressive Context Memory (PCM) is proposed to enable clinically feasible long-range dependency modeling for continuous risk monitoring. By incrementally compressing patient history into compact memory states, PCM maintains temporal awareness over extended ICU stays while reducing computational complexity. Meanwhile, tailor training strategies are proposed, with balanced sampling significantly enhancing detection sensitivity for early deterioration signals, and late batching resolving optimization instability in variable-length temporal modeling.

- An innovative continuous risk reassessment approach, aligned with ICU workflows, is presented. This method replaces traditional static prediction intervals by providing hourly updates on the likelihood of ARDS and the urgency of time-to-onset. The framework integrates infrequent chest X-ray results with high-frequency physiological data streams, converting delayed imaging into a valuable prognostic reference point. The model achieved outstanding performance across $24h/48h$ pre-onset windows (AUROC $0.91/0.87$), outperforming previous techniques (AUROC $0.78$-$0.85$) in a more demanding scenario. Particularly within the critical <6h pre-onset window, our system successfully detects $91\%$ of ARDS cases, achieving an AUROC of $0.94$. Additionally, it provides actionable risk categorization: cohorts identified as high-risk (thresholds ranging from $0.5$ to $0.7$) show a

20-fold increase in ARDS occurrence. Furthermore, our precise quantification (with MAE below 0.6) directly aids in optimizing resource allocation priorities.

## 1.4 Thesis Organization

Following this introduction, Chapter 2 provides a comprehensive literature review: the first section examines methodological approaches to flexible modality integration in medical multimodal learning, while the second section focuses specifically on the three target applications—multimodal MRI synthesis, Alzheimer's disease diagnosis, and ARDS risk monitoring. This thesis adopts a thesis-by-publication structure, with Chapters 3, 4, and 5 presenting the three core publications respectively. Each application chapter begins by establishing the problem background and research gaps. The method section details problem formulations and algorithmic innovations, while experiment-specific elements, including data sources, evaluation protocols, and implementation details, are consolidated in dedicated experiment settings sections. Results and their critical analysis follow in subsequent discussion sections. Finally, Chapter 6 synthesizes overall conclusions and outlines future work, encompassing refinements to the proposed frameworks and broader prospects for flexible modality integration in medical AI.

# Chapter 2

# Literature Review

## 2.1 Flexible Modality Integration in Medical Multi-Modal Learning

Multimodal learning leverages complementary information across diverse data sources to enhance performance in medical tasks. A core challenge involves developing fusion strategies that integrate modalities effectively. Early approaches, such as feature concatenation [27], evolved into tensor fusion [28] and low-rank factorization [29] to mitigate computational complexity. However, these methods assume complete modality availability during inference, limiting their applicability in real-world scenarios where data may be partially missing. Consequently, a critical focus in multimodal learning is the development of models that maintain effectiveness even when handling incomplete modality information [30, 31].

Historical literature addresses flexible input combinations through distinct paradigms. First, the fixed-modality setting requires all modalities during training but allows

one or more to be absent during inference [15]. Second, learning on incomplete multimodal data [32] permits missing modalities in both training and inference phases, albeit within a predefined maximum modality count. Third, open-modality frameworks [33, 34] eliminate fixed constraints, enabling dynamic integration of novel or scarce modalities.

The overarching aim is to maintain performance under missing modalities comparable to full-modality settings, which is a theoretically unattainable ideal. Pragmatically, robust models should: (1) outperform single-modality baselines by leveraging cross-modal synergies; (2) surpass models trained only on complete subsets when using all available data (including partial samples); and (3) generalize better to unseen modality combinations than approaches without modality-completion augmentation.

Two dominant paradigms address missing modalities. The first, imputation, reconstructs absent data before fusion. Naive deletion of incomplete samples wastes information; generative imputation often introduces artifacts and computational overhead [35]. For instance, Zhang et al. imputed missing electronic health records (EHR) by modeling patient-modality graphs via task-guided kernels and graph neural networks [35]. However, such inverse problems are ill-posed and may propagate errors. The second paradigm, flexible architectures, avoids reconstruction by dynamically adapting to input combinations. Subgrouping strategies train separate models for distinct modality subsets, framing learning as multi-task optimization [36]. Graph-based methods, like hypergraphs linking instances with identical missing patterns [32], facilitate information exchange between subgroups but scale poorly to large datasets. Latent space alignment techniques project modalities into a shared embedding, enabling arithmetic operations

for fusion [37, 38]. Transformers [39] further generalize this via self-attention, processing any modality tokenized into sequences (e.g., images [40], time-series). While promising, Ma et al. [41] revealed that Transformers' robustness varies significantly with fusion strategies and datasets, lacking a universal solution.

Research bifurcates along modality types due to task-specific demands. Multimodal images primarily target dense prediction tasks like segmentation, reconstruction, or generation [15]. These require pixel-level alignment and typically assume full-modality training, emphasizing spatial consistency. Conversely, image and non-image fusion centers on high-level decisions such as classification or prognosis [35]. Here, modalities are often highly heterogeneous, numerous, and imbalanced (e.g., scarce lab tests vs. abundant imaging). Challenges include integrating cross-modal semantics and managing combinatorial complexity during missing-data scenarios [33], necessitating architectures that scale efficiently amid modality diversity.

### 2.1.1 Multi-Modal Medical Images

Existing research on missing modality problems in medical imaging, particularly in MRI, underscores the significance and complexity of this challenge. Azad et al. [15] highlights that missing modalities commonly arise due to factors such as scan duration, varying clinical protocols across hospitals, limitations of imaging equipment, or patient allergies to contrast agents. These missing sequences result in the loss of complementary information that is often hard to fully recover from remaining modalities, , raising the question: Can this missing information be effectively reconstructed or compensated?

Notably, Gijs and Marleen [42] explored this intriguing question: although synthetic data does not introduce new information, why does it still enhance performance? In their experiments, they observed that training and testing using a single synthetic sequence yielded accuracy comparable to training on the original dataset without the synthetic sequence. Substituting the synthetic data with zeros produced similar outcomes. The process of imputing data enables fixed-architecture models to adapt more effectively to incomplete modalities, potentially increasing the effective size of the training data, which aids in improving the generalization capabilities of deep learning models. However, under identical dataset conditions, synthetic data did not contribute additional benefits. Furthermore, they discovered that restricted Boltzmann machines (RBMs) possess a practical advantage over neural networks. RBMs learn a joint probability distribution that can be utilized to predict any missing sequence, whereas neural networks are explicitly trained to predict one sequence given others, requiring a separate network for each sequence. Additionally, more complex classifiers tend to extract richer information from the original data and are therefore less likely to gain significant advantages from synthetic data.

To address this issue, a potential approach involves incorporating additional information during the inference process. An earlier study on PET/MRI attenuation correction [43] proposed a similar concept. To estimate attenuation by synthesizing CT images from MRI, the researchers utilized 17 paired MRI-CT datasets. They integrated both local and global matching techniques to leverage this data effectively. For the local matching, they established precise mappings at specific anatomical landmarks following image registration. Subsequently, they employed a patch similarity matching method in conjunction with Gaussian process regres-

sion for transformation. In terms of global matching, they aligned the patient's MRI and all reference MRIs to an atlas, applied the resulting deformation fields to the available CT scans, and then computed the average of these CTs as the estimated patient CT. To combine the local and global matching results, they used a simple weighted averaging technique with manually adjusted weighting factors.

In subsequent developments, deep learning techniques emerged to address this issue more effectively. Approaches such as learning shared latent representations across modalities [37, 44, 45, 46, 47, 48] have become prominent. For example, [37] proposed a model embedding each modality separately into a common latent space and computing first and second moments to facilitate segmentation, thus avoiding the need for multiple imputation models. However, such methods are often limited to simple statistical measures and do not explicitly model inter-modality correlations. More advanced techniques, like the dual-path framework introduced by Wang et al. [48], train dedicated models for each possible missing-modality scenario and share knowledge via a latent space, but tend to dilute modality-specific information.

Recent innovations have concentrated on explicitly modeling the relationships between modalities. Zhang et al. [49] utilized inter-modality transformers with alignment loss, and incorporated Bernoulli variables during training to simulate missing modalities, encouraging the model to learn robust, modality-invariant features. Meanwhile, approaches based on mutual information maximization [50, 51, 52] aim to retain as much information as possible by optimizing similarity metrics between available modalities during training. For instance, [52] introduced latent correlation representations derived via linear combinations in a learned latent space, allowing the estimation of missing modalities from available ones.

Another promising route involves generative adversarial networks (GANs) [53, 54, 55, 56, 57], which synthesize missing images conditioned on available modalities. Specifically, models like MM-GAN [53] and the recent unified synthesis framework by Zhang et al. [58] leverage conditional GANs to generate missing modalities with high anatomical consistency. However, these models often require extensive training data of complete modalities, which can be scarce in medical contexts, and may introduce artifacts that negatively affect downstream tasks. Additionally, the computational overhead associated with GAN-based synthesis can be substantial, limiting practical deployment.

In contrast to purely generative approaches, techniques like knowledge distillation [59, 60, 61] transfer knowledge from models trained on full data to more robust models that perform well even when some modalities are missing. For example, [60] employed a teacher-student framework where a teacher network trained on complete modalities guides a student network trained solely on partial data, with the guidance enhanced by hierarchical adversarial training to mimic intermediate features across multiple scales.

Overall, current research indicates a shift from simple imputation towards more sophisticated modeling of inter-modality relationships and robustness to missing data, leveraging latent correlation representations, attention mechanisms, and adversarial training. Nevertheless, challenges remain in designing models that are scalable, generalizable to unseen modality combinations, and capable of preserving critical diagnostic information with minimal computational cost.

### 2.1.2 Medical Image and Mon-Image Data

Existing research that primarily focuses on multimodal images generally addresses dense prediction tasks such as image synthesis and segmentation. In routine clinical practice, a single patient's visit often generates diverse types of digital data across multiple modalities, including image data, such as pathology slides, radiology scans, and camera images, and non-image data, such as laboratory test results and electronic health records. This heterogeneity in data sources offers complementary perspectives of the same patient, thereby enabling more comprehensive support for various clinical decisions, including disease diagnosis and prognosis.

The process of multimodal learning generally consists of three key components: single-modal preprocessing and feature extraction, cross-modal feature integration, and prediction modeling. The architecture of the feature extraction module is determined by the specific modality type. For instance, 2D/3D Convolutional Neural Networks (CNNs) and BERT are commonly used for extracting features from images and text, respectively. Additionally, Transformers can serve as a unified framework for feature extraction by converting elements from images, text, and tabular data into token representations, as demonstrated in recent studies [62]. For preprocessing, the methods varies in different detailed modalities. For pathology images, which are large 3-channel 2D images, the first step is to define the ROIs from the whole slide image, which could be manually annotated by experts, segmented by pre-trained segmentation model or selected by pixel density. Differently, radiology images typically go through skull-stripping [63], affine registration [63], foreground extraction [14], lesion segmentation [63, 64], then resized or cropped to a uniformed size, and then the intensities are standard-

ized. For 3D images, they could be converted to 2D through maximum intensity projection [65] and selection of representative slices [64, 66]. For functional MRI [67, 68], which benefits autism spectrum disorder and Alzheimer's disease. It is normally divided into regions using templates and then the correlation coefficient between regions is calculated to form functional connectivity matrix. For structured data, the categorical values will go through one-hot encoding or soft one-hot encoding [69]. For the high-dimensional genomic data, preprocessing includes feature selection with highest variance as the most informative ones. Missing data is a common problem in tabular data, and it can be alleviated at the preprocessing stage. Attributes with high missing rate are normally directly discarded, while the others are imputed by average, mode or K-nearest neighbors [65]. The missing status was also recorded as input features [70, 71].

The integration of different modalities can generally be categorized into decision-level and feature-level fusion. Decision-level approaches involve selecting (via voting or meta-classifiers) or weighting predictions from multiple unimodal models [63, 65, 72, 73, 74, 75]. These techniques demonstrate flexibility when dealing with missing modalities and, in some cases, outperform simple feature-level fusion methods [72, 75]. However, they lack interaction among heterogeneous modalities before consolidating them into a final prediction format. In contrast, feature-level fusion, when appropriately designed, holds the potential to uncover and leverage the intricate relationships between heterogeneous features for prediction tasks.

For feature-level fusion, the integration of different modalities typically relies on operations such as addition, multiplication, and concatenation. Among these, the most straightforward approaches—concatenation, element-wise addition, and element-wise multiplication—are widely adopted in early multimodal

studies [63, 65, 73, 76, 77, 78]. These three methods were evaluated in the context of breast cancer diagnosis [65], with results indicating no substantial differences among them. However, concatenation may suffer from the dominance of high-dimensional features, which can be mitigated by enhancing low-dimensional features [78]. Notably, an earlier study [14] addressed the issue of missing modalities. To maximize the use of available samples, the authors proposed a three-stage pipeline, where all stages are guided by classification tasks. In the first stage, unimodal deep neural networks (DNNs) are trained separately for each modality, allowing the utilization of all samples regardless of missing data. The second stage involves fusing pairs of modalities, leveraging samples with only one missing modality. Finally, in the third stage, features obtained from the second stage are combined, optimized using the smallest subset of samples with complete modalities. The flexibility of this approach stems from enumerating all possible combinations and enforcing classification tasks in each scenario. Nevertheless, there are certain limitations: 1) the computational cost increases proportionally with the number of modalities; 2) concatenation-based fusion is overly simplistic, particularly when dealing with a larger number of modalities; 3) separate supervision for each modality may align features from different sources toward a common task, akin to challenges observed in decision-level fusion.

These basic operations do not add extra network parameters on their own. However, element-wise operations require uniform dimensions across different modalities and assume well-aligned elements within these modalities. Additionally, concatenation may result in long feature vectors that are computationally expensive to process and prone to overfitting. To overcome the limitations of element-wise operations and capture complex relationships among heterogeneous

modalities, several advanced techniques have been introduced. Tensor-based fusion utilizes the outer product of augmented feature vectors from different modalities [28]. For modeling higher-order interactions, another tensor-based approach [79] performs P times self-outer product on concatenated and augmented multimodal features, generating a P-dimensional tensor that captures interactions up to order P+1. The computational cost is significantly reduced using low-rank tensor networks. Furthermore, to improve expressiveness, attention weights can be incorporated into both multiplication-based and addition-based methods, enabling interactions either between or within modalities. In the health domain, Kronecker product-based tensor fusion combined with gated-attention layers has been employed to integrate genomic data, cell graphs, and pathology images [80]. Another tensor-based method [75] applies outer products for both inter-modal and intra-modal interactions. Orthogonal regularization is applied to the learned features from four modalities to promote diversity and minimize redundancy in tensor-based fusion [81].

Besides, attention-based techniques represent a significant trend in multimodal fusion. For instance, attention scores can be allocated to each modality prior to concatenation, as demonstrated in [66]. In another study [82], pooled image features and tabular data were concatenated to derive element-wise attention scores for all features. By treating image features as the primary predictor, non-image features were utilized to provide channel-wise attention for image features, as described in [83]. A similar approach was adopted in [71], where tabular data modulated the 3D image feature map through affine transformation. Self-attention mechanisms and transformer architectures have become prominent in recent multimodal studies. In [84], tabular features were expanded to align with 3D image

features before concatenation for self-attention computation. The resulting outputs were subsequently flattened and concatenated once more with the original tabular features to produce the final prediction. Cross-attention derived from WSI and genomic features was employed to weigh WSI features, as shown in [80]. Symmetric cross-attention was also utilized to ensure equal treatment of both modalities, as seen in [69, 76]. With regard to textual modality, transformer-based language models have been increasingly applied in healthcare domains. BERT [85] has been used to extract text features, followed by fine-tuning after incorporating image tokens [86]. Moreover, visual-text transformers pretrained on general datasets were further fine-tuned using radiology images and their corresponding reports, as conducted in [87].[87] using radiology images and the corresponding reports.

Multimodal fusion has also been integrated into graph neural networks. In [70], every voxel of the 3D image features, which represents a specific region in the original 3D image, is treated as a node within the graph. The node feature is formed by combining multimodal features. By integrating non-image features into the graph propagation process, this method refines modality interactions at the regional level rather than the overall image level. This approach has also been applied to population-level graphs. Methods based on population graphs [67, 68] utilize patient features from both the labeled training set and the unlabeled test set to construct a large graph that connects each patient. Here, image features serve as nodes, while the correlation of tabular features is used as edge weights to establish a graph convolutional network. During the learning process, non-image data influences the extent to which information from image data can be transferred between patients' image data.

Recent studies have explored generative approaches to compensate for missing

modalities, either at the instance level or through embedding reconstruction. For example, Ma et al. [30] introduced a Bayesian meta-learning framework to reconstruct features of absent modalities. Hayat, Geras, and Shamout [88] employed an LSTM layer to create a representative vector suitable for general scenarios. Additionally, Zhang et al. [35] suggested leveraging auxiliary information in the latent space for imputation. However, these techniques often rely on prior knowledge or assume similarity among different modalities. There is also concern that methods depending on generated representations may lack robustness [89]. An alternative strategy involves separating shared and complementary information across modalities, utilizing the shared component for reconstruction or downstream tasks [90, 91]. Despite this, many existing works primarily address modalities with substantial shared information, such as using four MRI modalities for brain tumor segmentation. The challenge of managing missing modalities in highly heterogeneous contexts, like the integration of EHR and medical imaging, remains largely unresolved.

## 2.2 Multi-Modal Learning in Targeted Applications

### 2.2.1 Multi-Modal MRI synthesis

#### 2.2.1.1 Cross-modality Synthesis

Cross-Modality Synthesis (CMS) is a key task in the field of medical image translation. Initially, methods treated MRI translation as a regression problem. For instance, random forest [92] was utilized to perform nonlinear regression in feature space for predicting target modality intensity. In another study [93], the authors

suggested modeling a shared latent representation to accommodate multiple input modalities. While regression-based techniques focus on pixel-level accuracy, they often produce overly blurred results. On the other hand, generative approaches effectively capture the distribution of target images, maintaining finer details in synthesized outputs. Consequently, conditional GANs, exemplified by Pix2Pix [94], have emerged as the dominant solution for CMS. In [95], an edge-aware conditional GAN was introduced. This method highlights edges by incorporating an edge detector post-generator, where detected edges are fused with the generated image to serve as attention cues for the discriminator. Furthermore, the authors proposed a sample-adaptive GAN model [96], which enhances local spatial learning through dual pathways, allowing flexible adjustments for CMS tasks. In [53], a multi-input multi-output MR pulse sequence synthesizer was developed within the Pix2Pix framework, leveraging PatchGAN to ensure precise local feature synthesis. A comparable strategy was adopted in [97], where the translation module was integrated into segmentation and registration networks to improve overall performance in MRI translation.

The utilization of unpaired data has garnered significant attention in the field of medical image translation. Unsupervised CycleGAN, as introduced by Zhu et al. [98], demonstrates this approach by enabling cross-domain image translation without relying on paired datasets through the enforcement of cycle consistency. Despite advancements, some studies have integrated CycleGAN with unpaired data [99, 100], raising concerns regarding its suitability for medical image translation due to challenges such as generating multiple potential solutions [18] and vulnerability to severe systematic misalignments [21]. A comparative analysis between conditional GAN (cGAN) and CycleGAN in MRI-CT translation [101] revealed

that while CycleGAN can produce realistic CT images, it performs less effectively than cGAN, particularly in areas with weak MR signals, such as bone/air interfaces. This highlights the significance of incorporating a conditioning mechanism when employing generative models.

An alternative strategy is presented in MT-Net [102], which bypasses the use of CycleGAN by leveraging unpaired data to pre-train an edge-aware Vision Transformer (ViT) encoder. This method enhances performance even when limited paired training data is available. Another research avenue addresses misaligned data, where matched pairs may not be perfectly aligned during preprocessing. RegGAN [18] treats misaligned targets as noisy labels and adapts a registration network to accommodate the distribution of misaligned noise. In contrast, Honkamaa et al. [21] emphasized the concept of deformation equivariance, disentangling rigid misalignment from elastic deformation to mitigate systematic errors learned by the translation network.

Furthermore, certain studies explore feature modulation and style transfer within the context of cross-modality MRI synthesis. Qin et al. [100] integrate pix2pix and CycleGAN, utilizing Adaptive Instance Normalization (AdaIN) to facilitate style transfer. Zhan et al. [103] advocate for explicitly representing the target style within a style transfer network by incorporating an auxiliary GAN to produce a pseudo-target as the style reference. Hu et al. [104] introduce neural architecture search into MRI translation, proposing a GAN-based perceptual search loss to achieve a balance between performance and model complexity.

#### 2.2.1.2 Multi-Contrast MRI Super-Resolution

Multi-contrast MRI super-resolution (MCSR) typically surpasses single-contrast methods and has become an increasingly significant research area. In contrast to single-contrast super-resolution, the primary challenge in multi-contrast scenarios lies in effectively modeling the relationships between different modalities, with numerous approaches emphasizing cross-modal similarities. Early traditional techniques focused on modeling contrast-invariant representations, such as image gradients [105], local covariance [106], and non-local similarity graphs [107]. These representations were subsequently integrated into optimization-based frameworks [105, 107] or interpolation filters [106, 108]. Deep learning-based approaches have demonstrated superior outcomes compared to conventional methods, benefiting from the effective registration of multi-modal inputs during preprocessing [109, 110].

Following methods have concentrated on sophisticated attention mechanisms for capturing both local and global similarities. These include spatial-channel attention mechanisms [17, 111] as well as cross-attention Transformers [112, 113, 114, 115]. Drawing inspiration from reference-based super-resolution techniques [116], global matching strategies have been incorporated to explore global similarity within the reference modality. Such approaches are frequently augmented with Swin Transformer for unimodal feature extraction [117]. A recent study [118] has further refined this concept by integrating patch-based self-attention modules with channel-based ones [119] for unimodal feature extraction. By reducing the size of the search window under spatial alignment constraints, these methods address issues related to computational complexity and overfitting, which arise due

to the inherent differences between reference-based SR and MCSR [120]. Moreover, deformable attention in vision transformers [121] has been utilized in the backbone architecture to enhance computational efficiency [120].

Notably, the challenge of identifying fine-grained differences has been highlighted in prior studies, encompassing issues such as spatial misalignment [20] and structural variations [107]. In [20], the researchers addressed subtle spatial misalignments arising from imperfect pre-registration, a prevalent issue in practical applications that is often overlooked. To tackle this, they proposed an explicit modeling approach using an auxiliary registration network to generate a deformed reference. Multi-modal fusion was subsequently achieved via concatenation after establishing refined spatial correspondence. On the other hand, in [107], the authors suggested explicitly constructing a cross-modal correlation map by leveraging patch-level global similarity graphs. This map was then utilized to assign weights to information from various contrasts during interpolation. While the correlation map effectively captures coarse structural distinctions while mitigating noise, it requires significant computational resources. Additionally, the straightforward weighted interpolation method proves insufficient for achieving precise restoration.

Although generative methods share similar objectives with CMS, they have not yet been extensively adopted in MCSR. While direct optimization of pixel-wise loss can achieve higher quantitative outcomes compared to adversarial loss, it often results in overly smoothed outputs [122]. To tackle this issue, the study in [123] was among the first to apply conditional GANs in MCSR by integrating priors across various frequency levels. Recently, there has been growing interest in leveraging generative approaches for this task. For instance, TransMRSR [124]

combined a generative prior into the process, utilizing an unconditional generative model pretrained on extensive target-domain images as a static decoder. Nevertheless, this strategy might introduce additional uncertainty due to the nature of unconditional GAN generation. On the other hand, DisC-Diff [125] incorporated the diffusion model, employing a multi-input architecture with feature disentanglement to approximate the reverse distribution during each iteration. The advantages of diffusion models include stable training dynamics and the ability to estimate uncertainty; however, these models are resource-intensive and encounter challenges in achieving precise conditioning for fine-grained generation.

### 2.2.2 Alzheimer's Disease Diagnosis

Alzheimer's disease, a neurodegenerative disorder marked by amyloid-β plaques, tau tangles, and progressive cognitive decline, demands early and precise diagnostic strategies to enable timely intervention. Neuroimaging modalities such as structural MRI, amyloid-PET, and FDG-PET provide complementary insights into distinct pathological processes. While MRI identifies structural degeneration in regions like the hippocampus, PET imaging reveals molecular-level dysfunction years before symptomatic onset. Pioneering works in multimodal AD analysis concentrated on predetermined modality sets.

T1-weighted MRI and FDG-PET imaging data are the most commonly considered modalities in multimodal AD analysis. Liu et al. [126] developed a cascaded CNN framework for Alzheimer's disease diagnosis that hierarchically processes MRI and PET scans. The architecture employs parallel 3D-CNNs to extract local patch-level features from each modality, followed by a 2D-CNN that integrates

these multimodal representations into a unified feature space. The fused features are then fed into a softmax classifier for final diagnosis. This hierarchical design simultaneously captures local biomarker patterns while modeling their global cross-modal correlations, demonstrating improved diagnostic accuracy compared to single-modality approaches. Huang et al. [127] leveraged both T1-weighted MRI and FDG-PET imaging data by concatenating their features in a multi-modal 3D CNN network. The proposed approach demonstrates that high diagnostic accuracy can be achieved without the need for prior segmentation of brain structures, specifically focusing on the hippocampal region, which is highly relevant in AD progression. The results highlight the importance of integrating complementary information from different imaging modalities, as the combined use of MRI and PET significantly improves classification performance compared to single modalities. Shi et al. [128] proposed to explicitly model the coupled interactions at both the feature level and modality level, the coupled feature representation captures the inherent relationships among features from different brain regions, which are known to be anatomically and functionally interconnected. Shi et al. [129] proposed to construct adaptive similarity matrix that evolves with feature selection, where similarity learning is shared across modalities, and the local structure preservation enhances discriminative power.

Liu et al. [130] constructed graph convolutional network from structural MRI and functional MRI, demontrating that strategically combining structural discriminative features with functional connectivity patterns can significantly improve neuropsychiatric disorder classification, while also providing interpretable biomarkers of disease-related brain network alterations. Chen et al. [131] employs orthogonal projections to map multi-modal neuroimages into a discriminative latent

space, utilizing adaptive feature weighting to prioritize diagnostic biomarkers and joint graph learning to capture cross-modal correlations.

In addition to sMRI and PET, other heterogeneous types of data, such as cerebrospinal fluid (CSF) biomarkers and genetic information, have also been integrated into multi-modal learning approaches. Suk et al. [132] proposed a hierarchical deep architecture that recursively eliminates uninformative features through sparse multi-task learning, where optimal regression coefficients quantify feature importance and serve as weighting factors for subsequent hierarchies. The method further incorporates class distribution characteristics by employing clustering-induced subclass labels as target responses in the weighted sparse regression framework. Tong et al. [133] introduced a nonlinear graph fusion process to combine the similarity graphs from different modalities, generating a unified graph for structural MRI, FDG-PET, CSF biomarkers and genetic information. El-Sappagh et al. [134] further involved cognitive test scores, considering five different types of modalities in their study. They proposed a stacked CNN-BiLSTM architecture that processes each time-series modality separately to extract both local features (via CNN) and long-term temporal dependencies (via bidirectional LSTM) and a late fusion strategy of all learned features for joint prediction of multiclass progression status and four cognitive score regression tasks. Tu et al. [11] integrated MRI scans, clinical data (MMSE, CDR), APOE genotype, and demographic data. They proposed geometric algebra-based feature extension to enhance low-dimensional clinical/biological data, and influence degree-based feature filtration to remove non-discriminative features. Qiu et al. [5] conducted a large-scale multi-centre study involving neuro-imaging data, clinical assessments, functional evaluations, demographic data, and APOE genotype. They proposed a

multi-stage diagnostic pipeline including classifing NC vs MCI vs dementia, and then distinguishing AD from other dementia. Their method achieved $96.2\%$ accuracy for AD diagnosis, which exceeds clinician diagnosis accuracy and identifies biologically plausible biomarkers. Han et al. [135] integrated multiple biological modalities including mRNA expression data that provides gene activity information, DNA methylation data that offers epigenetic regulation insights, and miRNA expression data that reveals post-transcriptional regulatory mechanisms. They proposed a multimodal dynamics method introduces an innovative two-stage fusion approach that dynamically assesses both feature-level and modality-level informativeness. At the feature level, a sparse gating mechanism with L1 regularization identifies and weights the most informative molecular features for each sample. At the modality level, true class probability estimation evaluates the confidence and reliability of each omics data type for every individual case. These dynamic assessments are then integrated through a nested fusion architecture that combines the weighted features and confidence-adjusted modalities to produce more trustworthy predictions. The framework's key innovation lies in its sample-specific adaptive weighting of both individual biomarkers and entire modalities. Zhou et al. [13] leveraged deep canonical correlation analysis to maximize the correlation among structural MRI, function connectivity data, and SNPs in the latent space. This method identifies modality-specific discriminative features and highlights disease-relevant imaging-genetic relationships, which provides interpretable attention maps for clinical insights. Zhang et al. [136] introduced cross-attention in modeling inter-modality relationships among imaging data and CSF biomarkers, with modality alignment loss to reduce feature space discrepancy. This method achieved superior performance over traditional concatenation-based

fusion methods by effectively modeling cross-modal relationships through its attention mechanism. Kang et al. [137] adopted a dual-branch architecture including a vision branch processes sMRI scans via CNN and a attribute branch encodes clinical data. The prompt learning mechanism learns modality-specific prompts to bridge visual-clinical gaps, and generates dynamic prompts conditioned on input attributes. The framework demonstrates innovative use of prompt learning for neuro-clinical data fusion, showing particular strength in identifying high-risk pMCI cases. Qiu et al. [138] considered not only sMRI, fMRI but also Diffusion Tensor Imaging (DTI) for white matter integrity, and Amyloid PET for pathological protein deposition. They introduced an innovative three-level fusion architecture for multimodal integration. At the global level, a self-adaptive Transformer captures cross-modal relationships across the entire brain volume. The local level employs specialized convolutions to identify regional associations between modalities, while the latent-space learning module enhances representations through outer product operations to discover deeper multimodal interactions. The network further incorporates a disease-induced region-aware learning module that uses gradient weighting to highlight Alzheimer's-relevant brain regions, providing both improved performance and some degree of interpretability by identifying important pathological features.

Most of existing work on multi-modal AD analysis assumes full availability of all modaltiy data. However, missing modality is a common challenge in practical scenarios especially for this task with numerous and imbalanced modalities. However, most AD-related work only solves missing PET problem in T1 MRI and PET fusion. Pan et al. [139] addresses the common clinical challenge of missing PET data by a CycleGAN-based framework to synthesize missing PET

scans from available MRI, and demonstrated advantages over other imputation methods. Similarly, Jin et al. [140] also addresses missing PET data through a hybrid deep learning framework that integrates image synthesis and classification, it involves three different stages that pretrains on complete paired MRI-PET data, uses classification features to guide PET synthesis, and fine-tunes the classifier on synthesized PET images. Liu et al. [141] proposed a auto-encoder based multi-view completion framework. First, an auto-encoder network maps available complete modality data to a latent space representation, reducing noise and dimensionality while preserving structural information through graph regularization. Then, it uses these learned representations to complete missing PET data in kernel space rather than raw feature space, maintaining complex high-dimensional relationships. The method incorporates Hilbert-Schmidt Independence Criterion (HSIC) constraints to preserve inherent cross-modal associations during completion. Finally, a kernel-based multi-view approach integrates the completed data to produce robust common representations for diagnosis. Liu et al. [142] also addresses image synthesis and clinical prediction simultaneously. A generative adversarial network synthesizes missing PET scans from available MRI data, while a classification network integrates features from both real and synthesized images for progression prediction. These components are jointly optimized through shared feature representations, enabling the synthesized images to be specifically tailored for the prediction task. The method further incorporates transfer learning to address data scarcity by pretraining on the large ADNI dataset before fine-tuning on smaller clinical cohorts. This allows the model to leverage patterns learned from extensive multimodal data even when applied to datasets with limited samples or missing modalities. Differently, Abdelaziz et al. [143] considered three

modalities, including T1-weighted MRI, FDG-PET scans, and genetic data. They explicitly addresses missing data through a linear interpolation technique to fill missing features in incomplete samples, and they demonstrated the statistical similarity between synthetic and real features. These studies are based on a restricted and unchanging set of modalities, necessitating training with fully aligned multimodal data. In real-world scenarios, however, this assumption may not always be valid due to the uneven distribution of modalities. As the number of assumed modalities grows, the proportion of complete modality data will decrease, leading to inadequate samples for effective training. At the same time, the total number of potential combinations will increase dramatically, which might not correspond to the specific combination observed during the inference process. Furthermore, patients could have AD-related information that falls outside the predefined modalities.

### 2.2.3 ARDS Risk Prediction

Highly relevant works are first introduced in chronological order, followed by conclusions on clinical setting and prediction objectives, patient selection criteria, study scale and data sources, and approaches for handling missing data.

In early studies, simple rule-based electronic algorithms have been described that analyze EHR data to screen patients for ARDS [144, 145]. Gajic et al. [146] validated the effectiveness of Lung injury prediction score (LIPS) calculated from clinical records (weights adjusted based on logistic regression) as initial risk assessment of acute lung injury at the time of admission. It was reported that in 2500 patients, LIPS achieved AUROC of $80\%(95\%CI : 76\% - 83\%)$, with sensitivity

of $69\%(95\%CI : 64\% - 74\%)$, PPV of $18\%(95\%CI : 16\% - 20\%)$.

Later, Reamaroon et al. [147] addressed the challenge of label uncertainty in ARDS diagnosis, arguing that could lead to more efficiently learning and better generalize to new patient cases. They include only 401 patients moderate hypoxia patients and utilized a soft-margin support vector machine (SVM) where clinician-provided confidence levels for each diagnosis are integrated through a modified regularization term. Additional, they proposed a specialized sampling strategy designed to reduce correlations among multivariate clinical time-series features. While demonstrating improved generalization capabilities, the method's reliance on manual confidence annotations and its linear treatment of temporal patterns present opportunities for enhancement through more sophisticated modeling techniques.

Zeiberg et al. [148] pointed out the limited clinical utility of existing ARDS prediction models, attributing this to their reliance on manual chart abstraction and suboptimal performance. To address this, they trained and validated a logistic regression model on a single-center dataset comprising 1,621 (51 ARDS) and 1,122 (27 ARDS) patients, respectively, who presented with moderate hypoxia. The model achieved AUROC of $81\%$, and sensitivity of $56\%$, specificity of $86\%$, PPV of $9\%$, relative risk score of 17, using a strict threshold $85^{th}$ percentile of risk. However, their approach has several limitations. Firstly, it relies solely on EHR data, thereby overlooking complementary information from CXRs crucial for comprehensive lung injury assessment. Secondly, the model's reliance on only 6 hours of EHR data neglects longer-term ICU monitoring data, which is vital for capturing the progression of a patient's condition. Furthermore, using summary statistics to represent time-series data leads to a loss of crucial temporal correla-

tion and progression information. Additionally, defining ARDS onset as the point where all diagnostic criteria are met can be problematic, as this time point may be variable and lag behind the actual clinical deterioration, particularly due to potential delays in relevant examinations like CXRs. Finally, their prediction time is set when hypoxia has already progressed, limiting its utility for truly early or preventative interventions.

Reamaroon et al. [149] highlighted the critical role of chest X-rays (CXR) in ARDS diagnosis and treatment. They devised a 'directionality measure' descriptor to capture lung infiltrates, demonstrating its effectiveness against common descriptors and features from a pre-trained ResNet-50. Using a simple SVM as the classifier, their method achieved AUROC of $74\%$ and accuracy of $78\%$. However, their approach relied on a deep network pre-trained on non-medical data and focused on CXRs from patients with already developed ARDS, limiting its applicability for early warning purposes. Also focused on CXRs, Sjoding et al. [150] conducted a large-scale study employing a DenseNet model pre-trained on CheXpert and MIMIC-CXR for common findings, and subsequently fine-tuned on consecutive patients with hypoxemia from a single hospital. They utilized all CXRs acquired during the first seven days of hospitalization, which were reviewed by at least two physicians who rated the images on a scale of 1-8 for the presence of bilateral opacities consistent with ARDS, accounting for rater consistency and diagnostic uncertainty. Their approach focused on diagnosing ARDS at the clear manifestation of diagnostic features, rather than serving as an early warning system. The study involved 8,072 CXRs for training and 958 CXRs from 431 patients for testing, achieving an AUROC of $88\%(95\%CI : 85\% - 91\%)$.Pai et al. [151] proposed a multimodal approach for ARDS prediction, integrating EHR

and CXR data collected within the first 48 hours of admission. Their method employed a Convolutional Neural Network (CNN) for CXR classification and traditional machine learning models for EHR data, with ensemble methods combining these outputs to predict the final ARDS probability. Trained and validated on a relatively small (1577 in total, $20\%$ for validation) ad single-centre dataset, their model achieved an AUROC of $92.5\%$. However, their approach presents several limitations. Firstly, relying on mean values for time-series data overlooks crucial temporal progression information. Secondly, their single-pass prediction, based solely on the initial 48 hours of data, disregards patient-specific ARDS progression, thereby limiting its utility for dynamic monitoring or truly preventative interventions.

Xu et al. [152] included 16523 sepsis patients for ARDS prediction from MIMIC-IV database, used multivariate logistic regression from selected variables of vital signs and laboratory measurements within 48 hours after admission. they use only initial measurement of repeated examinations. achieved AUROC of $81.2\%(95\%CI : 79.8\% - 82.6\%)$, accuracy $70.5\%(95\%CI : 77.3\% - 82.3\%)$, sensitivity of $79.8\%(95\%CI : 77.3\% - 82.3\%)$, specificity of $68.2\%(95\%CI : 66.8\% - 69.7\%)$, PPV of $38.5\%(95\%CI : 36.4\% - 40.7\%)$. Recently, Tran et al. [153] conducted a systematic review that included 52 studies for prediction, classification and management of ARDS. They reported that only 8 studies leveraged neural network while gradient boosting remianed to be the main stream method. In addition to CXR and EHR data, CT imaging has also been utilized to enhance AI-based ARDS prediction. Zhou et al. [154] highlighted that the complex pathophysiology of ARDS poses significant challenges for early prediction. To address this, they incorporated CT for a more detailed assessment of lung pathology. Their

study included 928 patients for training and validation purposes. A UNset Transformer model was employed to segment lung lesions and predict ARDS, achieving an AUROC of $86.5\%(95\%CI : 77.4\% - 94.5\%)$. However, due to the reliance on CT imaging, the sample size remains relatively small.

Regarding clinical settings, studies primarily focus on: 1) identifying pathological features from medical imaging to enhance diagnostic accuracy [149, 150, 154], 2) leveraging EHR data for diagnosis (e.g., Reamaroon et al. [147] using 24 EHR variables sampled every 2 hours with pre-onset timestamps labeled as non-ARDS), and 3) predicting future ARDS risk using EHR or multimodal data [151]. Notably, most risk prediction studies employ fixed time windows from hospital admission [146, 151, 152], with only Zeiberg et al. [148] using moderate hypoxemia as a dynamic reference point. Patient selection strategies vary significantly: diagnostic studies predominantly involve hypoxemic patients [147, 149, 150] (excepting CT segmentation work including patients with any of four risk factors), while risk prediction studies employ heterogeneous criteria including hypoxemia at prediction [148], predisposing conditions [146], sepsis-specific cohorts [152], or no risk constraints [151]. Label accuracy remains challenging due to the subjective interpretation of imaging required by Berlin criteria. While most studies use this framework [148, 154], some address diagnostic uncertainty through clinical expert consensus [147, 150] at substantial human resource cost. Study scales show limited generalization: most utilize single-center data with small sample sizes (<2,000 train/test cases) [147, 151, 154], excepting Sjoding et al.'s large-scale effort [150]. Only a minority leverage multi-center databases like MIMIC-IV [148, 152]. For missing data, imaging studies exclude cases with unavailable scans, while longitudinal analyses rely on simple imputation methods [147, 151].

Crucially, no existing work employs deep learning-based time-series modeling to address temporal data gaps.

# Chapter 3

# Multimodal MRI Synthesis: Integration of Multi-Modal Images

This chapter studies flexible integration of multi-modal images for magnetic resonance imaging (MRI) synthesis, which aims at synthesizing high-quality MRI of target modality from its limited cues or other modalities. In this problem setting, it involves two homogeneous modalities as a starting case for this study on flexible modality integration. This task is a representative of dense prediction tasks, which focus on spatial fidelity, pixel-level precision, and preserving anatomical or structural details. In the following sections, problem background and research gap are introduced in Section 3.1. The method details are presented in Section 3.2, where explanations on how the proposed formulation guides the architecture design of the synthesis network to enable flexible modality integration and fine-grained difference learning in Section 3.2.1, and the technical details for modality difference modeling are elaborated in Section 3.2.2-3.2.3. Experimental details for data, model, training, and evaluation are provided in Section 3.3. Results are

presented in Section 3.4, and corresponding discussions in Section 3.5. Finally, the conclusion is provided in Section 3.6.

## 3.1 Problem Background and Research Gap

MRI is a radiation-free, non-invasive imaging modality renowned for its soft tissue contrast resolution, enabling precise differentiation of anatomical structures [16]. This is accomplished by acquiring multi-modal images of the same anatomical region with distinct tissue contrasts, achieved through varying pulse sequences and imaging parameters. Common MRI modalities include T1-weighted (T1w), T2-weighted (T2w), and FLAIR images. T1w imaging is particularly valuable for anatomical visualization and structural assessment, offering clear differentiation between gray and white matter in the brain, as well as enhanced contrast for fat-rich tissues such as meningiomas [155]. T2-weighted (T2w) and FLAIR images are especially sensitive to fluid-related structures and pathological changes. T2w imaging provides excellent contrast for cerebrospinal fluid (CSF) in the brain and spinal cord, facilitating clear visualization of ventricles and CSF-filled spaces. Meanwhile, FLAIR sequences are optimized to suppress CSF signal, thereby improving the detection of periventricular and white matter lesions, which is a key feature in diagnosing conditions such as multiple sclerosis [156]. Consequently, specific anatomical features, abnormalities, or lesions may exhibit distinct structural patterns in certain MRI modalities while appearing as signal-void or relatively homogeneous in others. For precise diagnosis and comprehensive assessment, these multi-modal images should be analyzed at high resolution to capture fine structural details [157].

However, acquiring high-resolution (HR) MRI images across multiple modalities simultaneously remains challenging due to inherent differences in acquisition times. T1-weighted sequences, with their shorter repetition (TR$<$500ms) and echo times (TE$<$30ms), enable high-resolution imaging within relatively brief scan durations. In contrast, T2-weighted and FLAIR sequences require substantially longer repetition times (TR$>$2000ms) and inversion times (TI$\approx$2000ms) [16]. The Prolonged acquisition time not only reduce patient comfort but also increase susceptibility to motion artifacts. In clinical practice, this often necessitates compromising on image quality, that is, more time-intensive modalities are frequently acquired at reduced resolution or, in certain cases, omitted entirely from the imaging protocol [17, 18].

To overcome these limitations, researchers have developed algorithms that leverage high-resolution (HR) reference-modality images to reconstruct either under-sampled or completely missing target-modality scans. In medical imaging literature, this problem is categorized into two distinct tasks: (1) cross-modality synthesis (CMS) when the target modality is entirely absent, and (2) multi-contrast super-resolution (MCSR) when dealing with under-sampled acquisitions. While both approaches aim to generate HR target-modality images within the same anatomical coordinate system, existing methods have primarily focused on exploiting structural similarities across different contrasts, ignoring the critical challenges arising from subtle inter-modality variations: (1) Non-linear spatial misalignment: Despite initial affine registration, persistent non-linear deformations occur due to patient motion, tissue elasticity, and imaging artifacts [19], which degrade reconstruction performance [20, 21] and diagnostic accuracy [22]. (2) Structural distinction: Even under ideal alignment conditions, accurate reconstruction remains

challenging because certain anatomical features may be entirely absent in the reference modality, and when available in the target modality, these structures often appear severely degraded in the input data.

While these two tasks share fundamental connections, they represent distinct methodological directions in medical image analysis. For CMS, state-of-the-art approaches primarily employ conditional generative frameworks that simultaneously learn the target modality's data distribution, and preserve precise structural alignment with the high-resolution reference image. Pix2Pix [94] and CycleGAN [98] are two representative methods. For structural consistency, CycleGAN-based approaches rely on cycle-consistent adversarial learning between domains, whereas Pix2Pix-based methods employ pixel-wise loss functions that enforce stricter alignment constraints on training pairs. While the pixel-level reconstruction loss in Pix2Pix has demonstrated superior accuracy over CycleGAN approaches [18], this comes at the cost of greater sensitivity to spatial misalignments. This trade-off has motivated increasing research attention [18, 21], with recent solutions frequently integrating registration networks to properly align reconstruction supervision with reference coordinates.

In contrast to CMS, MCSR is typically examined alongside single-modal super-resolution (SR), as high-resolution (HR) reference images have demonstrated significant potential for enhancing severely degraded target-modality scans. The fundamental challenge in MCSR lies in effectively modeling inter-modal relationships compared to single-modal SR. Current approaches predominantly exploit structural similarities between modalities, where methods assuming perfect anatomical alignment typically employ simple feature concatenation for modality fusion [20, 109, 110]. This strategy proves effective when alignment serves

as a reliable prior for leveraging shared structures against degradation artifacts, though spatial misalignment issues have been acknowledged [20]. Unlike CMS methods [18, 21] that maintain a fixed HR reference, the MCSR approach in [20] deforms the HR reference to match the low-resolution (LR) target - a particularly challenging task given the ill-posed nature of SR reconstruction. This process introduces alignment ambiguities and risks degrading the reference image quality. Recent advances have developed more sophisticated techniques to harness global similarity from references, including spatial-channel attention mechanisms [17, 111], global matching strategies [117, 118], and cross-attention Transformers [112, 113, 114, 115]. While these methods effectively utilize global repetitive patterns and improve overall performance, they do not explicitly enforce anatomical correspondence and still attempt to align references to degraded targets rather than producing properly aligned HR pairs. Another key distinction from CMS is MCSR's historical dominance by conventional reconstruction methods. The field has recently incorporated generative approaches through transformer-based priors [124] and diffusion models [125], though these currently lack sophisticated conditioning mechanisms to optimally balance structural accuracy with visual fidelity.

A critical yet understudied challenge in deep learning-based approaches is the handling of modality-specific structures. While these distinctive features often occupy limited spatial regions, they provide crucial diagnostic information [23] and present substantial technical difficulties. For methods relying on strict alignment assumptions, reference modality structures can generate artifactual responses in the reconstructed output. Global similarity-based approaches circumvent this issue through implicit matching strategies, but their performance degrades when processing distinct, severely downsampled target patches. Consequently, cur-

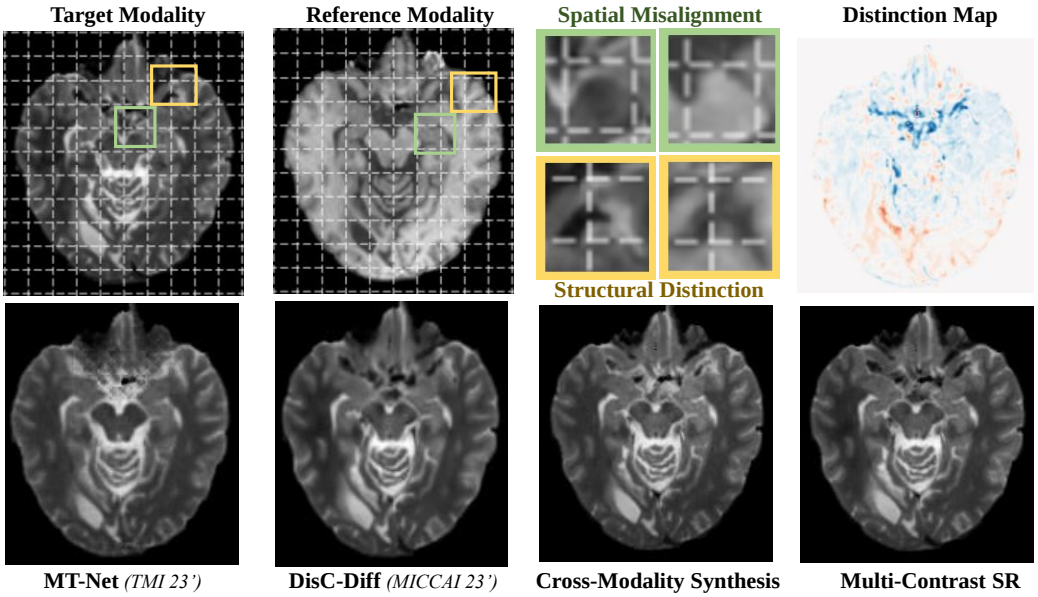



Figure 3.1: Illustration of fine-grained differences & Comparison with state-of-the-art methods.

rent reconstructions frequently exhibit either structural disappearance or excessive blurring (Fig. 3.1). Traditional methods have explicitly acknowledged structural distinction as a known limitation [106, 108]. The work in [107] proposed modeling inter-modal correlations through patch-level similarity graphs, using these to weight modality contributions during interpolation. While their correlation maps captured coarse structural differences robust to noise, the approach incurred substantial computational overhead and the basic weighted interpolation scheme proved inadequate for accurate feature restoration.

## 3.2 Methods

A unified framework is presented to address both CMS and MCSR through coordinated learning of spatial coordinate alignment and structural intensity differences.

Unlike prior MCSR approaches like [20] that deform HR references to match LR targets, the proposed method anchors both tasks to reference coordinates through a generalized down-sampling ratio. This formulation reveals that structural differences between reconstructed outputs and aligned ground truth peak under CMS conditions, formally defining modality-specific *structural distinction* as features irreducible from reference modalities. The framework comprises four synergistic components: 1) A label correction module establishing unified spatial coordinates; 2) A CMS base model providing structural distinction priors; 3) An SR branch leveraging increasing target cues ($k$ reduction) to approximate distinctions; 4) A difference projection discriminator with fine-grained conditioning on distinctive regions. For the SR branch, several novel modules are introduced: 1) Deformable convolution integration [158] enabling multi-scale feature alignment 2) Feature-regularized incremental modulation that characterizes structural distinctions via forward-pass analysis and guides spatially-variant generation through precision-controlled generative learning. This dual approach optimally utilizes misaligned LR targets while maintaining reference integrity, achieving unprecedented balance between structural fidelity and generative realism. The CMS module serves as both distinction prior and stability anchor, particularly crucial when handling severely degraded inputs.

### 3.2.1 Problem Formulation and Overall Architecture

Given a paired training dataset with HR reference images $\{I_{\text{ref}}\}$ and non-aligned target images $\{\tilde{I}_{\text{tar}}\}$, the LR target input is defined as $\{\mathcal{S}_k(\tilde{I}_{\text{tar}})\}$, where $\mathcal{S}$ represents the degradation operator (k-space truncation in the implementation) and

$k \in (1, \infty)$ denotes the down-sampling ratio. The unavailable aligned target images $\{I_{\text{tar}}\}$ are estimated through a label correction module:

$$\hat{I}_{\text{tar}} = \tilde{I}_{\text{tar}} \circ \mathcal{R}(\tilde{I}_{\text{tar}}, I_{\text{ref}}) \tag{3.1}$$

where $\mathcal{R}$ is a registration network approximating the non-linear transformation to reference coordinates, and $\circ$ denotes pixel-wise warping via the learned deformation field.

After label correction, both CMS and MCSR tasks can be expressed through a unified mapping function $\{\mathcal{F}_k(I_{\text{ref}}, S_k(\tilde{I}_{\text{tar}}))\}_k$, which transforms the reference image and varying levels of target information to the aligned target domain. Here, CMS emerges as the limiting case when $k \to \infty$ (complete absence of target information). The structural difference map:

$$D_k = \mathcal{F}_k(I_{\text{ref}}, S_k(\tilde{I}_{\text{tar}})) - \hat{I}_{\text{tar}} \tag{3.2}$$

where magnitude $\|D_k\|_1$, identifies challenging regions common to both tasks, which is addressed using a novel difference projection discriminator (Sec. 3.2.2).

The difference magnitude $\|D_k\|_1$ exhibits several fundamental properties that characterize structural distinction. First, it increases monotonically with the down-sampling ratio $k$, reaching its upper bound $\|D_\infty\|_1$ in the complete absence of target information ($k \in (1, \infty)$), which corresponds to the CMS task. This limiting case $D_\infty$ provides a formal pixel-level definition of modality-specific structural distinction, precisely localizing regions where information cannot be transferred from the reference modality. As $k$ decreases, the task shifts to better approximat-

ing these structural distinctions by leveraging the increasing target information available in $S_k(\tilde{I}_{\text{tar}})$, which represents the unique challenge for MCSR compared to CMS.

Accordingly, the proposed framework consists of two interconnected components designed to address both alignment and distinction challenges. The base network models $\mathcal{F}_\infty$ o solve the CMS task and estimate the structural distinction map $D_\infty$, with its down-sampled versions denoted as $d_i$ at each level $i$. Building upon this foundation, the SR branch processes the degraded target input $S_k(\tilde{I}_{\text{tar}})$ while incorporating features from the CMS module. To handle spatial misalignment between modalities, deformable convolutions [158] are employed that progressively align target features with the reference coordinates, as detailed in Section 3.2.3.1. The structural distinction is characterized through correlation analysis of cross-modal features (Section **??**), which is then embedded into an incremental modulation scheme to guide the spatially-variant generation of distinct anatomical structures (Section 3.2.3.3).

### 3.2.2 Difference Projection Discriminator

The structural difference map serves as a critical indicator of target regions that are insufficiently constrained by input data, representing the primary reconstruction challenge for both CMS and MCSR tasks. Current approaches exhibit complementary strengths and limitations: MCSR methods predominantly employ discriminative networks that achieve high precision for common structures but excessively smooth distinctive features, while CMS techniques typically leverage generative models that better synthesize target-aligned details but with greater uncertainty.

The proposed framework bridges this methodological divide by developing an enhanced conditional generative adversarial network (cGAN) architecture that combines their respective advantages through a novel conditioning mechanism.

Building upon the projection discriminator concept [159], which employs inner products between embedded conditional and feature vectors rather than simple concatenation [160], a multi-scale difference projection discriminator is introduced. This architecture extends the U-shaped discriminator design [161] while significantly advancing beyond prior projection approaches like [162] through pixel-level structural distinction mapping. The key innovation lies in using high-dimensional structural distinction maps to guide the projection, as these maps explicitly identify challenging regions with higher likelihood of synthetic artifacts.

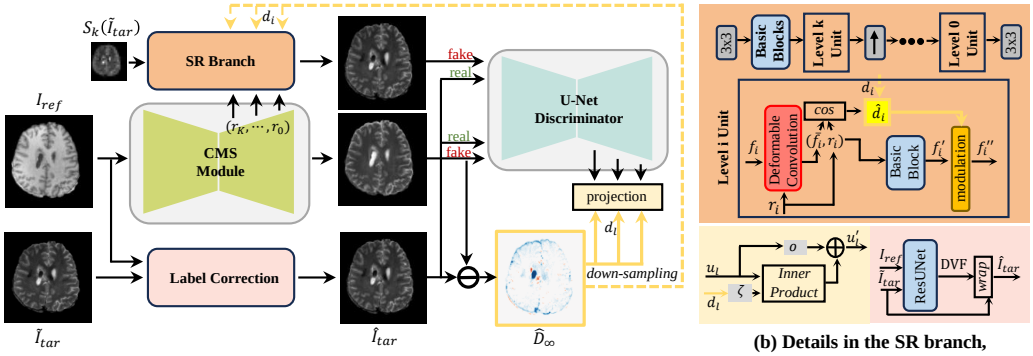
The proposed discriminator architecture comprises $L$-level encoder $U_{\text{enc}}$ components, with relativistic GAN loss [163] enhancing stability. The decoder performs multi-level discrimination between generated output $\mathcal{F}_k(I_{\text{ref}}, S_k(\tilde{I}_{\text{tar}}))$ and corrected target $\hat{I}_{\text{tar}}$. For decoder logits $u_l$ at level $l$, the projection updates:

$$u'_l = \langle \zeta(d_l), u_l \rangle + o(u_l) \tag{3.3}$$

where $d_l = D^l_\infty$ is the down-sampled distinction map, $\zeta$ represents convolutional transformation, and $o$ captures residual terms. The resulting discriminator loss becomes:

$$\begin{aligned} \mathcal{L}_{U_{\text{dec}}} = &-\mathbb{E}(\sum_{i,j,l} \log\left(\sigma(\dot{u}'_l - \mathbb{E}(\hat{u}'_l))\right)_{i,j}) \\ &-\mathbb{E}(\sum_{i,j,l} \log(1 - \sigma(\hat{u}'_l - \mathbb{E}(\dot{u}'_l)))_{i,j})). \end{aligned} \tag{3.4}$$

The training process employs a carefully designed two-stage approach: (1)




**(a) Overview of the composite architecture** **(b) Details in the SR branch, projection and label correction**

Figure 3.2: Overview of network architecture.

Initial CMS module training with progressively refined structural distinction approximation; (2) Subsequent SR branch optimization with frozen $D_\infty$ representation The difference projection discriminator participates in both stages, enabling comprehensive characterization of fine-grained data distributions across all structural regions. This dual-phase strategy ensures precise control over the generation process, achieving optimal balance between anatomical faithfulness and visual realism in the reconstructed outputs.

### 3.2.3 Difference Learning in SR branch

#### 3.2.3.1 Deformable Convolution for Feature-Level Alignment

To tackle the spatial misalignment between inputs, a recent approach [20] utilized an extra registration network to transform the high-resolution (HR) reference image into the unconstrained target coordinate system. However, this introduces a complex and ill-posed problem, significantly increasing computational costs for both training and inference. Additionally, the final outcomes heavily depend on the precision of a single displacement estimation. To resolve this challenge, cross-

modal spatial alignment is performed at the feature level. This is achieved by incorporating deformable convolutions, as proposed in [158], into the fundamental processing units within the super-resolution (SR) branch. Deformable convolutions adaptively modify the sampling locations of standard convolutions using learned offsets, which can be expressed through the following equation:

$$\bar{f}(\mathbf{p}) = \sum_{j=1}^{n^2} w\left(\mathbf{p}_j\right) \cdot f\left(\mathbf{p} + \mathbf{p}_j + \Delta\mathbf{p}_j\right), \tag{3.5}$$

where $f$ and $\bar{f}$ represent the input and output feature maps, respectively. $\mathbf{p}$ indicates the current spatial position within the feature map, while $\mathbf{p}_j$ corresponds to the $j_{th}$ standard sampling offset in the convolution kernel $w$ of size $n \times n$. In the SR branch, given multi-scale reference features $\{r_i, i = 0, 1, ..., K\}$, the target feature $f_i$ at each level $i$ is first aligned using a deformable convolutional layer before being passed into the fusion block. The offset $\Delta\mathbf{p}_j$ is estimated by the offset learning layer based on the multi-modal features $(f_i, r_i)$.

To elucidate the connection between explicit alignment and the proposed approach, the feature adjusted by the $j_{th}$ offset is represented as $f_j(\mathbf{p}) = f(\mathbf{p} + \mathbf{p}_j + \Delta\mathbf{p}_j)$, where $\mathbf{p}_j + \Delta\mathbf{p}_j$ denotes the deformation vector at position $\mathbf{p}$. This corresponds to feature-level spatial alignment, whereas explicit alignment focuses on learning a single deformation field at the image level. Consequently, Eq. 3.5 can be rewritten as

$$\bar{f}(\mathbf{p}) = \sum_{j=1}^{n^2} w\left(\mathbf{p}_j\right) \cdot f_j\left(\mathbf{p}\right) \tag{3.6}$$

This approach is equivalent to performing $n^2$ independent spatial deformations, succeeded by a $1 \times 1 \times n^2$ 3D convolution. Consequently, by integrating de-

formable convolutions into the hierarchical feature fusion process, it becomes possible to accomplish gradual multi-level alignment through varied and adaptable deformations, all while maintaining an end-to-end framework.

#### 3.2.3.2 Characterization of Structural Distinctions

Provided with $D_\infty$ derived from training pairs, the goal is to capture the structural differences arising from cross-modal inputs during the forward pass and integrate this information into an adaptive modulation mechanism to guide the generator. A simple solution might involve utilizing an auxiliary network to directly estimate the distinction map from $(S_k(\tilde{I}_{\text{tar}}), I_{\text{ref}})$. However, this approach not only increases the number of parameters but also introduces a highly unbalanced task that is difficult to optimize. To address this, a regularization technique is introduced that progressively models the distinction map by leveraging the correlation among multi-scale cross-modal features. As shown in Fig. 4.1, within the SR branch, the LR input undergoes processing through $k$ stages of processing modules. In each module, the target feature $f_i$ is initially aligned with the reference feature $r_i$, resulting in $\bar{f}_i$. Subsequently, for each pair of feature maps $\{\bar{f}_i, r_i\}$ of size $R^{\frac{H}{2^i}\times\frac{W}{2^i}\times C}$, the corresponding down-sampled distinction map $d_i = D^i_\infty$ (clipped within the range $(-1, 1)$) is generated and the feature regularization loss $L_D$ is reduced, formulated as follows:

$$\mathcal{L}_D = \sum_i \sum_{\mathbf{p}} \left\| \hat{d}_{i,\mathbf{p}} - d_{i,\mathbf{p}} \right\|^2, \hat{d}_{i,\mathbf{p}} = \frac{\bar{f}_{i,\mathbf{p}}^\top r_{i,\mathbf{p}}}{\|\bar{f}_{i,\mathbf{p}}\| \|r_{i,\mathbf{p}}\|} \tag{3.7}$$

where $\hat{d}_{i,\mathbf{p}} \in (-1, 1)$ represents the predicted difference at level $i$ for pixel position $\mathbf{p}$. This method provides an effective way to evaluate relative distinctions at the

pixel level, which is consistent with the goals for the SR branch features. By utilizing the known reference features, the target branch is motivated to investigate previously neglected details within the CMS module. Additionally, it highlights the direction of intensity variation by applying oriented linearity constraints on pixel features across different regions. The experimental findings indicate that feature regularization alone can enhance performance during the reconstruction phase.

#### 3.2.3.3 Fine-Grained Incremental Modulation.

Feature modulation, a technique that applies conditioned feature denormalization via learned affine transformations, is commonly incorporated into generative adversarial networks (GANs) to enable effective control over the generation process [164, 165, 166, 167]. The formulation of modulation should be adapted to suit the specific problem being addressed. For example, in style transfer tasks, affine parameters are utilized to govern the overall style of an input image and are typically represented as low-dimensional vectors (of size $1 \times 1 \times C$) derived from another image [164, 165]. In semantic synthesis, which focuses on generating detailed structures with semantic styles from input noise, the affine parameters assume the form of high-dimensional spatially adaptive maps (of size $H \times W \times C$) learned from semantic masks [166, 167]. In this work, the goal of modulation is to generate distinct details additively while maintaining the primary structure and global contrast. The approach involves learning spatially varying increments from the distinction map, which are subsequently superimposed on the original affine parameters to refine residual details.

In the SR branch, denote the updated feature after the basic block at level $i$

as $f_i'$, which undergoes instance normalization with a mean of $\mu_0$ and a standard deviation of $\sigma_0$. The modulation procedure can be rewritten as:

$$f_i'' = \left(1 + \frac{\Delta\sigma}{\sigma_0}\right) f_i' - \frac{\Delta\sigma}{\sigma_0}\mu_0 + \Delta\mu \tag{3.8}$$

, where $\Delta\sigma = \psi_1 \circ \phi\left(\hat{d}_i\right)$ and $\Delta\mu = \psi_2 \circ \phi\left(\hat{d}_i\right)$ represent spatially adaptive increments predicted through convolutional layers $\psi., \phi$ based on the approximated distinction map. Following the modulation step, $f_i''$ is upsampled to generate $f_{i+1}$ for use in the subsequent level. This method guarantees that the SR branch, acting as the generator, remains well-suited for progressively creating distinct structures.

## 3.3 Experiment Settings

### 3.3.1 Datasets

Three datasets were utilized for evaluation, with BraTSReg being pre-processed by the original providers.

**BraTSReg dataset.** The BraTSReg dataset [168] contains multi-institutional, multi-modal MRI scans of brain glioma patients. Pre-processing steps include skull stripping, bias field correction, and registration to a common anatomical template. All scans have a resolution of $256 \times 256$ pixels. T1-weighted (T1w) and T2-weighted (T2w) scans were selected, where T1w served as the reference contrast for multi-contrast super-resolution (MCSR). Experiments utilized 280 subjects from the training and validation sets.

**IXI dataset.** The IXI dataset provides healthy brain MRI scans across multiple modalities. T1w and T2w images were employed, with T1w designated as

the reference modality for MCSR. Following rigid registration, 577 subjects were retained. Axial slices were extracted, and images with insufficient content were filtered based on non-zero pixel counts.

**FastMRI dataset.** The FastMRI dataset [169] offers open-access knee MRI scans. Proton density-weighted (PDw) and fat-suppressed PDw (FSPDw) images from 1,054 subjects were selected. For MCSR experiments, PDw served as the reference contrast and FSPDw as the target.

### 3.3.2 Data Pre-Processing

Low-resolution (LR) target images were generated through $k$-space truncation, following established MCSR protocols [17, 117, 118]. The process involved: 1) Applying Discrete Fourier Transform (DFT) to $\tilde{I}_{\text{tar}}$, 2) Retaining the central $(1/\kappa, 1/\kappa)$ region of $k$-space data for down-sampling ratio $\kappa$, and 3) Reconstructing LR images via Inverse DFT.

### 3.3.3 Network Architecture

The framework incorporates modular components compatible with standard network blocks. Restormer blocks [119] were adopted as base units due to their feature dimension preservation. A 5-level U-shaped architecture formed the cross-modality synthesis (CMS) module, while the discriminator utilized a U-Net backbone [170]. The label correction module employed a ResUNet structure [171] with default configurations.

### 3.3.4 Training Details

Training proceeded in two phases. Phase I optimized the CMS module, label correction module, and difference projection discriminator using combined losses: 1) $\mathcal{L}_1$ and multi-scale SSIM [172] between reconstructions and aligned targets , and 2) Relativistic GAN losses [163] computed from discriminator high-level features. Phase II froze the CMS module while updating the SR branch with additional feature regularization $\mathcal{L}_D$. Implementation used PyTorch on NVIDIA 3090Ti GPUs, with the Ranger optimizer [173] ($lr = 1e^{-4}$) for both phases. A single CMS module handled all downsampling ratios across datasets. Comparative methods were evaluated using official implementations and recommended hyper-parameters, with synthetic $k$-space data generated via DFT where required.

### 3.3.5 Evaluation Metrics

Three metrics assessed reconstruction quality: 1) **PSNR**: Quantifies pixel-level fidelity through signal-to-noise ratio. 2) **SSIM**: Measures structural preservation using luminance, contrast, and pattern correlations. 3) **LPIPS** [174]: Evaluates perceptual similarity via deep feature comparisons. While PSNR/SSIM emphasize local pixel accuracy, they may favor overly smooth results. LPIPS addresses this limitation by aligning with human visual perception through learned feature correlations, demonstrating particular relevance in medical imaging validation [175].

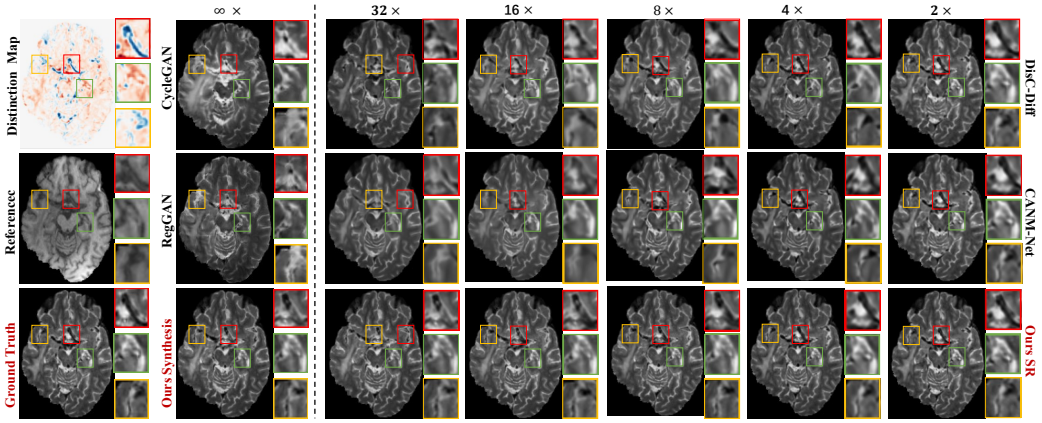



Figure 3.3: A qualitative comparison of both CMS and MCSR is presented, where the horizontal axis indicates progressively lower generalized down-sampling ratios. The distinction map and detailed views assist in the comparison, particularly for complex structures.

## 3.4 Results

### 3.4.1 Comparative Study

#### 3.4.1.1 Cross-Modality Synthesis

Four representative CMS methods were selected for comparison: Pix2pix [94] requiring pixel-aligned training pairs, CycleGAN [98] addressing unpaired data with increased output uncertainty, RegGAN [18] incorporating registration modules for result alignment, and MT-Net [102] leveraging unpaired pretraining for supervised learning. As shown in Tab. 3.3 and Tab. 3.4, RegGAN outperformed Pix2Pix and CycleGAN across metrics, demonstrating the benefits of explicit label alignment. The proposed method achieved superior LPIPS scores compared to all baselines while matching MT-Net in PSNR and SSIM. Visual comparisons in Fig. 3.3 (left) revealed that CycleGAN suffered from structural distortions and local artifacts, whereas RegGAN maintained better contrast alignment but failed

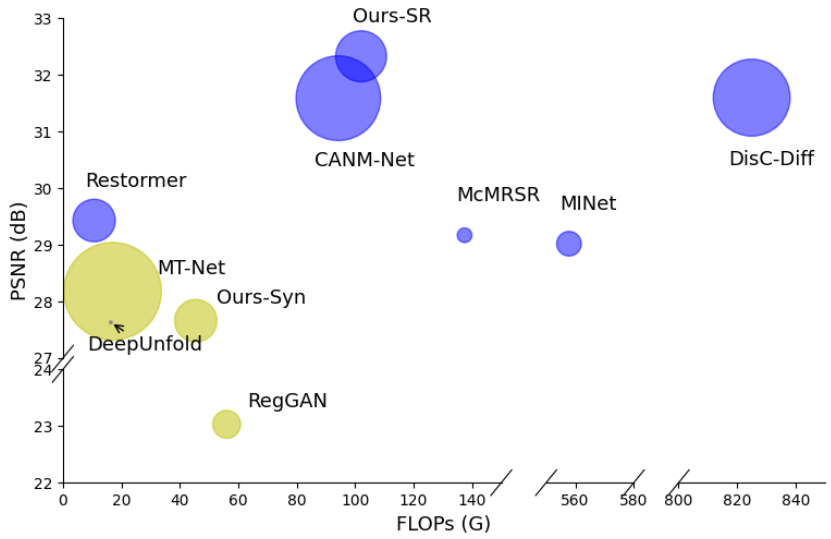



Figure 3.4: Analysis of computational efficiency. The SR techniques are evaluated at a $4\times$ scale, and the corresponding PSNR values are reported based on the IXI dataset. In the visualization, the size of each blob corresponds to the number of parameters (in millions, M) multiplied by a scaling factor to adjust for the figure dimensions.

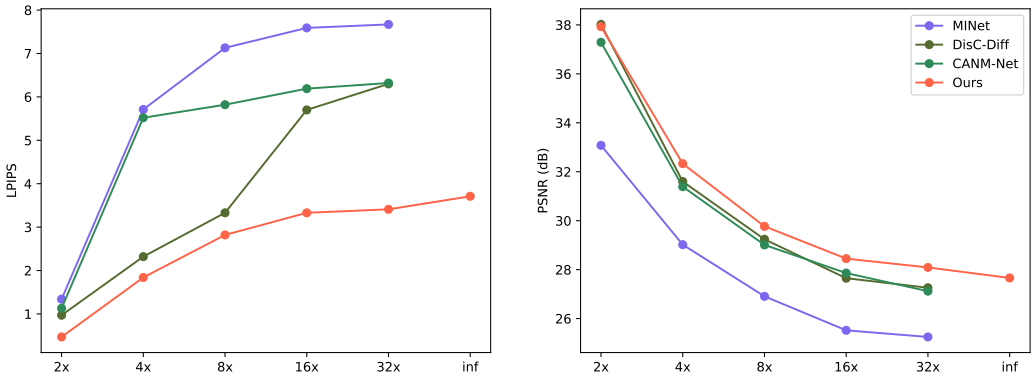



Figure 3.5: Evaluation of various down-sampling ratios using the IXI dataset.

Table 3.1: Quantitative evaluation of SR techniques applied to FastMRI (reference: PDw, target: PDFSw).The top and second-ranking outcomes are highlighted in red and blue, respectively. Findings that do not show a statistically significant difference compared to this method are indicated with an asterisk. The symbol * refers to the single-contrast SR approach.

| Method | FastMRI | | | | | |
|---|---|---|---|---|---|---|
| | 4× | | | 8× | | |
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
| Restormer* [119] | $30.61 \pm 0.56$ | $80.78 \pm 1.21$ | $20.30 \pm 1.05$ | $26.80 \pm 0.54$ | $78.14 \pm 1.16$ | $24.15 \pm 1.12$ |
| DeepUnfold [176] | $28.50 \pm 0.55$ | $77.23 \pm 1.38$ | $25.14 \pm 1.12$ | $25.98 \pm 0.56$ | $75.96 \pm 1.35$ | $25.60 \pm 1.11$ |
| MINet [17] | $30.11 \pm 0.48$ | $81.25 \pm 0.96$ | $20.32 \pm 0.88$ | $27.32 \pm 0.46$ | $78.88 \pm 1.02$ | $20.41 \pm 0.84$ |
| McMRSR [117] | $29.68 \pm 0.40$ | $79.92 \pm 1.13$ | $22.42 \pm 1.12$ | $26.48 \pm 0.47$ | $77.80 \pm 1.36$ | $23.45 \pm 1.08$ |
| WavTrans [115] | $31.78 \pm 0.48$ | $85.56 \pm 0.88$ | $16.02 \pm 0.80$ | x | x | x |
| DisC-Diff [125] | $31.60 \pm 0.46$ | $85.20 \pm 0.95$* | $11.36 \pm 0.68$ | $29.68 \pm 0.46$ | $80.10 \pm 1.32$ | $16.16 \pm 0.76$ |
| CANM-Net [118] | $32.48 \pm 0.43$* | $85.96 \pm 0.88$* | $16.15 \pm 0.89$ | $30.08 \pm 0.43$ | $80.23 \pm 1.10$* | $18.29 \pm 0.77$ |
| Proposed SR | $32.22 \pm 0.27$ | $85.12 \pm 1.40$ | $8.09 \pm 0.40$ | $30.80 \pm 0.28$ | $80.93 \pm 1.78$ | $12.30 \pm 0.49$ |

to reconstruct fine anatomical details.

#### 3.4.1.2 Multi-Contrast Super-Resolution

Quantitative evaluations at 4× and 8× downsampling (Tab. 3.1 and Tab. 3.2) compared the proposed method against five approaches: similarity-based MINet [17], McMRSR [117], and WavTrans [115]; anatomy-constrained CANM-Net [118]; diffusion-based DisC-Diff [125]; and model-guided DeepUnfold [176]. Generative methods (proposed, DisC-Diff) achieved the best LPIPS scores, indicating superior perceptual quality, while maintaining competitive PSNR/SSIM. Fig. 3.5 demonstrates performance trends across downsampling ratios on the IXI dataset. At 2×, all methods exhibited comparable results. However, as the ratio increased, the proposed method smoothly approached the CMS performance bound, outperforming others significantly at extreme ratios (e.g., 16×). Visual results in Fig. 3.3 confirmed precise recovery of structural boundaries and spatial coherence under high degradation.

Table 3.2: Quantitative evaluation of SR techniques applied to IXI (reference: T1w, target: T2w) datasets. The top and second-ranking outcomes are highlighted in red and blue, respectively. Findings that do not show a statistically significant difference compared to this method are indicated with an asterisk. The symbol * refers to the single-contrast SR approach.

| Method | IXI | | | | | |
|---|---|---|---|---|---|---|
| | 4× | | | 8× | | |
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
| Restormer* [119] | 29.45 ± 0.24 | 90.24 ± 0.30 | 7.09 ± 0.16 | 26.90 ± 0.28 | 86.35 ± 0.69 | 7.13 ± 0.34 |
| DeepUnfold [176] | 25.60 ± 1.11 | 27.64 ± 0.42 | 87.55 ± 0.36 | 23.56 ± 0.90 | 25.42 ± 0.32 | 81.91 ± 0.69 |
| 26.36 ± 0.92 | | | | | | |
| MINet [17] | 29.02 ± 0.39 | 91.10 ± 0.30 | 5.50 ± 0.21 | 26.91 ± 0.28 | 86.35 ± 0.58 | 7.13 ± 0.31 |
| McMRSR [117] | 29.15 ± 0.22 | 89.83 ± 0.33 | 7.01 ± 0.19 | 25.90 ± 0.24 | 81.96 ± 0.58 | 11.14 ± 0.36 |
| WavTrans [115] | 32.08 ± 0.22 | 92.03 ± 0.30 | 5.02 ± 0.21 | x | x | x |
| DisC-Diff [125] | 31.62 ± 0.24 | 92.42 ± 0.28 | 2.32 ± 0.09 | 29.24 ± 0.21 | 90.48 ± 0.22* | 3.33 ± 0.16 |
| CANM-Net [118] | 31.40 ± 0.24 | 92.76 ± 0.34* | 5.82 ± 0.21 | 29.01 ± 0.25 | 89.83 ± 0.56 | 5.76 ± 0.20 |
| Proposed SR | 32.41 ± 0.21 | 92.59 ± 0.23 | 1.82 ± 0.09 | 29.91 ± 0.23 | 90.86 ± 0.32 | 2.78 ± 0.15 |

Table 3.3: Quantitative comparison of CMS methods on FastMRI. The best and second-best results are in red and blue. Results with no significant difference from this method are marked with asterisk.

| Method | FastMRI | | | | | |
|---|---|---|---|---|---|---|
| | PDw→PDFSw | | | PDFSw→PDw | | |
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
| Pix2pix [94] | 23.68 ± 0.59 | 75.24 ± 1.12 | 20.15 ± 0.82 | 23.12 ± 0.60 | 74.83 ± 1.10 | 20.12 ± 0.65 |
| CycleGAN [98] | 21.02 ± 0.66 | 72.51 ± 1.38 | 19.98 ± 0.61 | 20.99 ± 0.67 | 71.32 ± 1.38 | 19.59 ± 0.60 |
| RegGAN [18] | 26.06 ± 0.50 | 75.98 ± 1.54 | 18.01 ± 0.66 | 25.27 ± 0.45 | 75.69 ± 1.50 | 18.05 ± 0.68 |
| MT-Net [102] | 28.06 ± 0.51* | 78.11 ± 1.42* | 18.12 ± 0.56 | 27.09 ± 0.40* | 76.18 ± 1.40* | 18.04 ± 0.54 |
| Proposed Synthesis | 27.85 ± 0.49 | 77.92 ± 1.48 | 12.98 ± 0.58 | 27.13 ± 0.38 | 75.43 ± 1.46 | 12.14 ± 0.50 |

#### 3.4.1.3 Computational Efficiency

Model complexity was evaluated via parameter counts (M) and FLOPs (G), as illustrated in Fig. 3.4. The proposed method reduced parameters by 38% compared to DisC-Diff while achieving 2.1× faster inference. In CMS tasks, it matched MT-Net's PSNR with 22% fewer parameters and 15% lower FLOPs. Notably, the shared synthesis network across varying downsampling ratios enabled efficient multi-task deployment without redundant branches.

Table 3.4: Quantitative comparison of CMS methods on IXI. The best and second-best results are in red and blue. Results with no significant difference from proposed method are marked with asterisk.

| Method | IXI | | | | | |
|---|---|---|---|---|---|---|
| | T1w→T2w | | | T2w→T1w | | |
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
| Pix2pix [94] | $23.35 \pm 0.56$ | $75.08 \pm 0.90$ | $9.18 \pm 0.21$ | $22.89 \pm 0.62$ | $73.92 \pm 1.11$ | $9.06 \pm 0.20$ |
| CycleGAN [98] | $21.11 \pm 0.70$ | $75.32 \pm 0.88$ | $8.12 \pm 0.18$ | $21.14 \pm 0.72$ | $74.64 \pm 0.90$ | $8.07 \pm 0.18$ |
| RegGAN [18] | $24.12 \pm 0.60$ | $78.92 \pm 0.89$ | $6.02 \pm 0.20$ | $23.96 \pm 0.61$ | $79.01 \pm 0.90$ | $6.05 \pm 0.20$ |
| MT-Net [102] | $28.18 \pm 0.46^*$ | $88.56 \pm 0.41^*$ | $5.98 \pm 0.20$ | $27.28 \pm 0.42^*$ | $87.95 \pm 0.45^*$ | $6.12 \pm 0.18$ |
| Proposed Synthesis | $27.80 \pm 0.21$ | $88.65 \pm 0.30$ | $3.80 \pm 0.12$ | $27.14 \pm 0.22$ | $88.04 \pm 0.32$ | $3.55 \pm 0.11$ |

### 3.4.2 Ablation Study

The proposed method addresses CMS and MCSR challenges through a unified difference learning framework, aiming to reconstruct complex structures while maintaining robustness against spatial misalignment between training pairs $(I_{\text{ref}}, \tilde{I}_{\text{tar}})$ and inputs $(\mathcal{S}_k(\tilde{I}_{\text{tar}}), I_{\text{ref}})$. This section evaluates the effectiveness of key designs in handling spatial misalignment and structural distinction, with results and analyses presented separately.

#### 3.4.2.1 Spatial Misalignment

Two key components were evaluated for robustness against spatial misalignment: (1) a label correction module addressing misalignment between training pairs, and (2) deformable convolutions [158] integrated into the SR branch for implicit feature-level alignment (Sec. 3.2.3.1). Experiments on the BraTSReg dataset utilized aligned target images as ground truth, with Gaussian deformation fields applied to simulate misaligned inputs. As shown in Tab. 3.5, the label correction module proved critical, its absence led to degraded structural distinction estimation and significant performance drops. Compared to SOTA MCSR methods (DisC-

Table 3.5: Quantitative analysis was conducted by comparing this approach with the top-performing MCSR methods. Additionally, an ablation study was performed on deformable convolutions (DCs) using different down-sampling rates and deformation strengths on the BraTSReg dataset. Here, 15% and 30% represent the relative deviation of the Gaussian deformable field. The variant "Proposed w/o both" refers to the model that excludes both the label correction module and deformable convolutions. The best and second-best outcomes are highlighted in red and blue, respectively. Results showing no significant difference compared to proposed method are indicated with an asterisk.

| Method | | Strength: 15% | | | Strength: 30% | | |
|---|---|---|---|---|---|---|---|
| | | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
| DisC-Diff [125] | 2 × | 29.65 ± 0.11 | 93.83 ± 0.26 | 2.32 ± 0.06 | 26.85 ± 0.08 | 89.46 ± 0.12 | 3.62 ± 0.06 |
| | 4 × | 28.52 ± 0.09 | 90.46 ± 0.15 | 4.08 ± 0.05 | 26.50 ± 0.12 | 89.10 ± 0.11 | 5.68 ± 0.10 |
| | 8 × | 25.67 ± 0.07 | 86.68 ± 0.13 | 5.80 ± 0.07 | 25.61 ± 0.09 | 88.99 ± 0.12 | 7.05 ± 0.14 |
| CANM-Net [118] | 2 × | 31.61 ± 0.11 | 95.50 ± 0.06 | 2.41 ± 0.05 | 27.28 ± 0.09 | 91.66 ± 0.12 | 3.91 ± 0.04 |
| | 4 × | 31.02 ± 0.11 | 94.54 ± 0.07 | 4.54 ± 0.12 | 27.16 ± 0.10 | 91.46 ± 0.11 | 6.05 ± 0.11 |
| | 8 × | 29.98 ± 0.09 | 92.82 ± 0.08 | 6.34 ± 0.15 | 27.33 ± 0.09 | 90.97 ± 0.09 | 8.64 ± 0.15 |
| Proposed w/o both | 2 × | 32.49 ± 0.08 | 96.12 ± 0.06 | 2.36 ± 0.04 | 27.35 ± 0.08 | 91.22 ± 0.10 | 3.85 ± 0.04 |
| | 4 × | 30.64 ± 0.10 | 93.85 ± 0.13 | 3.54 ± 0.07 | 26.82 ± 0.08 | 90.54 ± 0.11 | 5.49 ± 0.05 |
| | 8 × | 29.46 ± 0.15 | 92.18 ± 0.23 | 4.58 ± 0.11 | 26.91 ± 0.12 | 89.83 ± 0.20 | 7.33 ± 0.14 |
| Proposed w/o DCs | 2 × | 33.59 ± 0.08 | 96.30 ± 0.06* | 2.34 ± 0.04 | 30.61 ± 0.10 | 93.75 ± 0.09* | 3.36 ± 0.04 |
| | 4 × | 32.08 ± 0.13 | 94.57 ± 0.15 | 3.41 ± 0.08* | 29.97 ± 0.12* | 92.92 ± 0.12* | 4.15 ± 0.07* |
| | 8 × | 30.16 ± 0.17* | 92.69 ± 0.24* | 4.58 ± 0.10* | 28.78 ± 0.17* | 91.51 ± 0.25* | 5.07 ± 0.12* |
| Proposed | 2 × | 33.81 ± 0.06 | 96.35 ± 0.06 | 2.17 ± 0.04 | 30.74 ± 0.09 | 93.82 ± 0.07 | 3.28 ± 0.04 |
| | 4 × | 32.39 ± 0.12 | 94.87 ± 0.14 | 3.39 ± 0.08 | 30.01 ± 0.12 | 92.92 ± 0.12 | 4.15 ± 0.06 |
| | 8 × | 30.30 ± 0.17 | 92.87 ± 0.22 | 4.47 ± 0.11 | 28.82 ± 0.17 | 91.50 ± 0.24 | 5.05 ± 0.10 |

Diff [125], CANM-Net [118]), the proposed method with label correction achieved superior metrics under deformation, particularly in LPIPS under strong misalignment. Deformable convolutions provided marginal improvements, especially at higher downsampling ratios (4×, 8×).

#### 3.4.2.2 Structural Differences

Structural distinction handling was evaluated through conditional GAN designs: (1) high-dimensional difference projection in the discriminator and (2) feature regularization with incremental generator modulation. Tab. 3.6 demonstrates that difference projection reduced prediction uncertainty (lower deviation) and improved

Table 3.6: Ablation analysis of design choices for conditional generation of structural differences on the BraTSReg dataset at $8\times$ magnification. Each variant is evaluated against the preceding one and denoted with an asterisk if no substantial difference is observed.

| Variants | Discriminator | | Generator | | | MCSR | | CMS | |
|---|---|---|---|---|---|---|---|---|---|
| | $U_{\text{dec}}$ | $U'_{\text{dec}}$ | FR | SPADE | IM | PSNR ↑ | LPIPS ↓ | PSNR ↑ | LPIPS ↓ |
| (1) | ✓ | | | | | $29.15 \pm 0.35$ | $4.78 \pm 0.17$ | $26.64 \pm 0.38$ | $5.42 \pm 0.25$ |
| (2) | | ✓ | | | | $30.78 \pm 0.23$ | $3.65 \pm 0.11$ | $28.85 \pm 0.21$ | $3.72 \pm 0.11$ |
| (3) | | ✓ | ✓ | | | $31.36 \pm 0.19$ | $3.11 \pm 0.08$ | x | x |
| (4) | | ✓ | ✓ | ✓ | | $31.31 \pm 0.21^*$ | $3.10 \pm 0.09^*$ | x | x |
| (5) | | ✓ | ✓ | | ✓ | $32.05 \pm 0.19$ | $2.73 \pm 0.09$ | x | x |

performance across both tasks when comparing baseline $U_{\text{dec}}$ and modified $U_{\text{dec}}$. Feature regularization alone provided modest gains (variant (2) vs. (3)), while incremental modulation outperformed SPADE [177] in leveraging regularized features for condition-guided synthesis.

## 3.5 Discussion

The fundamental role of label correction stems from its ability to preserve structural correspondence between misaligned training pairs, a prerequisite for accurate difference learning. While deformable convolutions enhance feature alignment, their limited impact at larger downsampling ratios suggests that explicit input-level alignment becomes less effective as resolution decreases. The superior performance over DisC-Diff and CANM-Net highlights the advantage of integrating label correction into a unified framework, rather than treating misalignment as a standalone preprocessing step. Notably, CANM-Net's relatively better robustness compared to DisC-Diff implies that matching-based approaches may inherently tolerate mild spatial deviations better than generative diffusion models.

The effectiveness of difference projection underscores the importance of explicitly modeling structural discrepancies in high-dimensional feature space, rather than relying solely on low-level intensity differences. Feature regularization acts as an intermediate representation that disentangles structural distinctions from domain-specific variations, enabling more stable adversarial training. The superiority of incremental modulation over SPADE suggests that adaptive, multi-scale condition integration, tailored to the specificity of structural distinctions, is crucial for preserving anatomical fidelity in cross-modal SR tasks. This aligns with observations that affine transformations in SPADE may oversimplify the relationship between conditioning features and generator outputs when handling heterogeneous structural patterns.

In comparative studies, for CMS, the performance gap highlights two critical factors: (1) explicit alignment mechanisms (as in RegGAN) mitigate structural mismatches but remain limited by registration accuracy, and (2) difference learning enables direct modeling of inter-modal discrepancies without relying on auxiliary alignment modules. While MT-Net's dual-stage training improves stability, its dependency on separate pretraining phases restricts adaptability to joint CMS-MCSR optimization. The visual superiority in distinct structure recovery (Fig. 3.3) further corroborates the advantage of integrating difference projection into adversarial conditioning. For MCSR, the divergence in performance scaling stems from two design aspects: (1) unified feature regularization prevents error accumulation in cascaded restoration steps, unlike model-guided or similarity-based approaches, and (2) incremental modulation adapts to varying input fidelity levels more effectively than fixed conditioning in diffusion models. While DisC-Diff benefits from the generative prior of diffusion models, its simple condition-

ing mechanism struggles with anatomical consistency under severe degradation. CANM-Net's anatomical constraints improve robustness but limit adaptability to non-local structural variations observed in multi-contrast MRI.

The computational efficiency gains originate from three factors: (1) parameter-sharing mechanisms in difference learning eliminate task-specific modules, (2) deformable convolutions reduce the need for explicit alignment networks, and (3) lightweight incremental modulation replaces computationally intensive spatial transformers used in SPADE-like architectures. While diffusion models excel in sample diversity, their iterative denoising process inherently increases computational costs—a critical limitation for medical imaging applications requiring real-time throughput. The unified architecture further demonstrates that joint CMS-MCSR optimization need not incur proportional complexity increases compared to standalone task models.

## 3.6 Conclusion

A unified framework is presented that bridges cross-modality synthesis and multi-modal MRI super-resolution. Unlike existing approaches that focus solely on similarity modeling or high-level contrast transfer, proposed method focus on fine-grained differences involving the non-linear spatial misalignments and structural distinctions. For spatial misalignments, a label correction module and deformable convolutions are introduced into proposed network architecture. This approach leads us to define structural distinction at the pixel level, enabling us to develop a precisely controlled generation method. Experiments demonstrate that proposed method achieves state-of-the-art performance, particularly in perceptual quality

and at higher down-sampling ratios. The proposed unified approach not only economizes computational expenses across various scenarios but also enhances the consistency of generated results at different down-sampling ratios.

Despite these advantages, the proposed method has several limitations in design and application. First, the benefits of using deformable convolutions diminish for small deformations at higher down-sampling ratios, as the degree of misalignment between inputs decreases with the loss of local structures. Second, proposed modality translation model currently processes only one pair of modalities at a time. In clinical practice, where multiple modalities may be available, it would be advantageous to extend the model to handle multiple inputs simultaneously and address more complex modality differences. Lastly, the experiments indicated that traditional metrics often fail to detect errors in small but important structures. Therefore, developing more effective metrics in the future to assess semantic accuracy in clinical settings is crucial.

# Chapter 4

# Alzheimer's Disease Diagnosis: Integration of Image and Non-Image Data

This chapter explores the adaptive integration of diverse multi-modal data for diagnosing Alzheimer's Disease. The approach utilizes both multi-modal imaging and various forms of tabular data to accurately classify normal cases, mild cognitive impairment (MCI), and Alzheimer's Disease (AD). In subsequent sections, challenges arising from these heterogeneous data modalities are discussed in the problem background and research gap (Section 4.1). Section 4.2.1 presents a symbolic formulation that defines the data framework and goals of this research. Sections 4.2.2 and 4.2.3 detail the design strategies for achieving flexible multi-modal learning, focusing on network architecture and fusion techniques. Experimental methodologies are described in Section 4.3. Results and analyses, including ablation studies, comparative evaluations, and the adaptability to novel modalities,

are presented in Sections 4.4 and 4.5. Finally, concluding remarks are provided in Section 4.6.

## 4.1 Problem Background and Research Gap

Recent advances in diagnostic technologies have facilitated the integration of diverse biomedical data modalities to enhance clinical decision-making, particularly for complex disorders like Alzheimer's Disease (AD). This progression has driven interest in multi-modal deep learning frameworks for computer-aided diagnosis [178]. However, medical multi-modal learning exhibits distinct characteristics compared to conventional multi-modal applications in vision, language, and audio domains [4]. Medical modalities encompass heterogeneous biosensor-derived data that lack semantic alignment and exhibit complex missing patterns in real-world scenarios.

AD diagnosis exemplifies these challenges, involving complementary yet imbalanced modalities such as T1-MRI, FDG-PET, clinical assessments, and biomarker measurements. Existing studies predominantly focus on fixed modality combinations (e.g., T1-MRI & FDG-PET [9], T1-MRI & tabular data [11]), requiring matched multi-modal data during both training and inference. Such approaches suffer from critical limitations: (1) The combinatorial explosion of potential modality configurations reduces the available training data per combination as modality count increases; (2) Clinical settings often encounter unseen modality combinations not addressed during model development; (3) Integration of novel modalities necessitates architectural modifications and retraining.

Architectural inflexibility constitutes a primary research gap. Current Transformer-

based solutions [2, 24] either incur quadratic computational costs through parallel modality processing or lack permutation invariance via serial cross-attention mechanisms [25]. Furthermore, their reliance on modality-specific components (e.g., separate backbone networks) increases parameter counts, exacerbating training instability and overfitting risks in data-scarce medical contexts.

A second critical gap lies in modality alignment strategies. In multi-modal learning, alignment serves a dual purpose: first, it involves establishing the relationships between different modalities within a shared metric space, which supports the fusion module in understanding inter-modal interactions [2, 24]; 2) second, it strengthens modality invariance to ensure robust performance despite variations in input (missing modalities) [**drfuse2024**, 91, 179]. Cross-modal alignment, a technique that promotes precise matching between features of co-occurring modalities, is commonly applied for both purposes [**drfuse2024**, 2, 24]. Fundamentally, this method restricts the modeling of interactions among different modalities to a unified framework (as illustrated in Fig. 4.2), enabling it to serve both alignment purposes concurrently. Nevertheless, this strategy falls short when addressing the unique requirements within the medical domain. In contrast to typical modalities such as vision, language, and audio, medical modalities lack inherent semantic connections. Instead, their relationships are task-dependent and frequently associated with a specific disease under investigation. Consequently, imposing direct cross-modal alignment may impede the discovery of distinct yet complementary information from diverse modalities, which is crucial for differential diagnosis.

## 4.2 Methods

### 4.2.1 Problem Formulation

Drawing from [5], the task is defined as a 3-way classification problem distinguishing Normal Cognition (NC), Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD). The objective is to utilize diverse modality combinations in the training data to predict AD classifications for any combination of known modalities while ensuring adaptability to novel modalities with minimal computational overhead. Let $M = \{m_1, m_2, \ldots, m_K\}$ represent the set of $K$ modalities observed during training, and $M_u = \{m_{u1}, \ldots\}$ denote unseen modalities. The non-empty power set $2^M = \{X \mid X \subseteq M, X \neq \emptyset\}$ encapsulates all possible combinations of known modalities. The specific combinations encountered during training are denoted as $C = \{X_i \mid X_i \in 2^M, \cup X_i = M\}$. During inference, two types of unseen scenarios are addressed: 1) novel combinations of known modalities, $C_u = 2^M \setminus C$, which can be directly inferred post-primary training; and 2) combinations involving unseen modalities, $C'_u = \{X \mid X \subseteq M \cup M_u, X \cap M_u \neq \emptyset\}$.

A combination $c \in C_*$ is termed a 'combination,' with $|c| = 1$ treated as a special case representing a single modality. A function $f$ is introduced to process samples from any modality combination, expressed as $f(\mathcal{X}_c; \theta, \cup\theta_{m_i})$, where $m_i \in c$ and $\mathcal{X}_c = \{x_{m_i}\}$ denotes the sample for combination $c \in C_*$. This function is implemented as a deep neural network with shared parameters $\theta$ across all scenarios and modality-specific parameters $\theta_{m_i}$, where $|\theta_{m_i}| \ll |\theta|$ to minimize modality-specific overhead. After training on $C$, the network can handle test samples $X_{c'}, c' \in 2^M$. For samples with unseen modalities, only the single unseen modalities $m'_i \in M_u$ require additional training to obtain $\theta_{m'_i}$, enabling direct in-

ference on any unseen combination in $C'_u$.

### 4.2.2 Architecture Design

The proposed architecture integrates two Transformer-based components for projection and fusion, as illustrated in Fig. 4.1. By incorporating modality-specific processing, the dimensionality of $\theta_{m_i}$ is significantly reduced, particularly for: a) initial feature integration and b) distance estimation in the projection phase. For a), modality-specific embedding layers $\mathcal{E}_{m_i}$ are designed to accommodate diverse modality formats. For 3D volumetric data, these layers utilize two 3D ResNet blocks to compress redundant information, transforming the resulting 3D features directly into initial tokens. For tabular data, the embedding layer expands each attribute to the feature dimension. Following [180], continuous values are processed via a linear layer, while categorical values use a look-up table. Specifically, each output token is computed as $e = b + h(x) \in \mathbb{R}^d$, where $x$ is a tabular value, $b$ is the feature bias, and $h$ is an element-wise multiplication with a vector $w \in \mathbb{R}^d$ for numerical values or a look-up table in $\mathbb{R}^{S\times d}$ for categorical values.

Inspired by [181], a modality-agnostic Transformer, denoted $\mathcal{G}(E_{m_i}, Q_{m_i}; \theta_G)$, is proposed. The architecture of $\mathcal{G}$ alternates between cross-attention layers and Transformer blocks. Each cross-attention layer incorporates a trainable, modality-specific query $Q_{m_i}$, enabling tailored distance modeling for different modalities. Notably, except for $Q_{m_i}$, the parameters $\theta_G$ are shared across all modalities, ensuring efficient and unified processing.

In the proposed framework, all input modalities $\{x_{m_i}\}$ are first converted into a set of initial tokens $\{E_{m_i}\}$. A modality-agnostic Transformer is then applied

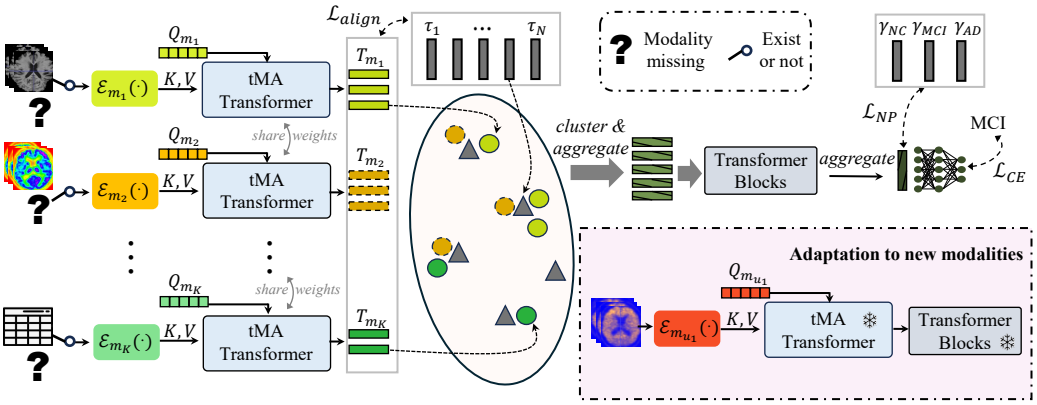
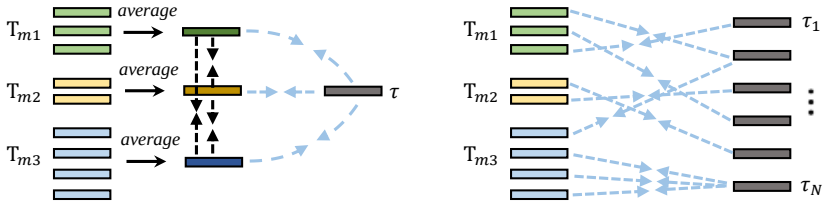



Figure 4.1: Pipeline of the proposed method, which involves a projection step that maps raw data into a shared metric space, and a fusion step that combines features to perform the final task.




(a) Prior work (b) Ours

Figure 4.2: Comparison with prior work with respect to multi-modal alignment in representation learning.

independently to each modality, producing a set of multi-modal feature tokens $t_j^i \in T_{m_i}$, where $t_j^i$ denotes the $j^{th}$ feature for modality $m_i$. This process is mathematically expressed as $\{T_{m_i} \mid T_{m_i} = \mathcal{G}(E_{m_i}, Q_{m_i})\}$. These feature tokens are subsequently projected into a fixed-length representation of size $N$. The resulting representations are processed by a fusion Transformer to generate class embeddings for final classification. Notably, the framework imposes no restrictions on the number of modalities, ensuring both projection and fusion Transformers handle all modalities uniformly. After pretraining, new modalities can be integrated by adding $Q_{m_u}$ and corresponding embedding layers $\mathcal{E}_{m_i}$.

### 4.2.3 Task-Oriented Fusion

The proposed unified multi-modal fusion approach emphasizes task-invariant elements. It assumes the existence of $N$ task-oriented factors, with implicit references $\{\tau_1, \tau_2, \ldots, \tau_N\}$ defined for these factors. The model is trained to align each feature token $t_j^i$ to one of these references using the alignment loss:

$$\mathcal{L}_{\text{align}} = -\sum_{ij} \log \left( \frac{\max \left\{ e^{(t_j^i)^\top \tau_1}, \ldots, e^{(t_j^i)^\top \tau_N} \right\}}{\sum_{n=1}^{N} e^{(t_j^i)^\top \tau_n}} \right). \tag{4.1}$$

Figure 4.2 illustrates how this alignment strategy differs from prior methods [1, 2, 3]. Traditional approaches enhance feature similarity across modality pairs (black arrows), effectively using a single anchor point (blue arrows). However, these methods face two limitations: 1) condensing feature tokens into a single representation per modality may overlook intra-modal variations, and 2) minimizing inter-modal distances during feature extraction can conflict with preserving complementary modality characteristics. In contrast, the proposed method uses multiple implicit task anchors to model both intra-modal and inter-modal similarities and differences concurrently.

For fusion, each feature token is assigned to a task-related factor via $v_j^i = \underset{n}{\text{argmax}}((t_j^i)^\top \tau_n)$. Features within each cluster $n$ are aggregated as:

$$t'_n = \sum_{v_i^j = n}^{i,j} \omega_i^j t_i^j, \quad \omega_i^j = \frac{e^{(t_j^i)^\top \tau_n}}{\sum_{v_{j'}^{i'} = n}^{i',j'} e^{(t_{j'}^{i'})^\top \tau_n}}. \tag{4.2}$$

This results in a fixed-size set of aggregated features $|\{t'_n\}| = N$, processed consistently by Transformer layers. The fusion Transformer's outputs are averaged to

produce the class embedding $\mu_y$, where $y$ is the class label. Task-level alignment is achieved using class-specific task anchors $\{\gamma_i\}$ for each class $i$, optimized via the N-pair Loss [182]:

$$\mathcal{L}_{\text{NP}}(\mu_y, \gamma_y, \{\gamma_i\}_{i \neq y}) = \log \left[ 1 + \sum_{i \neq y} e^{((\mu_y)^\top \gamma_i - (\mu_y)^\top \gamma_y)} \right]. \tag{4.3}$$

For samples of different combinations within the same class $y$, their embeddings are aligned closer to the shared class anchor $\gamma_y$, eliminating the need to sample multiple combinations per training step and enhancing training stability.

## 4.3 Experiment Setup

### 4.3.1 Dataset

The study leverages the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset https://adni.loni.usc.edu/, a longitudinal, multi-site study tracking Alzheimer's disease progression. It includes cognitively normal individuals, those with mild cognitive impairment (MCI), and Alzheimer's dementia patients, with repeated assessments. The dataset comprises 1.5T/3T structural MRI (T1, T2), PET scans (FDG, amyloid, tau), clinical/cognitive evaluations (e.g., ADAS-Cog, MMSE), CSF biomarkers (Aβ42, tau), and APOE genotyping. Collected under Institutional Review Board (IRB) approval and HIPAA compliance, the data is de-identified and accessible via the Laboratory of Neuro Imaging (LONI) after registration.

### 4.3.2 Data Preprocessing

Eight modalities were extracted from ADNI [183]: T1-weighted MRI (T1), T2-weighted MRI (T2), FDG-PET (F), Amyloid-PET (A), MMSE (Mm), MoCA (Mo), NeuroBat (Ne), and NPI-Q (Np). The unseen modality set is defined as $M_u = \{\text{A}, \text{Ne}\}$. For MRI, 6,231 MPRAGE T1w scans from 1,266 patients and 6,399 FLAIR T2w scans from 1,099 patients were used. MRI preprocessing involved spatial normalization to the MNI152 template ($91 \times 109 \times 91$), intensity normalization to $(0, 1)$ via min-max scaling, and skull stripping using FSL tools. Preprocessed PET scans ($160 \times 160 \times 96$), including 2,297 FDG-PET from 1,140 patients and 209 Amyloid-PET from 103 patients, were used as provided by ADNI. Tabular data included attributes from neuropsychiatric symptom assessment (NPI-Q), neuropsychological assessment (NeuroBat), and cognitive screening tools (MMSE, MoCA). The cohort with at least one modality totaled 3,881 patients.

All modality time points were used, with examinations within a six-month window paired to form multi-modal combination samples. The dataset was split at the patient level into training (80%) and testing (20%) sets for each combination, ensuring no data leakage.

### 4.3.3 Implementation Details

The model was implemented in PyTorch and trained on a single NVIDIA RTX A6000 GPU with 48GB memory. The classification task used cross-entropy loss ($\mathcal{L}_{CE}$), with the total loss as a weighted sum of $\mathcal{L}_{CE}$, $\mathcal{L}_{\text{align}}$, and $\mathcal{L}_{\text{NP}}$. Optimization employed AdamW with an initial learning rate of $3e-4$, following a Cosine

Annealing schedule with linear warm-up. To address modality imbalance, image augmentations included random noise, blurring, anisotropy, bias field simulation, ghosting, spike artifacts, motion distortions, affine transformations, and elastic deformations. Task anchors were represented as learnable parameters in $\mathbb{R}^{128}$. Performance was evaluated using weighted F1-score and Accuracy (ACC). For new modalities, the model was updated via supervised learning on all available training data.

## 4.4 Results

### 4.4.1 Ablation Studies

Ablation experiments were conducted on the modality set $\{\mathrm{T1}, \mathrm{F}, \mathrm{Mo}, \mathrm{Np}\}$, evaluating the impact of excluding specific components from the full model.

### 4.4.2 Ablation of Architectural Components.

The proposed architecture comprises: 1) a tunable modality-agnostic (MA) Transformer for shared projection parameters, and 2) a clustering mechanism in the fusion module to achieve fixed-length representations. For the first component, the tunable property of the MA Transformer was disabled by sharing query vectors $Q_{m_i}$, and then parameter sharing was entirely removed by employing modality-specific projection (MSP) models. As shown in Table 4.1, the proposed model achieved the highest average performance (F1: 0.598, ACC: 0.593) while using only $15\%$ of the parameters required by MSP (11.09M vs. 74.83M). MSP exhibited slower convergence ($3\times$ slower than the proposed model), unstable outcomes

Table 4.1: Ablation study results on projection architectures. MSP denotes modality-specific projection.

| | T1 | | F | | T1&F | | T1&F&Mo | | T1&Mo&F&Np | | Mean | | PARAMS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F1 | ACC | F1 | ACC | F1 | ACC | F1 | ACC | F1 | ACC | F1 | ACC | |
| MSP | 0.296 | 0.408 | **0.633** | **0.622** | **0.660** | **0.661** | **0.687** | **0.686** | 0.732 | **0.735** | 0.589 | 0.585 | 74.83 M |
| share $Q_{m_i}$ | 0.465 | 0.461 | 0.433 | 0.518 | 0.434 | 0.525 | 0.680 | 0.674 | 0.727 | 0.725 | 0.588 | 0.582 | 11.08 M |
| Proposed | **0.512** | **0.515** | 0.603 | 0.595 | 0.606 | 0.618 | 0.675 | 0.677 | **0.734** | 0.732 | **0.595** | **0.590** | 11.09 M |

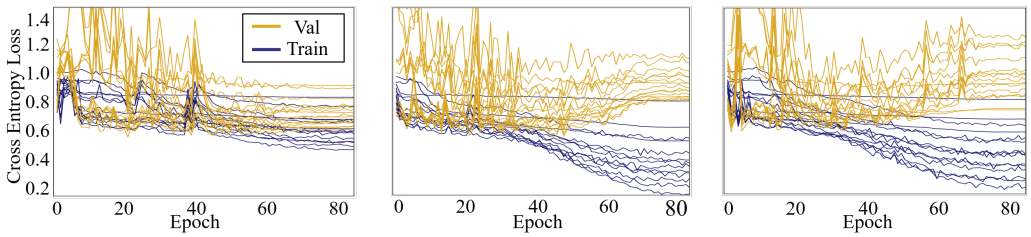
Table 4.2: Ablation study results on the fusion module. 'c' denotes clustering, $\mathcal{L}_a$ represents $\mathcal{L}_{\text{align}}$. 'Mean' reflects the average performance across all tested modalities.

| | | T1&F | | T1&Mo | | F&Mo | | T1&F&Mo | | Mo&F&Np | | T1&Mo&F&Np | | Mean | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F1 | ACC | F1 | ACC | F1 | ACC | F1 | ACC | F1 | ACC | F1 | ACC | F1 | ACC |
| w/ | w/ c | 0.606 | **0.618** | **0.657** | **0.654** | **0.680** | 0.674 | **0.675** | **0.677** | 0.722 | **0.721** | **0.734** | **0.732** | **0.595** | **0.590** |
| $\mathcal{L}_a$ | w/o c | **0.612** | 0.610 | 0.609 | 0.606 | 0.676 | 0.668 | **0.675** | 0.671 | **0.723** | 0.718 | 0.732 | 0.728 | 0.590 | 0.585 |
| w/o | w/ c | 0.600 | 0.592 | 0.620 | 0.615 | 0.657 | 0.654 | 0.653 | 0.635 | 0.674 | 0.671 | 0.661 | 0.660 | 0.582 | 0.576 |
| $\mathcal{L}_a$ | w/o c | 0.581 | 0.579 | 0.606 | 0.598 | 0.678 | **0.694** | 0.631 | 0.631 | 0.671 | 0.665 | 0.669 | 0.666 | 0.586 | 0.579 |
| w/o $\mathcal{L}_{NP}$ | | 0.578 | 0.577 | 0.623 | 0.628 | 0.679 | 0.668 | 0.631 | 0.635 | 0.720 | 0.712 | 0.665 | 0.666 | **0.595** | 0.589 |

with high oscillations, and modality collapse on T1. For the second component, Table 4.2 demonstrates that clustering reduces computational complexity without compromising performance.

### 4.4.3 Decoupled Alignment and Modality Imbalance.

Training across diverse modality combinations poses challenges due to modality imbalance, with varying convergence rates and overfitting tendencies [184]. As shown in Table 4.2, representation-level alignment via $\mathcal{L}_{\text{align}}$ significantly improves performance, especially for longer modality combinations. However, $\mathcal{L}_{\text{align}}$ alone is insufficient without task-level alignment through $\mathcal{L}_{NP}$. Figure 4.3 illustrates that $\mathcal{L}_{NP}$ enhances robustness across modality combinations, reducing scattered validation losses. Relying solely on $\mathcal{L}_{NP}$ mitigates performance gaps but risks overfitting.




(i) Our Complete Model (ii) Remove $\mathcal{L}_{align}$ (iii) Remove $\mathcal{L}_{NP}$

Figure 4.3: Loss curves with each line from a single combination.

### 4.4.4 Comparative Analysis

Given the novel setting, direct comparisons with prior work are unavailable. Instead, the proposed model was compared against flexible architectures for multimodal learning with missing modalities: Everything [2], a cross-modal alignment method using parallel Transformers, and CasAD [25], which employs cascaded Transformers for flexible fusion. Separate models were also trained for each modality combination, using 3D ResNet and late fusion for 3D volumes, FT-Transformer [180] for tabular data, and parallel Transformers for combined tabular and 3D volumes [12].

Figure 4.4 shows that the proposed model outperforms separate models as modality count increases, achieving a 5.4% higher mean accuracy than Everything for four modalities. The performance gap over separate models widened to 2.9% for four-modality combinations, while CasAD's performance declined significantly with more modalities, and Everything plateaued at top performance levels.

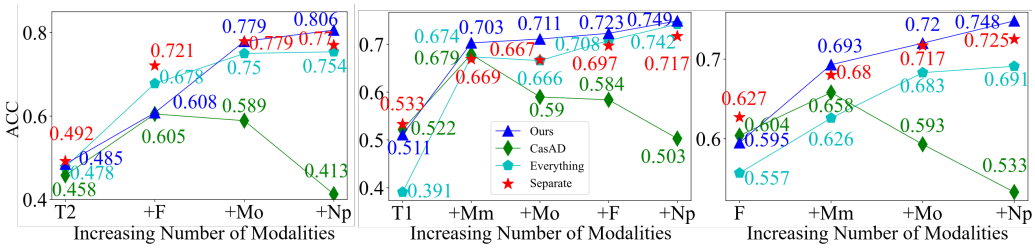



Figure 4.4: Results of comparative studies.

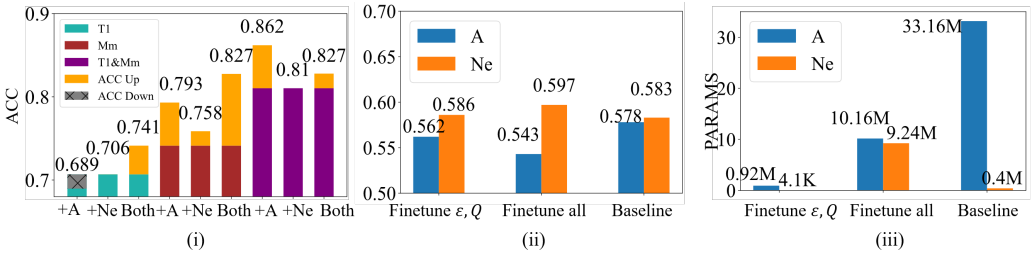



Figure 4.5: Results on new modalities and unseen combinations.

### 4.4.5 Adaptation to New Modalities

Following the training on all combinations of the six observed modalities, the two Transformers were locked in the proposed model and a small number of additional parameters $(\mathcal{E}_{m_u}, Q_{m_u})$ were introduced for the novel modalities $m_u \in \{A, Ne\}$. The second graph in Fig. 4.5 evaluates the performance differences among three approaches: updating only $(\mathcal{E}_{m_u}, Q_{m_u})$, updating the entire model, and using a baseline model specifically trained for the new modalities. Meanwhile, the third graph contrasts the parameter costs associated with these three strategies. The findings indicate that similar performance levels can be attained with considerably lower computational expenses. The first graph illustrates the proposed model's effectiveness on previously unseen modality combinations without any additional training. The performance improvements or declines resulting from incorporating A or Ne into three foundational combinations are highlighted. In most instances,

performance enhancements are indicated in orange, with universally positive outcomes when both new modalities are added. However, it is important to note that adding modalities does not guarantee performance gains in every case.

## 4.5 Discussion

The parameter efficiency of the modality-agnostic Transformer stems from its hybrid design that combines shared projection parameters with tunable query vectors. This architecture addresses the inherent trade-off between model specialization and combinatorial scalability. The tunable queries prevent modality interference observed in the shared $Q_{m_i}$) variant, particularly benefiting heterogeneous data pairs. The dual alignment strategy mitigates modality imbalance through complementary mechanisms. Representation-level alignment ensures stable gradient propagation across modalities with differing convergence rates, as evidenced by the performance improvement on four-modality combinations. Task-level alignment prevents overfitting to dominant modalities by enforcing combination-agnostic decision boundaries, reducing validation loss dispersion ompared to single-alignment configurations (Figure 4.3. This synergistic effect enables robust learning across various unique modality combinations.

In comparative studies, the CasAD model suffers from significant performance drop due to its sequential modeling of long modality combinations, which process only two modalities for each cascaded cross-attention, and is not permutation-invariant. In contrast, the growing performance advantage over separately trained models (Figure 4.4 highlights the framework's suitability for clinical environments with variable data availability. The linear scaling of parameters versus exponential

growth in traditional approaches makes the method particularly advantageous for handling diverse ($> 3$) modalities. Furthermore, the consistent performance gains when adding new modalities (Figure 4.5, orange bars) suggest emergent combinatorial benefits not explicitly trained for, which is a critical feature for real-world deployment where novel modality combinations frequently arise from evolving diagnostic protocols.

A primary strength of the proposed framework is its flexibility to incorporate diverse data modalities, which is a necessary step toward personalized medicine. However, the development of clinically equitable and generalizable AI models requires careful consideration of population heterogeneity and the compositional biases inherent in the datasets used for training.

In this work, the model was developed and evaluated on the dataset as a whole, without explicit analysis of performance across demographic subgroups such as those defined by sex, age, or comorbidity burden. This is a recognized limitation, as Alzheimer's Disease incidence and presentation can vary across populations [185]. If performance disparities were to exist across such subgroups, the real-world utility of the diagnostic tool would be compromised.

Our flexible architecture, however, provides a foundation for implementing fairness-aware mitigation strategies in future work. The modality-agnostic design could be adapted to incorporate group-specific decision thresholds at the output layer, calibrating sensitivity and specificity for different populations. Furthermore, the training procedure could be enhanced with techniques such as reweighting the loss function to balance the influence of underrepresented subgroups or employing adversarial debiasing to learn features invariant to protected attributes. Exploring these mitigation options is an essential next step to ensure the equitable application

of our model in diverse clinical populations.

## 4.6 Conclusion

This work addresses the critical challenge of unified multi-modal learning in medical imaging through a framework designed for diverse, imbalanced modalities and dynamic combination scenarios. The proposed architecture achieves three fundamental advancements: 1) Combinatorial scalability through modality-agnostic projection with tunable queries, enabling parameter-efficient adaptation while preventing modality interference; 2) Alignment stability via decoupled representation-task alignment that reduces validation loss dispersion compared to single-alignment strategies; 3) Dynamic adaptability supporting unseen combinations and new modalities through lightweight component additions, achieving comparable performance with full fine-tuning and specifically-trained baselines.

# Chapter 5

# ARDS Risk Monitoring: Integration of Asynchronous Modalities

## 5.1 Problem Background and Research Gap

Forecasting the risk of severe adverse events, marked by sudden onset and elevated mortality rates, is a pivotal use of artificial intelligence in clinical settings [186, 187]. Among these, Acute Respiratory Distress Syndrome (ARDS) poses a significant challenge within intensive care units (ICUs), particularly following the COVID-19 pandemic [153]. ARDS affects approximately 10–15% of ICU patients [188], with mortality rates approaching 40% [188, 189]. A major factor contributing to this high mortality is the frequent failure to promptly identify ARDS, which hinders timely life-saving interventions [188]. The challenge is intensified by the vast and intricate nature of real-time ICU data, which often surpasses clinicians' ability to derive actionable insights [190]. Thus, leveraging the rich, continuous, and multi-modal data streams to produce accurate and timely ARDS

risk predictions holds immense clinical potential.

Early identification of ARDS depends on integrating diverse clinical data sources, yet current methodologies remain disjointed. Traditional statistical and machine learning approaches have utilized structured electronic health record (EHR) data, focusing on vital signs (VS) and laboratory results (LAB) within fixed time frames. For example, logistic regression models analyzing EHR-derived features from the initial 24–48 hours of ICU admission achieved AUROCs of $0.78$–$0.81$, but these approaches often overlooked temporal trends and imaging data [148, 152, 191]. Chest X-rays (CXRs), which reveal lung-specific findings such as bilateral opacities, offer critical diagnostic precision but are limited by infrequent acquisition, subjective interpretation, and delayed manifestation of lung injury [150, 192]. Recent progress in deep learning has shown convolutional neural networks (CNNs) can identify subtle CXR patterns predictive of ARDS, achieving AUROCs of $0.82$–$0.85$ [149, 192]. However, these image-centric models fail to incorporate the dynamic physiological context provided by high-frequency VS and irregularly sampled LAB data. In [151], a late fusion approach attempted to integrate CXR and EHR data, but it relied on oversimplified temporal summaries, reducing time-series data to static statistics, and fixed prediction windows.

Such approaches neglect the asynchronous nature of ICU data sampling and the clinical need for continuous risk reassessment as patient conditions change. As a result, existing systems fall short in addressing two key deficiencies: 1) inadequate modeling of interdependencies among irregularly sampled multi-modal data, and 2) static, one-off predictions that do not align with the dynamic monitoring requirements of ICUs, where clinicians need regularly updated risk evaluations to inform time-critical interventions.

## 5.2 Methods

### 5.2.1 Data Sources

This research leverages retrospective datasets from two publicly accessible repositories from Beth Israel Deaconess Medical Center in Boston, MA: 1) MIMIC-IV [193] https://mimic-iv.mit.edu/, encompassing electronic health record (EHR) data from 53,130 ICU admissions spanning 2008 to 2019; and 2) MIMIC-CXR [194] https://www.physionet.org/content/mimic-cxr/2.0.0/, which includes 377,110 de-identified chest radiographs linked with free-text radiology reports collected between 2011 and 2016.

Patient eligibility was established through a series of exclusions. Initially, pediatric patients (under 18 years), admissions involving multiple ICU stays, and inter-unit transfers were removed, reducing the MIMIC-IV cohort from 53,130 to 45,127 admissions. Further filtering excluded admissions without corresponding chest radiographs in MIMIC-CXR, resulting in a final cohort of 10,056 ICU admissions.

#### 5.2.1.1 ARDS Case Definition

Following prior studies [148, 152], ARDS cases were identified using the 2012 Berlin criteria [195], which mandates mechanical ventilation for hypoxemia evaluation. However, this study adopted an updated ARDS definition [196] to align with contemporary respiratory support practices. Key modifications included: 1) incorporating high-flow nasal oxygen with a flow rate exceeding 30 L/min, and 2) defining hypoxemia using PaO2:FIO2 $< 300$ mm Hg or SpO2:FIO2 $< 315$ (for SpO2 $< 97\%$).

ARDS onset was marked as the earliest time point satisfying the hypoxemia criteria (modification 2). To exclude pre-existing ARDS and ensure sufficient baseline data, 457 admissions with ARDS onset within 6 hours of ICU admission were excluded. The final cohort consisted of 9,599 unique patients, including 463 confirmed ARDS cases.

#### 5.2.1.2 Data Pre-processing

Chest X-ray (CXR) images, irrespective of view type, underwent preprocessing steps including intensity normalization, resizing, center-cropping to $224 \times 224$ pixels, and data augmentation via random horizontal flips and affine transformations. Laboratory assessments encompassed a wide array of metabolic markers (e.g., lactic acid, glucose in serum and whole blood, anion gap, pH from arterial, venous, and urine samples, serum bicarbonate, and arterial base excess). Hematological parameters included platelet counts, leukocyte counts, hemoglobin, hematocrit, and differential counts for eosinophils, monocytes, neutrophils, basophils, and lymphocytes. Coagulation profiles covered prothrombin time, international normalized ratio, and partial thromboplastin time. Cardiac markers included creatine kinase-MB, troponin-T, arterial oxygen and carbon dioxide pressures, and total CO2 in arterial blood. Renal function was evaluated through blood urea nitrogen, serum creatinine, and urine specific gravity. Hepatic and inflammatory markers included aspartate aminotransferase, alanine aminotransferase, total bilirubin, albumin, lactate dehydrogenase, and alkaline phosphatase. Electrolyte profiles comprised sodium, chloride, ionized and non-ionized calcium, magnesium, phosphorus, and potassium in both whole blood and serum. Point-of-care glucose monitoring via fingerstick was also included. Vital signs included invasive

arterial blood pressure (systolic, diastolic, mean), noninvasive oscillometric blood pressure (systolic, diastolic, mean), core temperature (Fahrenheit, with site documentation), heart rate, rhythm characterization, and ventricular ectopy parameters. All laboratory measurements and vital signs were normalized to standard ranges [88]. The cohort was divided into training (70%) and test (30%) sets using stratified sampling to maintain outcome distribution, with results reported on the test set of 2,880 patients, including 139 ARDS cases.

### 5.2.2 Problem Formulation

**Heterogeneous Multi-Modal Data.** This study focuses on three ARDS-relevant modalities in the ICU: chest X-rays (CXRs), vital signs (VS), and laboratory results (LAB), with VS and LAB defined by a predefined parameter set $\mathcal{P}$. Unlike CXRs, which are acquired at discrete time points, VS and LAB parameters exhibit varying observation times. Vital signs are typically recorded at high frequency with regular intervals, while laboratory tests are sampled irregularly with a broader parameter range. CXRs, in contrast, are acquired sparsely and irregularly.

These heterogeneous data streams are unified under a single time coordinate, defining the asynchronous multi-modal input at prediction time $t_i$ over period $T$ as $I_{t_i} = \{D_t, t_i - T < t < t_i\}$, where each data point is denoted as $D_t^m$ for $m = \text{CXR}$, or $D_t^{m_p}, p \in \mathcal{P}$ for $m \in \{\text{VS}, \text{LAB}\}$, with $t$ indicating the observation time during ICU stay.

**Continuous Risk Monitoring.** Rather than relying on fixed data windows for ARDS diagnosis, the proposed model dynamically tracks ARDS risk through periodic predictions at 6-hour intervals. Starting at time $t_0$, the model generates

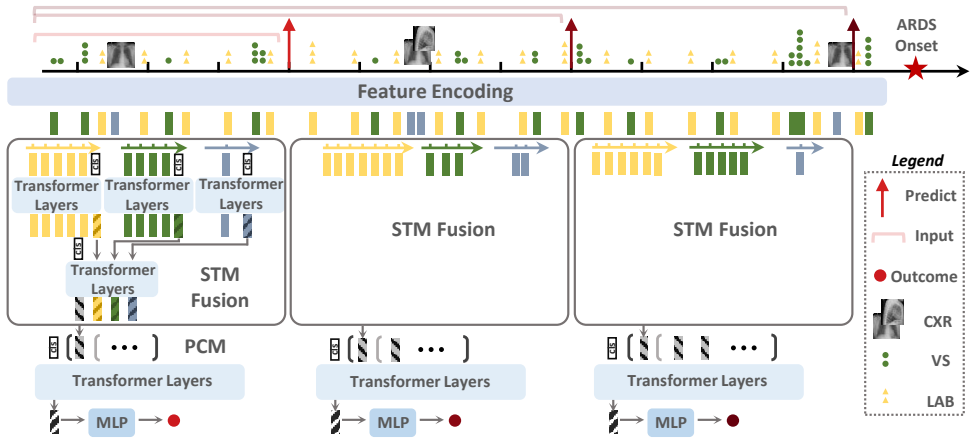



Figure 5.1: Schematic of the proposed pipeline.

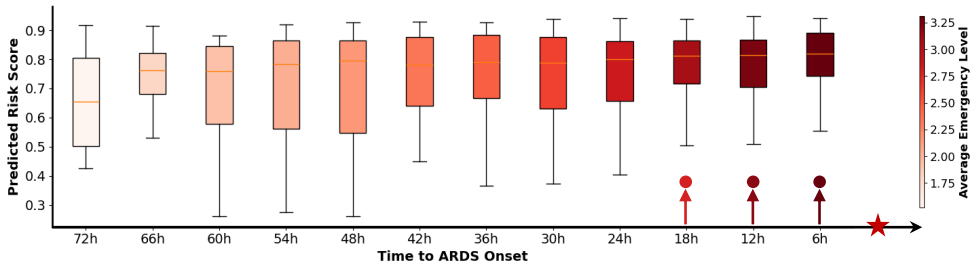



Figure 5.2: Visual representation of ARDS patient monitoring results, showing risk scores and emergency levels (color-coded).

predictions at $t_i = t_0 + i \times 6$ hours, where $i = 1, 2, \ldots$. At each $t_i$, the model processes asynchronous multi-modal data $I_{t_i}$ over a period $T = \textbf{min}(72 \text{ hours}, t_i - \text{admission time})$. Outputs include an ARDS risk score ($r \in (0, 1)$) indicating likelihood and an emergency level ($e \in \{1, 2, 3, 4\}$) denoting time to onset, guiding intervention urgency. Emergency level 4 indicates the highest urgency (onset within 12 hours), followed by level 3 (12–24 hours), level 2 (24–48 hours), and level 1 (beyond 48 hours).

### 5.2.3 Network Design

The proposed architecture comprises three integrated stages: 1) feature encoding, transforming raw data $\{D_i\}$ into unified latent features $\{f_{\tilde{t}}^m\}$; 2) feature fusion, capturing cross-modal and sequential dependencies to produce a summary feature $\hat{f}_i$; and 3) a task head, using a multilayer perceptron (MLP) to predict risk scores and emergency levels from $\hat{f}_i$.

**Feature Encoding.** Asynchronous multi-modal inputs $\{D_i\}$ are processed into unified latent features $\{f_{\tilde{t}}^m\}$ through modality-specific pipelines. CXR spatial features are extracted using established convolutional neural network (CNN) architectures with direct timestamp alignment. For tabular data (VS and LAB), an adaptive sliding window algorithm groups observations temporally. Starting from an initial timestamp $t_0$, the algorithm aggregates observations within a tolerance threshold $\tau$, avoiding duplicates by tracking marked timestamps. The window $\mathcal{W}$ includes all unmarked $t_k$ where $|t_k - t_0| \leq \tau$, with the effective timestamp set as $\tilde{t} = \min_{t_k \in \mathcal{W}} t_k$. This ensures non-redundant grouping while preserving temporal coherence.

Following [180], windowed data is embedded via linear projection $\mathbf{W}_{\text{cont}} \in \mathbb{R}^{d \times 1}$ for continuous parameters and an embedding table $\mathbf{E} \in \mathbb{R}^{d \times |\mathcal{C}|}$ for categorical values. Each parameter is represented as a $d$-dimensional token ($d = 128$ in experiments), with attention masks handling missing values in Transformer layers. A CLS token captures cross-parameter interactions, yielding modality feature $f_{\tilde{t}}^m$.

**Temporal Positional Encoding.** To maintain temporal relationships across asynchronous measurements, a modality-aware positional encoding is developed. For each feature $f_{\tilde{t}}^m$, the modality type $m$ is encoded as a categorical variable $C_m$,

and the normalized relative timestamp is computed as $R_{\tilde{t}} = (\tilde{t} - t_i)/\Delta T$ within the observation window $[t_i, t_i + \Delta T]$. These are combined via an MLP with ReLU activation to produce a joint positional encoding $P = \text{MLP}(R_{\tilde{t}}, C_m) \in \mathbb{R}^d$. This encoding is added element-wise to $f_{\tilde{t}}^m$, preserving both temporal and modality-specific context while mitigating synchronization issues from fixed-interval aggregation.

**Feature Integration.** Modeling interactions across asynchronous modalities and sequential predictions is addressed through a hierarchical fusion framework using Transformer architectures [39, 197]. Two specialized mechanisms enhance efficiency: Staged Temporal-Modal Fusion (STM Fusion) and Progressive Context Memory (PCM).

STM Fusion processes $\{f_{\tilde{t}}^m\}$ in two phases: modality-specific temporal Transformer layers compress each modality's irregular sequence into a summary vector using modality-aware positional encodings, followed by cross-modal fusion via multimodal Transformer layers with modality embeddings. This reduces computational complexity from $O((MT)^2)$ to $O(M \cdot T^2)$, minimizing interference between modalities.

PCM addresses sequential prediction by modeling dependencies across prediction times $\{t_0, \ldots, t_i\}$. Instead of processing all historical features, PCM uses: 1) Incremental Encoding, processing only new observations since $t_{i-1}$:

$$\Delta \hat{f}_i = \text{STM-Fusion}\left(\{f_t : \mid t_{i-1} < t \leq t_i\}\right), \tag{5.1}$$

and 2) Memory-Augmented Attention, maintaining a memory bank $\mathcal{M} = \{\hat{f}_0, \ldots, \Delta\hat{f}_{i-1}\}$ updated with $\hat{f}_i$. A Transformer layer over $\mathcal{M}$ learns attention weights for con-

text aggregation, reducing temporal complexity to $O(L^2 + MT^2)$ while preserving long-range dependencies.

### 5.2.4 Training Strategy

The training process tackles irregular sampling and class imbalance through late batching and balanced sampling.

**Late Batching.** To manage irregular temporal sequences without information leakage, a patient-centric late batching protocol processes one patient's data per batch, sequentially handling all prediction time points $\{t_0, \ldots, t_L\}$. Gradient accumulation over $B$ patients ensures stable updates:

$$\theta \leftarrow \theta - \eta \frac{1}{B} \sum_{k=1}^{B} \nabla_\theta \mathcal{L}(I_{t_0:t_L}^{(k)}), \tag{5.2}$$

where $I_{t_0:t_L}^{(k)}$ represents temporal samples for patient $k$.

**Balanced Sampling.** A balanced sampler addresses class imbalance by selecting half the batch from positive and negative classes, shuffling them to ensure equal representation. Implemented as a dynamic generator, it maintains dataset integrity while optimizing for imbalanced learning.

### 5.2.5 Training Implementation

The model uses cross-entropy loss for risk scores and MSE loss for emergency levels, equally weighted. Optimization employs AdamW with an initial learning rate of 3e-4, using cosine annealing with linear warm-up. Training ran for 20,000 steps with a batch size of 16, selecting the final model based on balanced perfor-

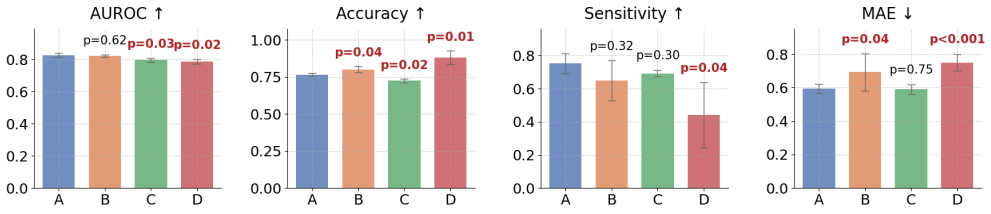



Figure 5.3: Performance comparison of fusion strategies, with notable differences emphasized. **A**: with both STM and PCM; **B**: without PCM; **C**: without STM; **D**: without both.

Table 5.1: Impact of training strategies on model performance.

| **Training Method** | | Slope↑ | Volatility↓ | AUROC% ↑ | ACC% ↑ | SEN% ↑ |
|---|---|---|---|---|---|---|
| Late Batching | Balanced Sampling | | | | | |
| ✗ | ✗ | 5.48e-6 | 0.089 | 70.62 | 95.45 | 0 |
| ✓ | ✗ | 2.06e-5 | 0.078 | 78.32 | 95.45 | 0 |
| ✗ | ✓ | 3.76e-5 | 0.070 | 75.52 | 72.21 | 59.12 |
| ✓ | ✓ | 7.41e-4 | 0.046 | 82.69 | 76.55 | 75.20 |

mance on the validation set. Experiments were conducted using PyTorch on an NVIDIA RTX 3090 Ti GPU with 24GB memory.

Table 5.2: Performance evaluation of asynchronous modality integration.

| **Modalities** | | AUROC% ↑ | ACC% ↑ | SEN% ↑ | SPE% ↑ | MAE↓ | MSE↓ |
|---|---|---|---|---|---|---|---|
| CXR | EHR | | | | | | |
| ✓ | ✗ | 73.71±1.73 | 79.92±1.17 | 45.38±2.34 | 82.15±1.18 | 1.12±0.03 | 1.85±0.11 |
| ✗ | ✓ | 79.03±1.19 | 78.04±0.96 | 67.03±1.80 | 79.13±0.98 | 1.26±0.04 | 2.42±0.12 |
| ✓ | ✓ | 83.26±0.96 | 76.43±0.29 | 73.70±0.20 | 76.53±0.08 | 0.58±0.01 | 0.70±0.02 |

## 5.3 Results

### 5.3.1 Design Choices Evaluation

**Fusion Strategies.**

The proposed fusion architecture was evaluated by comparing: (1) STM fusion against interleaved fusion on concatenated asynchronous multi-modal tokens,

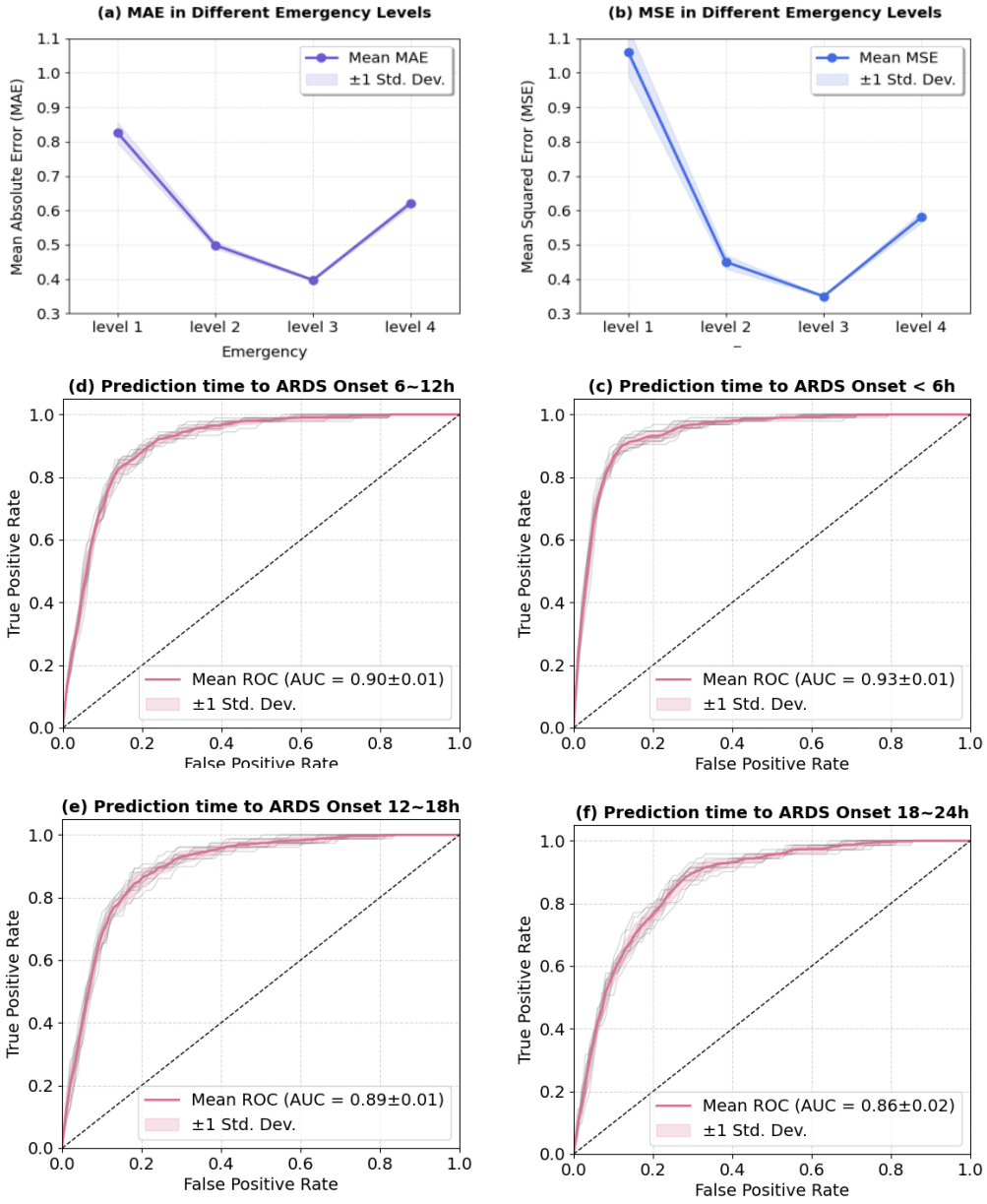



Figure 5.4: Detailed assessment of continuous risk monitoring performance.

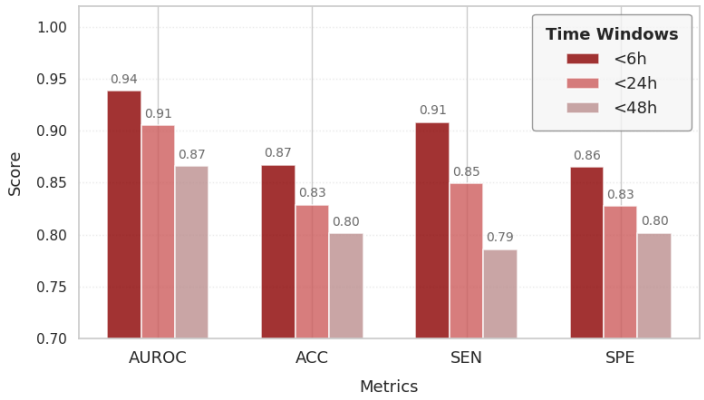



Figure 5.5: Per-prediction performance across different time intervals.

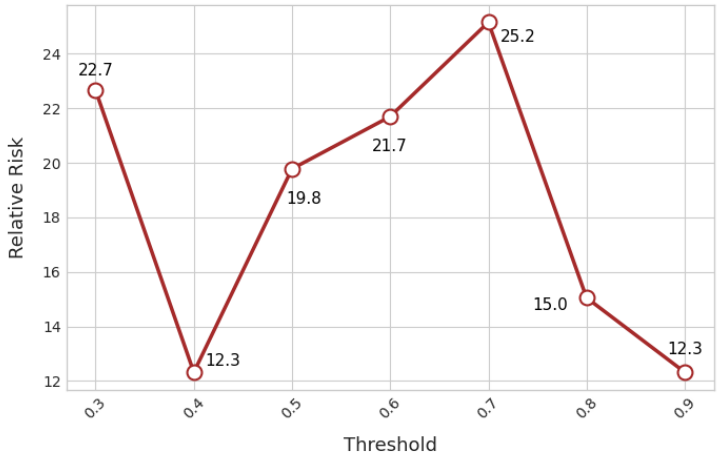



Figure 5.6: Relative risk trend analysis.

and (2) PCM-augmented attention against a baseline processing all historical data without discrimination. Performance metrics included Accuracy (ACC), Sensitivity (SEN), AUROC, and Mean Absolute Error (MAE) for emergency level predictions, averaged over sequential prediction points with 3-fold cross-validation and a fixed threshold of 0.5. As depicted in Fig. 5.3, STM significantly enhances AUROC, while PCM improves emergency level prediction accuracy and reduces variability, likely due to its context-aware modeling. The combination of STM and PCM markedly improves sensitivity for ARDS cases, critical for risk prediction. Without these fusion strategies, baseline **D** achieves high ACC but significantly lower SEN under class imbalance conditions.

**Training Strategy Effectiveness.** Ablation studies validated the necessity of tailored training approaches using absolute loss slope, volatility (relative mean absolute change), and diagnostic metrics. Without balanced sampling, the negative class overshadows predictions, resulting in low variance and near-zero outputs. Late batching alone boosts AUROC by approximately $+8\%$ by minimizing excessive padding for irregular data. Combined, these strategies ensure robust training with significant improvements in diagnostic performance (Tab. 5.1).

**Asynchronous Modality Integration.** Table 5.2 shows that EHR data outperforms CXR in risk prediction due to its sensitivity to early physiological changes, while CXR slightly excels in emergency prediction for positive cases through post-symptom imaging cues. Their integration demonstrates substantial complementary benefits, enhancing performance in this complex task.

### 5.3.2 Dynamic Risk Monitoring Performance

**Risk Score Diagnostic Accuracy.** Figure 5.2 visually illustrates monitoring outcomes. Further evaluation of temporal prediction performance is provided via ROC curves for risk scores across 6-hour windows within 24 hours before ARDS onset (Fig. 5.4(c-f)). Predictions were generated from randomly sampled time points (negative cases: random windows; positive cases: pre-onset windows) over 10 experiments. AUROC gradually declined from 0.94 ($< 6h$) to 0.87 ($18$-$24h$), reflecting decreasing clinical urgency.

Aggregated performance over broader intervals (Fig. 5.5) achieved AUROCs of 0.92 ($< 24h$) and 0.88 ($< 48h$) across 10 repeated samplings, surpassing prior studies (AUROC: $0.78$-$0.85$).

**Emergency Level Prediction Accuracy.** The model's emergency level predictions escalate as ARDS onset nears (Fig. 5.2). Stratification by true urgency levels (Fig. 5.4a-b) shows Level 1 ($> 48h$) predictions maintain acceptable error margins ($< 1$ level), preventing resource misallocation. Levels 2–3 demonstrate improved accuracy but tend to cluster in mid-range values (predicted: 1.5–3.3 vs. true: 1–4; Fig. 5.2). This conservative bias in predicting extreme levels suggests a need for recalibration to improve discrimination at critical urgency thresholds.

**Risk Stratification Analysis.** Patient-level risk stratification compared ARDS incidence between high-risk (threshold-exceeding) and low-risk groups using relative risk (RR), calculated as the ratio of ARDS incidence in the high-risk group to the low-risk group. Figure 5.6 shows RR values consistently above 12 across thresholds ($0.3$–$0.9$), indicating robust stratification. A threshold of 0.4 marks a shift from low-risk to high-risk dominance. Peak performance occurs at a thresh-

old of 0.7 (RR$= 25.83$), with stable RRs ($19.92$–$25.83$) in the moderate range ($0.5$–$0.7$), effectively identifying high-risk patients for efficient resource allocation.

## 5.4 Discussion

The STM-PCM fusion approach highlights the importance of specialized architectures for handling asynchronous, multi-modal longitudinal data. STM's AUROC improvements likely arise from its ability to capture temporal dependencies in token sequences, while PCM's context-aware attention reduces noise in sparse clinical data. Late batching and balanced sampling address key ICU data challenges: irregular sampling and class imbalance. These design choices collectively enhance model stability and sensitivity, overcoming limitations of naive baselines (e.g., high accuracy but low sensitivity in **D**). The complementary strengths of EHR (temporal dynamics) and CXR (pathological specificity) support multi-modal integration, aligning with clinical workflows where acute changes and historical context guide decisions.

The model's high AUROC ($0.94$ near onset) and robust RR values ($> 12$) demonstrate reliable risk stratification, outperforming existing methods and supporting its potential as an early-warning system, particularly within the critical 6-hour window.

The sharp RR drop at a threshold of 0.4 reflects a transition from low-risk to high-risk dominance. At 0.3, the low-risk group is small (32%) but highly accurate, leading to low incidence and high RR. At 0.4, the low-risk group expands but becomes less accurate, while the high-risk group remains insufficiently precise,

causing the RR drop. Beyond 0.5, the high-risk group dominates with greater accuracy.

The conservative bias in emergency level predictions (underestimating extremes) is a notable limitation. While this may prevent over-triage in resource-constrained settings, it risks delaying interventions for rapidly deteriorating patients. This bias may stem from loss function asymmetry or underrepresentation of extreme cases, necessitating recalibration to enhance discrimination at critical thresholds.

This study did not explicitly assess performance across demographic or clinical subgroups (e.g., by sex, age, or comorbidities), which is critical for equitable AI application. The MIMIC-IV and MIMIC-CXR datasets, sourced from a single center, may limit generalizability due to demographic constraints. Factors such as sex-specific ARDS presentation, age-related lung compliance changes, or comorbidities (e.g., chronic obstructive pulmonary disease) could impact performance [198]. Future validation on diverse cohorts is needed to identify potential disparities. Mitigation strategies include reweighting training data for underrepresented groups or applying group-specific thresholds to ensure equitable sensitivity and specificity.

## 5.5 Conclusion

This study advances ARDS risk prediction by introducing continuous monitoring using asynchronous multi-modal ICU data. Tailored hierarchical fusion and training strategies, validated through ablation studies, achieve significantly higher AUROCs than prior work. The model's potential for early ARDS detection and

resource prioritization is supported by robust emergency level predictions and effective patient risk stratification, offering substantial clinical value.

# Chapter 6

# Conclusion and Future Work

## 6.1 Conclusion

This thesis establishes unified frameworks for flexible modality integration in medical multimodal learning, systematically addressing the critical limitations of conventional multimodal paradigms constrained by static modality availability. By resolving domain-specific challenges across three clinically consequential applications, the research demonstrates that robust adaptation to dynamic modality subsets and asynchronous sequences is both technically feasible and essential for real-world clinical deployment.

The selection of these three applications, which are multimodal MRI synthesis, Alzheimer's diagnosis with heterogeneous data, and dynamic ARDS risk monitoring, is deliberate and representative. They collectively span the two dominant medical AI paradigms: (1) dense prediction tasks (exemplified by MRI synthesis), demanding pixel-level anatomical fidelity for spatially aligned multimodal images; and (2) decision-centric tasks (illustrated by AD diagnosis and ARDS

monitoring), requiring integration of semantically misaligned or temporally irregular data for clinical decision support. Furthermore, they encapsulate the three cardinal challenges of flexible integration: (i) multimodal image fusion (MRI sequences with spatial misalignment and modality-specific structures), (ii) heterogeneous modality fusion (non-imaging tabular data with neuroimaging in AD diagnosis, which is diverse and imbalanced), and (iii) asynchronous longitudinal fusion (sparse imaging and high-frequency physiological streams in ICU monitoring). These applications were chosen not only for their clinical prevalence but also for their distinct technical demands, providing rigorous testbeds for evaluating generalized solutions.

In multimodal MRI synthesis, the integration of fine-grained difference learning with multi-scale deformable convolutions reconciles the historical fragmentation between cross-modality synthesis and super-resolution. This approach achieves consistent high-fidelity reconstruction under extreme degradation (2–16$\times$ undersampling) while preserving clinically critical modality-specific structures, overcoming spatial misalignment through feature-level warping. For Alzheimer's diagnosis, the AnyMod architecture introduces representation-task decoupled alignment to handle arbitrary combinations of imaging and non-imaging data. By mapping heterogeneous inputs to class-invariant prototypes via modality-agnostic Transformer projectors and dynamic token clustering, the framework preserves task-specific semantics while ensuring computational scalability, demonstrating increasing performance advantages as modality count grows. In dynamic ARDS risk monitoring, modality-wise encoding with adaptive spatiotemporal embeddings and staged temporal-modal fusion resolves the asynchronicity between sparse chest X-rays and high-frequency vital signs. The progressive context memory mech-

anism enables efficient long-range dependency modeling, delivering actionable hourly risk stratification and precise time-to-onset quantification critical for ICU workflows.

Empirical validation across public datasets (IXI, BraTS, ADNI, MIMIC-IV) confirms consistent superiority with significant gains in reconstruction fidelity, diagnostic accuracy, and prognostic precision over state-of-the-art methods. Through integration of dynamic adaptation, task-driven alignment, and anatomical/temporal prior exploitation, this work advances medical AI toward clinically resilient and data-efficient multimodal learning.

### 6.1.1 Future Work

While the models were validated on established public datasets, their generalizability to new clinical environments with different patient demographics, clinical protocols, and data acquisition systems requires further investigation. For instance, models trained on ADNI, which has known underrepresentation of diverse ethnic groups, may not perform equitably across global populations. Similarly, models developed on MIMIC-IV's single-institution data may be susceptible to site-specific biases in clinical practice. Future work should include explicit sensitivity analyses and subgroup assessments across race, ethnicity, and hospital sites to quantify and mitigate these biases, ensuring model fairness. Furthermore, for operationally-focused models like the ARDS risk monitor, the goal should not always be a single generalizable model, but the ability to share 'recipes' for effective local retraining and validation.

Furthermore, to bridge the gap between high performance and clinical trust,

future work must prioritize explainable artificial intelligence (XAI). While our model demonstrates strong predictive accuracy, its"black-box" nature remains a limitation for clinical adoption. Future iterations should incorporate XAI techniques, such as saliency maps for neuroimaging modalities and Shapley values for tabular data, to visualize which features and regions most influenced the diagnosis. This is not just for transparency; it is a crucial step for validating that our model's decisions are based on biologically plausible markers like hippocampal atrophy or specific pathological signatures. For a system as flexible as ours, designed to handle any modality combination, an advanced framework like an "XAI Orchestrator" could be developed to dynamically manage and generate coherent explanations across different data types and time points, making the AI a more interpretable partner for clinicians.

There are several possible enhancements to further improve the proposed algorithms. For MRI synthesis, generalizing the current dual-modality approach to handle any number of input modalities will address clinical needs for comprehensive protocols. This requires mechanisms to efficiently combine anatomical priors from multiple references while managing computational complexity. For Alzheimer's diagnosis, introducing performance guarantees for different modality subsets will ensure reliable deployment, and integrating longitudinal multimodal trajectories to model preclinical disease progression will enable earlier disease risk prediction. The ARDS risk prediction framework would benefit from quantifying prediction uncertainty, allowing alerts to dynamically balance early warning sensitivity and false alarms based on clinical urgency.

In addition to application-specific improvements, three overarching frontiers are worth investigating. First, Integrating medical foundation models could enable

zero-shot adaptation to novel modalities by projecting them into unified semantic spaces. Coupled with multi-task optimization, where shared encoders maintain combinatorial flexibility while task-specific heads preserve diagnostic precision, this approach would create a scalable backbone for diverse clinical scenarios. Crucially, lightweight adapters must mitigate domain shifts to ensure robust deployment. Another promising direction the development of unified medical vision-language models. Such a model could use a large language model (LLM) as a core reasoning engine, while our proposed modality-agnostic projectors serve as adapters, translating diverse medical data modalities—be it an MRI scan, a tabular lab value, or a time-series vital sign—into a shared semantic space understandable by the LLM. This would enable zero-shot or few-shot adaptation to novel diagnostic tasks and modality combinations, while possibly enhance interpretability through the reasoning process. The key research challenge would be to develop efficient fine-tuning techniques, such as Low-Rank Adaptation (LoRA), to mitigate domain shift and align these general-purpose models with the precise requirements of medical diagnosis without catastrophic forgetting.

Second, AI agent ecosystems could operationalize flexible multimodal learning in clinical workflows. For instance, dense prediction outputs might dynamically trigger decision-centric agents for risk assessment, which in turn request additional tests based on uncertainty thresholds. Such chained reasoning, where agents iteratively gather modalities to resolve ambiguities, would embody the full potential of adaptive integration, transforming static AI tools into collaborative clinical partners.

Lastly, quantifying multimodal prediction reliability, particularly for critical decisions, would directly enhance clinical trustworthiness. Beyond flagging high-

risk cases (e.g., low-confidence ARDS alerts), uncertainty metrics could actively guide stepwise modality acquisition: ambiguous predictions trigger requests for additional tests, mirroring clinical workflows. This closed-loop interaction would dynamically demonstrate flexible integration's value in optimizing diagnostic resource utilization.

Building on uncertainty quantification, the next critical step is prospective validation in real-world longitudinal settings. Our current validation, while robust, is retrospective. A powerful method to demonstrate long-term predictive validity is time-to-event analysis (or survival analysis). This framework is ideal for modeling the time until a critical event, such as the conversion from Mild Cognitive Impairment (MCI) to Alzheimer's Disease . By applying models like the Cox proportional hazards model, we could not only predict who will convert but also estimate the hazard ratio and time-to-conversion for different patient subgroups, providing a dynamic and clinically highly relevant risk assessment. To operationalize this, we recommend a prospective cohort study design, such as Trials within Cohorts (TwiCs), where our model could be embedded into an ongoing observational study. Patients' data would be processed in real-time, and predictions delivered to clinicians, allowing for a direct evaluation of the model's utility in optimizing clinical workflows and improving patient outcomes .

# References


[1] Jiawen Yao et al. “Deep correlational learning for survival prediction from multi-modality data”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 406–414.

[2] Nina Shvetsova et al. “Everything at once-multi-modal fusion transformer for video retrieval”. In: *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 2022, pp. 20020–20029.

[3] Jiali Duan et al. “Multi-modal alignment using representation codebook”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 15651–15660.

[4] Yi-Lun Lee et al. “Multimodal Prompting with Missing Modalities for Visual Recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 14943–14952.

[5] Shangran Qiu et al. “Multimodal deep learning for Alzheimer’s disease dementia assessment”. In: *Nature communications* 13.1 (2022), p. 3404.

[6] Muhammad Adeel Azam et al. “A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases,

fusion techniques and quality metrics". In: *Computers in biology and medicine* 144 (2022), p. 105253.

[7] Julián N Acosta et al. "Multimodal biomedical AI". In: *Nature medicine* 28.9 (2022), pp. 1773–1784.

[8] Robert Kaczmarczyk et al. "Evaluating multimodal AI in medical diagnostics". In: *npj Digital Medicine* 7.1 (2024), p. 205.

[9] Yechong Huang et al. "Diagnosis of Alzheimer's Disease via Multi-Modality 3D Convolutional Neural Network". In: *Frontiers in Neuroscience* 13 (2019).

[10] Juan Song et al. "An effective multimodal image fusion method using MRI and PET for Alzheimer's disease diagnosis". In: *Frontiers in digital health* 3 (2021), p. 637386.

[11] Yue Tu et al. "Alzheimer's disease diagnosis via multimodal feature fusion". In: *Computers in Biology and Medicine* 148 (2022), p. 105901.

[12] Luoyao Kang et al. "Visual-attribute prompt learning for progressive mild cognitive impairment prediction". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 547–557.

[13] Rong Zhou et al. "Attentive deep canonical correlation analysis for diagnosing alzheimer's disease using multimodal imaging genetics". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 681–691.

[14] Tao Zhou et al. “Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis”. In: *Human brain mapping* 40.3 (2019), pp. 1001–1016.

[15] Reza Azad et al. “Addressing missing modality challenges in MRI images: A comprehensive review”. In: *Computational Visual Media* 11.2 (2025), pp. 241–268.

[16] Girish Katti, Syeda Arshiya Ara, and Ayesha Shireen. “Magnetic resonance imaging (MRI)–A review”. In: *International journal of dental clinics* 3.1 (2011), pp. 65–70.

[17] Chun-Mei Feng et al. “Multi-Contrast MRI Super-Resolution via a Multi-Stage Integration Network”. In: *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. 2021.

[18] Lingke Kong et al. “Breaking the dilemma of medical image-to-image translation”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 1964–1978.

[19] Juan Eugenio Iglesias et al. “Joint super-resolution and synthesis of 1 mm isotropic MP-RAGE volumes from clinical MRI exams with scans of different orientation, resolution and contrast”. In: *Neuroimage* 237 (2021), p. 118206.

[20] Kai Xuan et al. “Multimodal MRI Reconstruction Assisted With Spatial Alignment Network”. In: *IEEE Transactions on Medical Imaging* 41.9 (2022), pp. 2499–2509.

[21] Joel Honkamaa et al. “Deformation equivariant cross-modality image synthesis with paired non-aligned training data”. In: *Medical Image Analysis* (2023), p. 102940.

[22] Balint Kovacs et al. “Addressing image misalignments in multi-parametric prostate MRI for enhanced computer-aided diagnosis of prostate cancer”. In: *Scientific Reports* 13.1 (2023), p. 19805.

[23] DR Warakaulle and P Anslow. “Differential diagnosis of intracranial lesions with high signal on T1 or low signal on T2-weighted MRI”. In: *Clinical radiology* 58.12 (2003), pp. 922–933.

[24] Peng Wang et al. “ONE-PEACE: Exploring One General Representation Model Toward Unlimited Modalities”. In: *arXiv preprint arXiv:2305.11172* (2023).

[25] Linfeng Liu et al. “Cascaded multi-modal mixing transformers for alzheimer’s disease classification with incomplete data”. In: *NeuroImage* 277 (2023), p. 120267.

[26] Yuxuan Wang et al. “Deep time series models: A comprehensive survey and benchmark”. In: *arXiv preprint arXiv:2407.13278* (2024).

[27] Soujanya Poria et al. “Convolutional MKL based multimodal emotion recognition and sentiment analysis”. In: *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE. 2016, pp. 439–448.

[28] Amir Zadeh et al. “Tensor fusion network for multimodal sentiment analysis”. In: *arXiv preprint arXiv:1707.07250* (2017).

[29] Zhun Liu et al. “Efficient low-rank multimodal fusion with modality-specific factors”. In: *arXiv preprint arXiv:1806.00064* (2018).

[30] Mengmeng Ma et al. “Smil: Multimodal learning with severely missing modality”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 3. 2021, pp. 2302–2310.

[31] Yao-Hung Hubert Tsai et al. “Learning factorized multimodal representations”. In: *arXiv preprint arXiv:1806.06176* (2018).

[32] Jiayi Chen and Aidong Zhang. “Hgmf: heterogeneous graph-based fusion for multimodal data with incompleteness”. In: *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2020, pp. 1295–1305.

[33] Tao Jin et al. “Rethinking missing modality learning from a decoding perspective”. In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, pp. 4431–4439.

[34] Yuanhuiyi Lyu et al. “Omnibind: Teach to build unequal-scale modality interaction for omni-bind of all”. In: *arXiv preprint arXiv:2405.16108* (2024).

[35] Chaohe Zhang et al. “M3care: Learning with missing modalities in multimodal healthcare data”. In: *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*. 2022, pp. 2418–2428.

[36] Lei Yuan et al. “Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data”. In: *NeuroImage* 61.3 (2012), pp. 622–632.

[37] Mohammad Havaei et al. “Hemis: Hetero-modal image segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer. 2016, pp. 469–477.

[38] Yunhua Zhang, Hazel Doughty, and Cees Snoek. “Learning unseen modality interaction”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 54716–54726.

[39] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[40] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[41] Mengmeng Ma et al. “Are multimodal transformers robust to missing modality?” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 18177–18186.

[42] Gijs Van Tulder and Marleen de Bruijne. “Why does synthesized data improve multi-sequence classification?” In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2015, pp. 531–538.

[43] Matthias Hofmann et al. “MRI-based attenuation correction for PET/MRI: a novel approach combining pattern recognition and atlas registration”. In: *Journal of nuclear medicine* 49.11 (2008), pp. 1875–1883.

[44] T Varsavsky et al. *PIMMS: permutation invariant multi-modal segmentation, CoRR, vol. abs/1807.06537 (2018)*. 1807.

[45] MRI Missing Brain. "RS-Net: Regression-Segmentation 3D CNN for Synthesis of Full Resolution". In: *Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings*. Vol. 11037. Springer. 2018, p. 119.

[46] Yan Shen and Mingchen Gao. "Brain tumor segmentation on MRI with missing modalities". In: *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*. Springer. 2019, pp. 417–428.

[47] Reuben Dorent et al. "Hetero-modal variational encoder-decoder for joint modality completion and segmentation". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer. 2019, pp. 74–82.

[48] Yixin Wang et al. "Acn: Adversarial co-training network for brain tumor segmentation with missing modalities". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24*. Springer. 2021, pp. 410–420.

[49] Sanaz Karimijafarbigloo et al. "Mmcformer: Missing modality compensation transformer for brain tumor segmentation". In: *Medical imaging with deep learning*. PMLR. 2024, pp. 1144–1162.

[50] Tristan Sylvain et al. "Cross-modal information maximization for medical imaging: Cmim". In: *arXiv preprint arXiv:2010.10593* (2020).

[51] Tongxue Zhou et al. “Conditional generator and multi-sourcecorrelation guided brain tumor segmentation with missing MR modalities”. In: *arXiv preprint arXiv:2105.13013* (2021).

[52] Tongxue Zhou et al. “Latent correlation representation learning for brain tumor segmentation with missing MRI modalities”. In: *IEEE Transactions on Image Processing* 30 (2021), pp. 4263–4274.

[53] Anmol Sharma and Ghassan Hamarneh. “Missing MRI pulse sequence synthesis using multi-modal generative adversarial network”. In: *IEEE transactions on medical imaging* 39.4 (2019), pp. 1170–1183.

[54] Biting Yu et al. “3D cGAN based cross-modality MR image synthesis for brain tumor segmentation”. In: *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE. 2018, pp. 626–630.

[55] Bing Cao et al. “Auto-GAN: self-supervised collaborative learning for medical image synthesis”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 07. 2020, pp. 10486–10493.

[56] Hongwei Li et al. “DiamondGAN: unified multi-modal generative adversarial networks for MRI sequences synthesis”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*. Springer. 2019, pp. 795–803.

[57] Shuwei Qian and Chongjun Wang. “COM: Contrastive Masked-attention model for incomplete multimodal learning”. In: *Neural Networks* 162 (2023), pp. 443–455.

[58] Yue Zhang et al. “Unified multi-modal image synthesis for missing modality imputation”. In: *IEEE Transactions on Medical Imaging* (2024).

[59] Reza Azad, Nika Khosravi, and Dorit Merhof. “SMU-Net: Style matching U-Net for brain tumor segmentation with missing modalities”. In: *International conference on medical imaging with deep learning*. PMLR. 2022, pp. 48–62.

[60] Saverio Vadacchino et al. “Had-net: A hierarchical adversarial knowledge distillation network for improved enhanced tumour segmentation without post-contrast images”. In: *Medical imaging with deep learning*. PMLR. 2021, pp. 787–801.

[61] Qi Wang et al. “Multimodal learning with incomplete modalities by knowledge distillation”. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 1828–1838.

[62] Paul Pu Liang et al. “High-modality multimodal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning”. In: *arXiv preprint arXiv:2203.01311* (2022).

[63] Youngjin Yoo et al. “Deep learning of brain lesion patterns and user-defined clinical and MRI features for predicting conversion to multiple sclerosis from clinically isolated syndrome”. In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 7.3 (2019), pp. 250–259.

[64] Can Cui et al. “Survival prediction of brain cancer with incomplete radiology, pathology, genomic, and demographic data”. In: *International Con-*

*ference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 626–635.

[65] Gregory Holste et al. "End-to-end learning of fused image and non-image features for improved breast cancer classification from mri". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 3294–3303.

[66] Stefan Schulz et al. "Multimodal deep learning for prognosis prediction in renal cancer". In: *Frontiers in oncology* 11 (2021), p. 788740.

[67] Sarah Parisot et al. "Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease". In: *Medical image analysis* 48 (2018), pp. 117–130.

[68] Menglin Cao et al. "Using DeepGCN to identify the autism spectrum disorder from multi-site resting-state data". In: *Biomedical Signal Processing and Control* 70 (2021), p. 103015.

[69] Gan Cai et al. "A multimodal transformer to fuse images and metadata for skin disease classification". In: *The Visual Computer* 39.7 (2023), pp. 2781–2793.

[70] Hui Cui et al. "Co-graph attention reasoning based imaging and clinical features integration for lymph node metastasis prediction". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*. Springer. 2021, pp. 657–666.

[71] Sebastian Pölsterl, Tom Nuno Wolf, and Christian Wachinger. "Combining 3d image and tabular data via the dynamic affine feature map transform". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*. Springer. 2021, pp. 688–698.

[72] Shih-Cheng Huang et al. "Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection". In: *Scientific reports* 10.1 (2020), p. 22147.

[73] Jeremy Kawahara et al. "Seven-point checklist and skin lesion classification using multitask multimodal neural nets". In: *IEEE journal of biomedical and health informatics* 23.2 (2018), pp. 538–546.

[74] Changhee Lee and Mihaela Van der Schaar. "A variational information bottleneck approach to multi-omics data integration". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 1513–1521.

[75] Hongzhi Wang, Vaishnavi Subramanian, and Tanveer Syeda-Mahmood. "Modeling uncertainty in multi-modal fusion for lung cancer survival analysis". In: *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*. IEEE. 2021, pp. 1169–1172.

[76] Zilin Lu, Mengkang Lu, and Yong Xia. "M 2 f: A multi-modal and multi-task fusion network for glioma diagnosis and prognosis". In: *International workshop on multiscale multimodal medical imaging*. Springer. 2022, pp. 1–10.

[77] Pooya Mobadersany et al. “Predicting cancer outcomes from histology and genomics using convolutional networks”. In: *Proceedings of the National Academy of Sciences* 115.13 (2018), E2970–E2979.

[78] Jordan Yap, William Yolland, and Philipp Tschandl. “Multimodal skin lesion classification using deep learning”. In: *Experimental dermatology* 27.11 (2018), pp. 1261–1267.

[79] Ming Hou et al. “Deep multimodal multilinear fusion with high-order polynomial pooling”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[80] Richard J Chen et al. “Multimodal co-attention transformer for survival prediction in gigapixel whole slide images”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 4015–4025.

[81] Nathaniel Braman et al. “Deep orthogonal fusion: multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*. Springer. 2021, pp. 667–677.

[82] Hang Li et al. “Multi-modal multi-instance learning using weakly correlated histopathological images and tabular clinical information”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*. Springer. 2021, pp. 529–539.

[83] Hongyi Duanmu et al. “Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using deep learning with integrative imaging, molecular and demographic data”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*. Springer. 2020, pp. 242–252.

[84] Yulu Guan et al. “Predicting esophageal fistula risks using a multimodal self-attention network”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*. Springer. 2021, pp. 721–730.

[85] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019, pp. 4171–4186.

[86] Grzegorz Jacenków, Alison Q O’Neil, and Sotirios A Tsaftaris. “Indication as prior knowledge for multimodal disease classification in chest radiographs with transformers”. In: *2022 IEEE 19th international symposium on biomedical imaging (ISBI)*. IEEE. 2022, pp. 1–5.

[87] Yunxiang Li et al. “Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge”. In: *Cureus* 15.6 (2023).

[88] Nasir Hayat, Krzysztof J Geras, and Farah E Shamout. "MedFuse: Multimodal fusion with clinical time-series data and chest X-ray images". In: *Machine Learning for Healthcare Conference*. PMLR. 2022, pp. 479–503.

[89] Lei Li et al. "Multi-modality cardiac image computing: A survey". In: *Medical image analysis* 88 (2023), p. 102869.

[90] Cheng Chen et al. "Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*. Springer. 2019, pp. 447–456.

[91] Hu Wang et al. "Multi-Modal Learning With Missing Modality via Shared-Specific Feature Modelling". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 15878–15887.

[92] Amod Jog et al. "Random forest regression for magnetic resonance image synthesis". In: *Medical image analysis* 35 (2017), pp. 475–488.

[93] Thomas Joyce, Agisilaos Chartsias, and Sotirios A Tsaftaris. "Robust multimodal MR image synthesis". In: *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*. Springer. 2017, pp. 347–355.

[94] Phillip Isola et al. "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.

[95] Biting Yu et al. “Ea-GANs: edge-aware generative adversarial networks for cross-modality MR image synthesis”. In: *IEEE transactions on medical imaging* 38.7 (2019), pp. 1750–1762.

[96] Biting Yu et al. “Sample-adaptive GANs: linking global and local mappings for cross-modality MR image synthesis”. In: *IEEE transactions on medical imaging* 39.7 (2020), pp. 2339–2350.

[97] Qianye Yang et al. “MRI cross-modality image-to-image translation”. In: *Scientific reports* 10.1 (2020), p. 3753.

[98] Jun-Yan Zhu et al. “Unpaired image-to-image translation using cycle-consistent adversarial networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.

[99] Salman UH Dar et al. “Image synthesis in multi-contrast MRI with conditional generative adversarial networks”. In: *IEEE transactions on medical imaging* 38.10 (2019), pp. 2375–2388.

[100] Zhiwei Qin et al. “Style transfer in conditional GANs for cross-modality synthesis of brain magnetic resonance images”. In: *Computers in Biology and Medicine* 148 (2022), p. 105928.

[101] Yinglin Peng et al. “Magnetic resonance-based synthetic computed tomography images generated using generative adversarial networks for nasopharyngeal carcinoma radiotherapy treatment planning”. In: *Radiotherapy and Oncology* 150 (2020), pp. 217–224.

[102] Yonghao Li et al. “Multi-scale Transformer Network with Edge-aware Pre-training for Cross-Modality MR Image Synthesis”. In: *IEEE Trans-*

*actions on Medical Imaging* (2023), pp. 1–1. DOI: 10.1109/TMI.2023.3288001.

[103] Bo Zhan et al. "D2FE-GAN: Decoupled dual feature extraction based GAN for MRI image synthesis". In: *Knowledge-Based Systems* 252 (2022), p. 109362.

[104] Xiaobin Hu et al. "AutoGAN-Synthesizer: Neural Architecture Search for Cross-Modality MRI Synthesis". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 397–409.

[105] Junzhou Huang, Chen Chen, and Leon Axel. "Fast multi-contrast MRI reconstruction". In: *Magnetic Resonance Imaging* 32.10 (2014), pp. 1344–1352. ISSN: 0730-725X. DOI: https://doi.org/10.1016/j.mri.2014.08.025. URL: https://www.sciencedirect.com/science/article/pii/S0730725X14002562.

[106] Hong Zheng et al. "Multi-contrast brain magnetic resonance image super-resolution using the local weight similarity". In: *BMC medical imaging* 17 (2017), pp. 1–13.

[107] François Rousseau, Alzheimer's Disease Neuroimaging Initiative, et al. "A non-local approach for image super-resolution using intermodality priors". In: *Medical image analysis* 14.4 (2010), pp. 594–605.

[108] José V Manjón et al. "MRI superresolution using self-similarity and image priors". In: *Journal of Biomedical Imaging* 2010 (2010), pp. 1–11.

[109] Qing Lyu et al. "Multi-contrast super-resolution MRI through a progressive network". In: *IEEE transactions on medical imaging* 39.9 (2020), pp. 2738–2749.

[110] Lei Xiang et al. “Deep-learning-based multi-modal fusion for fast MR reconstruction”. In: *IEEE Transactions on Biomedical Engineering* 66.7 (2018), pp. 2105–2114.

[111] Chun-Mei Feng et al. “Exploring Separable Attention for Multi-Contrast MR Image Super-Resolution”. In: *arXiv preprint arXiv:2109.01664* (2021).

[112] Chaowei Fang et al. “Cross-modality high-frequency transformer for MR image super-resolution”. In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 1584–1592.

[113] Chun-Mei Feng et al. “Multi-modal transformer for accelerated mr imaging”. In: *IEEE Transactions on Medical Imaging* (2022).

[114] Runhan Wang et al. “Multi-contrast High Quality MR Image Super-Resolution with Dual Domain Knowledge Fusion”. In: *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2022, pp. 2127–2134.

[115] Guangyuan Li et al. “WavTrans: Synergizing Wavelet and Cross-Attention Transformer for Multi-contrast MRI Super-Resolution”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VI*. Springer. 2022, pp. 463–473.

[116] Liying Lu et al. “Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 6368–6377.

[117] Guangyuan Li et al. “Transformer-empowered Multi-scale Contextual Matching and Aggregation for Multi-contrast MRI Super-resolution”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 20636–20645.

[118] Wenxuan Chen et al. “Compound Attention and Neighbor Matching Network for Multi-contrast MRI Super-resolution”. In: *arXiv preprint arXiv:2307.02148* (2023).

[119] Syed Waqas Zamir et al. “Restormer: Efficient transformer for high-resolution image restoration”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5728–5739.

[120] Beiji Zou et al. “Multi-scale deformable transformer for multi-contrast knee MRI super-resolution”. In: *Biomedical Signal Processing and Control* 79 (2023), p. 104154.

[121] Zhuofan Xia et al. “Vision transformer with deformable attention”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 4794–4803.

[122] Manal AlAmir and Manal AlGhamdi. “The Role of generative adversarial network in medical image analysis: An in-depth survey”. In: *ACM Computing Surveys* 55.5 (2022), pp. 1–36.

[123] Salman UH Dar et al. “Prior-guided image reconstruction for accelerated multi-contrast MRI via generative adversarial networks”. In: *IEEE Journal of Selected Topics in Signal Processing* 14.6 (2020), pp. 1072–1087.

[124] Shan Huang et al. "TransMRSR: Transformer-based Self-Distilled Generative Prior for Brain MRI Super-Resolution". In: *arXiv preprint arXiv:2306.06669* (2023).

[125] Ye Mao et al. "DisC-Diff: Disentangled Conditional Diffusion Model for Multi-Contrast MRI Super-Resolution". In: *arXiv preprint arXiv:2303.13933* (2023).

[126] Manhua Liu et al. "Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis". In: *Neuroinformatics* 16 (2018), pp. 295–308.

[127] Yechong Huang et al. "Diagnosis of Alzheimer's disease via multi-modality 3D convolutional neural network". In: *Frontiers in neuroscience* 13 (2019), p. 509.

[128] Yinghuan Shi et al. "Leveraging coupled interaction for multimodal Alzheimer's disease diagnosis". In: *IEEE transactions on neural networks and learning systems* 31.1 (2019), pp. 186–200.

[129] Yuang Shi et al. "ASMFS: Adaptive-similarity-based multi-modality feature selection for classification of Alzheimer's disease". In: *Pattern Recognition* 126 (2022), p. 108566.

[130] Liangliang Liu et al. "An enhanced multi-modal brain graph network for classifying neuropsychiatric disorders". In: *Medical image analysis* 81 (2022), p. 102550.

[131] Zhi Chen et al. "Orthogonal latent space learning with feature weighting and graph learning for multimodal Alzheimer's disease diagnosis". In: *Medical Image Analysis* 84 (2023), p. 102698.

[132] Heung-Il Suk et al. “Deep sparse multi-task learning for feature selection in Alzheimer’s disease diagnosis”. In: *Brain Structure and Function* 221.5 (2016), pp. 2569–2587.

[133] Tong Tong et al. “Multi-modal classification of Alzheimer’s disease using nonlinear graph fusion”. In: *Pattern recognition* 63 (2017), pp. 171–181.

[134] Shaker El-Sappagh et al. “Multimodal multitask deep learning model for Alzheimer’s disease progression detection based on time series data”. In: *Neurocomputing* 412 (2020), pp. 197–215.

[135] Zongbo Han et al. “Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 20707–20717.

[136] Jin Zhang et al. “Multi-modal cross-attention network for Alzheimer’s disease diagnosis with multi-modality data”. In: *Computers in Biology and Medicine* 162 (2023), p. 107050.

[137] Luoyao Kang et al. “Visual-attribute prompt learning for progressive mild cognitive impairment prediction”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 547–557.

[138] Zifeng Qiu et al. “3D multimodal fusion network with disease-induced joint learning for early Alzheimer’s disease diagnosis”. In: *IEEE Transactions on Medical Imaging* (2024).

[139] Yongsheng Pan et al. “Synthesizing missing PET from MRI with cycle-consistent generative adversarial networks for Alzheimer’s disease diag-

nosis". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III 11*. Springer. 2018, pp. 455–463.

[140] Leiming Jin et al. "A hybrid deep learning method for early and late mild cognitive impairment diagnosis with incomplete multimodal data". In: *Frontiers in Neuroinformatics* 16 (2022), p. 843566.

[141] Yanbei Liu et al. "Incomplete multi-modal representation learning for Alzheimer's disease diagnosis". In: *Medical Image Analysis* 69 (2021), p. 101953.

[142] Yunbi Liu et al. "Assessing clinical progression from subjective cognitive decline to mild cognitive impairment with incomplete multi-modal neuroimages". In: *Medical image analysis* 75 (2022), p. 102266.

[143] Mohammed Abdelaziz, Tianfu Wang, and Ahmed Elazab. "Alzheimer's disease diagnosis framework from incomplete multimodal data using convolutional neural networks". In: *Journal of biomedical informatics* 121 (2021), p. 103863.

[144] Vitaly Herasevich et al. "Validation of an electronic surveillance system for acute lung injury". In: *Intensive care medicine* 35 (2009), pp. 1018–1023.

[145] Helen C Koenig et al. "Performance of an automated electronic acute lung injury screening system in intensive care unit patients". In: *Critical care medicine* 39.1 (2011), pp. 98–104.

[146] Ognjen Gajic et al. "Early identification of patients at risk of acute lung injury: evaluation of lung injury prediction score in a multicenter cohort

study". In: *American journal of respiratory and critical care medicine* 183.4 (2011), pp. 462–470.

[147] Narathip Reamaroon et al. "Accounting for label uncertainty in machine learning for detection of acute respiratory distress syndrome". In: *IEEE journal of biomedical and health informatics* 23.1 (2018), pp. 407–415.

[148] Daniel Zeiberg et al. "Machine learning for patient risk stratification for acute respiratory distress syndrome". In: *PloS one* 14.3 (2019), e0214465.

[149] Narathip Reamaroon et al. "Automated detection of acute respiratory distress syndrome from chest X-Rays using Directionality Measure and deep learning features". In: *Computers in biology and medicine* 134 (2021), p. 104463.

[150] Michael W Sjoding et al. "Deep learning to detect acute respiratory distress syndrome on chest radiographs: a retrospective study with external validation". In: *The Lancet Digital Health* 3.6 (2021), e340–e348.

[151] Kai-Chih Pai et al. "Artificial intelligence–aided diagnosis model for acute respiratory distress syndrome combining clinical data and chest radiographs". In: *Digital Health* 8 (2022), p. 20552076221120317.

[152] Chi Xu et al. "A prediction model for predicting the risk of acute respiratory distress syndrome in sepsis patients: a retrospective cohort study". In: *BMC Pulmonary Medicine* 23.1 (2023), p. 78.

[153] Tu K Tran et al. "A systematic review of machine learning models for management, prediction and classification of ARDS". In: *Respiratory Research* 25.1 (2024), p. 232.

[154] Yang Zhou et al. “Development and validation of a deep learning-based framework for automated lung CT segmentation and acute respiratory distress syndrome prediction: a multicenter cohort study”. In: *Eclinicalmedicine* 75 (2024).

[155] Marco Ganzetti, Nicole Wenderoth, and Dante Mantini. “Mapping pathological changes in brain structure by combining T1-and T2-weighted MR imaging data”. In: *Neuroradiology* 57 (2015), pp. 917–928.

[156] M Kitajima et al. “Comparison of 3D FLAIR, 2D FLAIR, and 2D T2-weighted MR imaging of brain stem anatomy”. In: *American journal of neuroradiology* 33.5 (2012), pp. 922–927.

[157] Chang-Woo Ryu et al. “High-resolution MRI of intracranial atherosclerotic disease”. In: *Neurointervention* 9.1 (2014), p. 9.

[158] Jifeng Dai et al. “Deformable convolutional networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 764–773.

[159] Takeru Miyato and Masanori Koyama. “cGANs with Projection Discriminator”. In: *International Conference on Learning Representations*. 2018. URL: `https://openreview.net/forum?id=ByS1VpgRZ`.

[160] Yang Liu, Lu Meng, and Jianping Zhong. “MAGAN: mask attention generative adversarial network for liver tumor CT image synthesis”. In: *Journal of Healthcare Engineering* 2021 (2021), pp. 1–11.

[161] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. “A u-net based discriminator for generative adversarial networks”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 8207–8216.

[162] Ligong Han et al. “Dual projection generative adversarial networks for conditional image generation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 14438–14447.

[163] Alexia Jolicoeur-Martineau. “The relativistic discriminator: a key element missing from standard GAN”. In: *arXiv preprint arXiv:1807.00734* (2018).

[164] Tero Karras, Samuli Laine, and Timo Aila. “A style-based generator architecture for generative adversarial networks”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 4401–4410.

[165] Xun Huang and Serge Belongie. “Arbitrary style transfer in real-time with adaptive instance normalization”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 1501–1510.

[166] Zhengyao Lv et al. “Semantic-shape adaptive feature modulation for semantic image synthesis”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 11214–11223.

[167] Chong Mou et al. “Metric learning based interactive modulation for real-world super-resolution”. In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*. Springer. 2022, pp. 723–740.

[168] Bhakti Baheti et al. “The brain tumor sequence registration challenge: establishing correspondence between pre-operative and follow-up MRI scans of diffuse glioma patients”. In: *arXiv preprint arXiv:2112.06979* (2021).

[169] Florian Knoll et al. “fastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning”. In: *Radiology: Artificial Intelligence* 2.1 (2020), e190007.

[170] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer. 2015, pp. 234–241.

[171] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. “Road extraction by deep residual u-net”. In: *IEEE Geoscience and Remote Sensing Letters* 15.5 (2018), pp. 749–753.

[172] Zhou Wang et al. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[173] Less Wright. *Ranger - a synergistic optimizer*. `https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer`. 2019.

[174] Richard Zhang et al. “The unreasonable effectiveness of deep features as a perceptual metric”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595.

[175] Anna Breger et al. “A study on the adequacy of common IQA measures for medical images”. In: *arXiv preprint arXiv:2405.19224* (2024).

[176] Gang Yang et al. “Model-Guided Multi-Contrast Deep Unfolding Network for MRI Super-resolution Reconstruction”. In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 3974–3982.

[177] Taesung Park et al. “Semantic image synthesis with spatially-adaptive normalization”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 2337–2346.

[178] Can Cui et al. “Deep multi-modal fusion of image and non-image data in disease diagnosis and prognosis: a review”. In: *Progress in Biomedical Engineering* (2023).

[179] Haolin Zuo et al. “Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities”. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[180] Yury Gorishniy et al. “Revisiting deep learning models for tabular data”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 18932–18943.

[181] Andrew Jaegle et al. “Perceiver: General perception with iterative attention”. In: *International conference on machine learning*. PMLR. 2021, pp. 4651–4664.

[182] Kihyuk Sohn. “Improved deep metric learning with multi-class n-pair loss objective”. In: *Advances in neural information processing systems* 29 (2016).

[183] Clifford R Jack Jr et al. “The Alzheimer’s disease neuroimaging initiative (ADNI): MRI methods”. In: *Journal of Magnetic Resonance Imaging: An*

*Official Journal of the International Society for Magnetic Resonance in Medicine* 27.4 (2008), pp. 685–691.

[184] Yunfeng Fan et al. "PMR: Prototypical Modal Rebalance for Multimodal Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 20029–20038.

[185] Aaron C Lim et al. "Quantification of race/ethnicity representation in Alzheimer's disease neuroimaging research in the USA: a systematic review". In: *Communications medicine* 3.1 (2023), p. 101.

[186] Nenad Tomašev et al. "A clinically applicable approach to continuous prediction of future acute kidney injury". In: *Nature* 572.7767 (2019), pp. 116–119.

[187] Davide Placido et al. "A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories". In: *Nature medicine* 29.5 (2023), pp. 1113–1122.

[188] Giacomo Bellani et al. "Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries". In: *Jama* 315.8 (2016), pp. 788–800.

[189] Nuala J Meyer, Luciano Gattinoni, and Carolyn S Calfee. "Acute respiratory distress syndrome". In: *The Lancet* 398.10300 (2021), pp. 622–637.

[190] Brendan J Clark and Marc Moss. "The acute respiratory distress syndrome: Dialing in the evidence?" In: *JAMA* 315.8 (2016), pp. 759–761.

[191] Zhenzhen Jiang et al. “Machine learning for the early prediction of acute respiratory distress syndrome (ARDS) in patients with sepsis in the ICU based on clinical data”. In: *Heliyon* 10.6 (2024).

[192] Mehak Arora et al. “Uncertainty-Aware Convolutional Neural Network for Identifying Bilateral Opacities on Chest X-rays: A Tool to Aid Diagnosis of Acute Respiratory Distress Syndrome”. In: *Bioengineering* 10.8 (2023), p. 946.

[193] Alistair EW Johnson et al. “MIMIC-IV, a freely accessible electronic health record dataset”. In: *Scientific data* 10.1 (2023), p. 1.

[194] Alistair Johnson et al. “Mimic-cxr database”. In: *PhysioNet10* 13026 (2024), C2JT1Q.

[195] Niall D Ferguson et al. “The Berlin definition of ARDS: an expanded rationale, justification, and supplementary material”. In: *Intensive care medicine* 38 (2012), pp. 1573–1582.

[196] Michael A Matthay et al. “A new global definition of acute respiratory distress syndrome”. In: *American journal of respiratory and critical care medicine* 209.1 (2024), pp. 37–47.

[197] Yidan Feng et al. “Unified Multi-modal Learning for Any Modality Combinations in Alzheimer’s Disease Diagnosis”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 487–497.

[198] Jiaxi Lin et al. “Development and validation of multimodal models to predict the 30-day mortality of icu patients based on clinical parameters and

chest x-rays". In: *Journal of Imaging Informatics in Medicine* 37.4 (2024), pp. 1312–1322.