UNDERSTANDING AND MITIGATING MISSING-MODALITY EFFECTS IN MEDICAL MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Missing modalities are common in medical settings and may hinder clinical decision-making. While prior work has explored ways to mitigate this issue, its impact on medical multimodal large language models (MLLMs) remains underexplored. This raises two questions: (1) how existing medical MLLMs perform under missing modalities; (2) how to improve their performance when modalities are missing. To address the first, we introduce MedGap, a MIMIC-IV-based benchmark with CXR, ECG, and laboratory data to systematically evaluate MLLMs under missing modalities. Unlike prior benchmarks that primarily focus on arbitrary modality combinations, MedGap further examines missingness within disease-relevant pairs, avoiding interference from irrelevant modalities. Our evaluation shows that more modalities do not automatically translate into better performance, while the paired diagnosis evaluation further shows that the impact of a missing modality depends on its relevance to the diagnosis target. Motivated by these findings, we address the second question with MIRI, a two-stage framework that combines probabilistic cross-modal recovery in latent space with teacher-guided integration of the recovered information into the LLM. Experiments on MedGap show that MIRI improves average performance across missing-modality settings and multiple model backbones. Further analysis shows that cross-modal recovery captures missing-modality information and downstream integration enables its effective use, with recovery becoming more beneficial as fewer modalities remain.

1 INTRODUCTION

Medical multimodal large language models (Medical MLLMs) have rapidly evolved from medical image–text understanding toward broader clinical capabilities across heterogeneous data sources, including medical imaging, physiological signals, and electronic health records (Dai et al., 2025; Lin et al., 2025; Xu et al., 2025; Jiang et al., 2025). In real clinical settings, however, these modalities are not always jointly available, making incomplete multimodal inputs common in practice (Le et al., 2025; Yao et al., 2024). Although many Medical MLLMs remain capable of inference when some modalities are unavailable, their performance under such incomplete inputs remains underexplored, limiting our understanding of their reliability in practice.

Prior benchmarks cover different aspects of this problem: MDS-ED (Lopez Alcaraz et al., 2025) studies multimodal prediction within an early emergency department (ED) window; CareBench (Yin et al., 2026) and MedFuse (Hayat et al., 2022) evaluate clinical models under missing or partially available modalities; and MedTVT-QA (Zhang et al., 2026) evaluates Medical MLLMs across modality-availability settings, but mainly treats missingness as an ablation without considering how the relevance of a missing modality varies across diagnosis targets. For example, CXR provides direct diagnostic evidence for pneumonia, Lab offers complementary evidence of inflammation or impaired oxygenation, whereas ECG is generally less informative (Patterson & Loebinger, 2012). Accordingly, existing benchmarks leave missing-modality behavior in Medical MLLMs unclear, particularly how performance varies across modality-availability settings and whether missing different relevant modalities has different effects.

To address these questions, we introduce **MedGap**, a benchmark for systematically evaluating Medical MLLMs under missing modalities, with its key differences from existing benchmarks summarized in Table 1. Built from MIMIC-IV (Johnson et al., 2023b), MedGap uses CXR, ECG, and Lab from the early ED encounter, whose short duration keeps modality acquisition and diagnosis temporally close and limits variation from treatment and disease progression. MedGap contains 20,549 ED encounters from 13,629 patients and covers three clinical prediction tasks: multi-label ED discharge

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

diagnosis over 12 disease families, hospital length of stay, and 30-day mortality. MedGap includes two complementary evaluations: all three tasks are assessed across seven modality-availability settings, while diagnosis is further evaluated at the disease level by comparing each clinically relevant modality pair with either modality alone on the same cases and diagnosis targets.

Table 1: Comparison of related medical multimodal benchmarks. Early ED Alignment indicates a short interval between modality acquisition and diagnosis.

Benchmark	MLLM Eval.	Missing-Modality Eval.	Early ED Alignment	Disease-Level Missing Eval.
MDS-ED	✗	✗	✗	✗
CareBench	✗	✓	✗	✓
MedFuse	✗	✓	✗	✗
MedTVT-QA	✓	✓	✗	✗
MedGap	✓	✓	✓	✓

Based on our evaluation with MedGap, we find that missing-modality effects are both model-dependent and target-dependent, where target dependence means that the impact of a missing modality varies with its relevance to the prediction target. The model-dependent variation suggests that modality availability alone does not determine performance and that how models use multimodal information also matters. Meanwhile, target dependence indicates that useful predictive information can be lost when a target-relevant modality is missing.

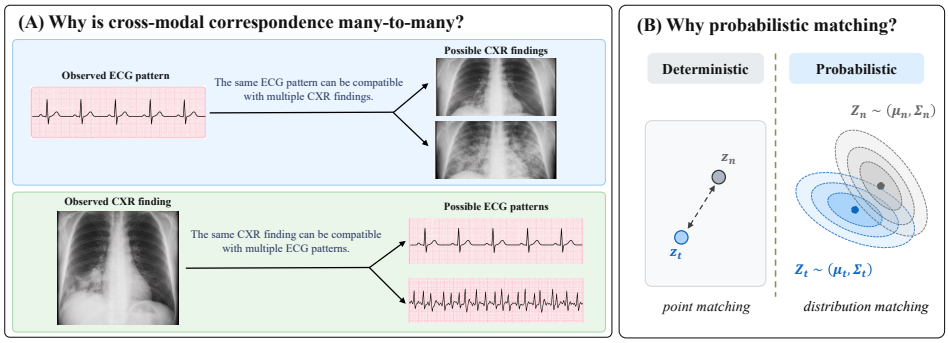

Figure 1: Many-to-many correspondence motivates probabilistic distribution matching.

Building on these findings, we introduce a two-stage **M**issing-Modality **I**nformation **R**ecovery and **I**ntegration (**MIRI**) framework for Medical MLLMs. As illustrated in Figure 1, cross-modal clinical correspondence is often many-to-many: the same pattern in one modality can correspond to multiple manifestations in another, and vice versa (Gao et al., 2025; Javed et al., 2024). This uncertainty makes deterministic point prediction restrictive, motivating MIRI to formulate missing-modality recovery as probabilistic distribution matching. Inspired by joint-embedding predictive learning (Assran et al., 2023; Lei et al., 2025), the recovery stage first learns modality-specific probabilistic latent targets through within-modality self-supervision and then predicts them from observed modalities through probabilistic cross-modal predictive learning. To make the recovered information usable for downstream prediction, MIRI further employs teacher-guided integration, using supervision from a full-modality teacher to adapt the Medical MLLM to combine recovered representations with observed inputs. Evaluations on MedGap show that MIRI improves average performance across missing-modality settings and multiple Medical MLLM backbones.

Our main contributions are summarized as follows:

- We introduce **MedGap**, a benchmark for missing-modality evaluation of Medical MLLMs using early-ED CXR, ECG, and Lab with short acquisition-to-diagnosis intervals, combining seven modality-availability settings with disease-level diagnosis evaluation.
- We propose **MIRI**, a two-stage framework for Medical MLLMs that performs probabilistic cross-modal recovery in a learned latent space and teacher-guided integration of the recovered representations into the LLM.
- Using MedGap, we systematically analyze Medical MLLM performance across missing-modality conditions and clinical targets, and show that MIRI achieves the best average performance across incomplete-modality settings.

2 Related Work

Early Medical MLLMs primarily extended general-purpose vision–language models to biomedical image–text understanding, enabling open-ended visual question answering and multimodal interaction in the medical domain (Li et al., 2023; Moor et al., 2023). More recent models have further strengthened medical visual understanding and reasoning (Sellergren et al., 2025), while Medical MLLMs have also expanded toward heterogeneous clinical inputs beyond images, including physiological signals and electronic health records (Dai et al., 2025; Lin et al., 2025; Xu et al., 2025; Jiang et al., 2025; Zhang et al., 2026). Although some of these models can operate when one or more modalities are unavailable, missing-modality robustness is not a primary focus of their design or evaluation, leaving their behavior under incomplete multimodal inputs insufficiently understood.

Existing work on missing modalities in medical multimodal learning spans both evaluation and learning (Le et al., 2025; Wu et al., 2026). On the evaluation side, MDS-ED (Lopez Alcaraz et al., 2025) benchmarks multimodal clinical prediction within an early emergency department (ED) observation window, but does not explicitly study modality missingness. CareBench (Yin et al., 2026) and MedFuse (Hayat et al., 2022) evaluate clinical prediction under missing or partially available modalities. MedTVT-R1 (Zhang et al., 2026) further evaluates a Medical MLLM across different combinations of CXR, ECG, and Lab data. However, existing evaluations do not directly examine how the impact of missing modalities varies across disease-relevant targets.

On the learning side, existing medical approaches address missing modalities through disentangled or robust fusion (Yao et al., 2024; Wang et al., 2025b), modality synthesis and imputation (Zhang et al., 2025; Qi et al., 2025), and context-driven recovery for medical prediction (Liu et al., 2026). Most of these methods are developed for task-specific discriminative models rather than Medical MLLMs. A recent exception is MissRAG (Pipoli et al., 2025), which addresses missing modalities in multimodal LLMs through retrieval-based augmentation. In contrast, we recover missing-modality information from available clinical inputs and integrate it into the LLM for downstream prediction.

3 Benchmarking Under Missing Modalities

MedGap is constructed around emergency department (ED) encounters, with CXR, ECG, and Lab restricted to the first 12 hours after ED arrival. We consider three prediction tasks: ED discharge diagnosis (DX), hospital length of stay (LOS), and 30-day mortality (M30). MedGap includes two complementary evaluations. First, the modality-availability evaluation assesses all three tasks across seven modality-availability settings while keeping the prediction target fixed, measuring how performance changes under different missing-modality patterns. Second, for diagnosis, where relevant modalities can be identified by disease, the modality-paired evaluation selects the two most clinically relevant modalities for each disease and compares their combination with each modality alone on the same cases, testing whether missing different disease-relevant modalities has different effects on the same diagnosis. Figure 2 summarizes MedGap construction and dataset distribution.

3.1 MedGap: Benchmark Construction

The ED encounter provides a common temporal anchor, keeping multimodal acquisition and diagnosis temporally close within the same acute presentation, before variation from subsequent inpatient treatment and disease progression. MIMIC-IV (Johnson et al., 2023b) and its linked datasets (Johnson et al., 2019b; Gow et al., 2023; Johnson et al., 2023a) are well suited to this design because they provide diverse clinical modalities that can be linked across datasets by subject identity and timestamp. We construct MedGap by linking each ED encounter across the available CXR, ECG, and laboratory records. Using ED arrival as the time anchor, we restrict all clinical inputs to the first 12 hours, ensuring that CXR, ECG, and laboratory data fall within the same early acute-care window. CXR inputs consist of eligible chest radiographs, ECG inputs of rendered 12-lead recordings, and Lab inputs of chronologically ordered measurements from a predefined test set. We additionally include age, sex, and arrival transport as limited encounter context.

MedGap includes three prediction tasks. **Diagnosis** uses ED discharge diagnoses rather than subsequent hospital discharge diagnoses, which may incorporate later tests, treatments, and disease progression beyond the 12-hour input window. This keeps the diagnosis target aligned with the same ED encounter as the multimodal inputs. Diagnosis is formulated as multi-label prediction over 12 diagnosis families: pneumonia, dyspnea, heart failure, arrhythmia, acute kidney injury, electrolyte

disorders, anemia, sepsis, acute myocardial infarction, pleural disease, acute respiratory failure, and diabetic ketoacidosis. **Length of stay** uses four categories: $[0, 1)$, $[1, 3)$, $[3, 7)$, and $[7, \infty)$ days. **30-day mortality** predicts whether the patient dies within 30 days after the end of the 12-hour window. We use patient-disjoint train/test splits to prevent leakage. The resulting benchmark contains 20,549 ED encounters from 13,629 patients, with 15,439 for training and 5,110 for testing.

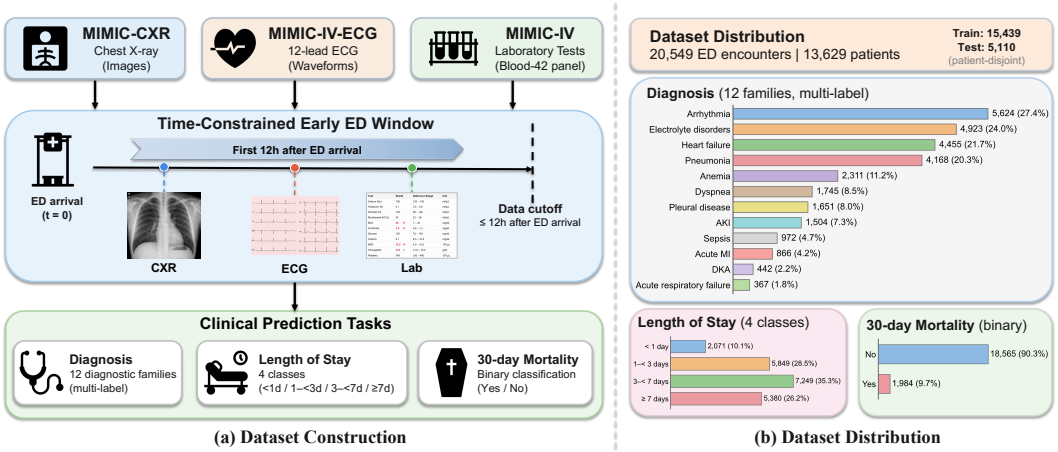

Figure 2: Overview of MedGap construction and dataset distribution. (a) ED-centered cohort construction within the first 12 hours after arrival and the three prediction tasks. (b) Distributions of diagnosis, length of stay, and 30-day mortality.

3.2 Missing-Modality Evaluation

MedGap includes two complementary missing-modality evaluations. First, we evaluate all three prediction tasks across seven modality-availability settings: CXR+ECG+Lab, CXR+ECG, CXR+Lab, ECG+Lab, CXR only, ECG only, and Lab only. For each test encounter, all settings use the same prediction targets, so only the available input modalities change. This within-encounter design measures how overall performance changes across missing-modality patterns. However, for diagnosis, aggregate performance mixes diseases with different modality relevance: a small performance change may reflect either model robustness or limited relevance of the missing modality, while a large drop may reflect either model difficulty under missingness or strong dependence of the diagnosis on that modality. Thus, this evaluation alone cannot determine whether the effect of a missing modality depends on its relevance to the disease being diagnosed.

Second, to examine modality relevance at the disease level, we perform a paired evaluation that fixes the diagnosis targets and compares the effects of missing different disease-relevant modalities. We focus on diagnosis because relevant modalities can be identified for individual diseases, whereas LOS and M30 are broader prognostic outcomes. For each modality pair, we define a fixed set of diagnosis targets for which both modalities are clinically relevant and evaluate three settings on the same encounters and targets: both modalities available, the first modality alone, and the second modality alone. The resulting disease–modality assignments are summarized in Table 2.

Table 2: Diagnosis target scopes for modality-paired evaluation.

Modality Pair	**Diagnosis Targets**
CXR–ECG	Pneumonia, heart failure, arrhythmia, acute myocardial infarction, pleural disease
CXR–Lab	Pneumonia, heart failure, acute respiratory failure, pleural disease, electrolyte disorders
ECG–Lab	Heart failure, arrhythmia, acute myocardial infarction, electrolyte disorders

4 Missing-Modality Recovery and Integration

We propose MIRI, a two-stage framework motivated by the many-to-many nature of cross-modal clinical correspondence, where the same observed inputs can be compatible with multiple manifestations in a missing modality. MIRI therefore formulates missing-modality recovery as probabilistic cross-modal predictive learning rather than deterministic point prediction, predicting latent distributions for missing modalities in modality-specific target spaces learned through within-modality

self-supervision. To make the recovered information useful for clinical prediction, MIRI further introduces teacher-guided integration, mapping the recovered tokens into the LLM's input embedding space and adapting the student under full-modality teacher supervision. This preserves the original input pathways for observed modalities while supplementing only the missing ones.

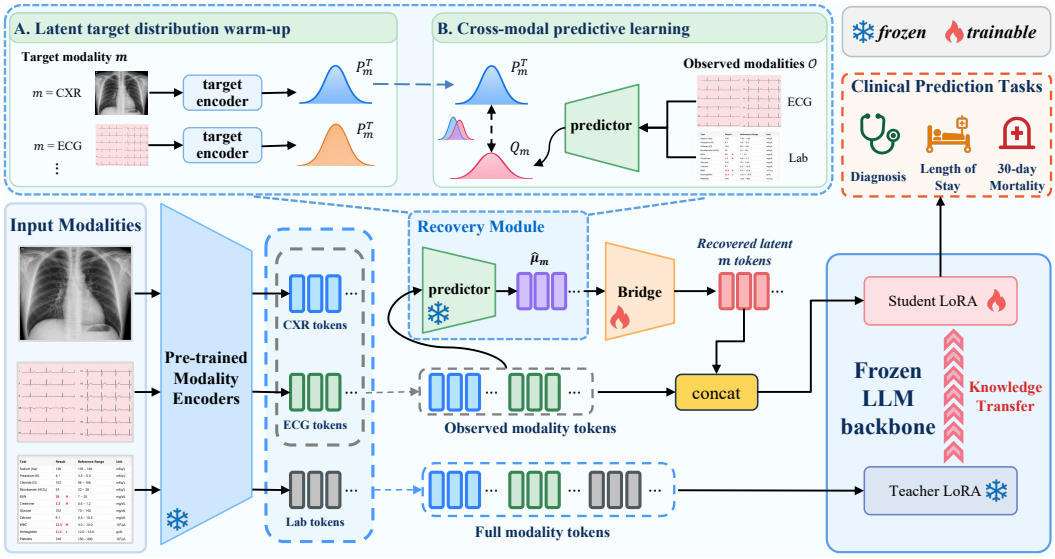

Figure 3: Overview of MIRI, which learns probabilistic latent targets via within-modality self-supervision (A) and recovers missing-modality tokens via cross-modal prediction (B). Recovered tokens are integrated with observed inputs under full-modality teacher supervision.

4.1 Probabilistic Cross-Modal Recovery

Let $\mathcal{M}$ denote the set of clinical modalities and $\mathbf{H}_m \in \mathbb{R}^{N_m \times d}$ the frozen pre-LLM token embeddings of modality $m \in \mathcal{M}$, where N_m is the number of tokens. Rather than directly predicting the full modality-token sequence, MIRI compresses $\mathbf{H}_m$ into a compact set of K_m latent tokens (Jaegle et al., 2021). These compact latent tokens define the target space for cross-modal recovery. Since clinical modalities often exhibit many-to-many correspondence, the same ECG pattern may be associated with different CXR findings, while the same CXR finding may coexist with different ECG patterns (Gao et al., 2025; Javed et al., 2024). We therefore model each target latent token as a Gaussian distribution rather than a deterministic point (Chun et al., 2021; Gao et al., 2025). Specifically, letting $\mathbf{Z}_m = \{\mathbf{z}_{m,k}\}_{k=1}^{K_m}$ denote the latent tokens of modality m, we define the probabilistic target representation as

$$P_m^T(\mathbf{Z}_m \mid \mathbf{H}_m) = \left\{\mathcal{N}\left(\mathbf{z}_{m,k}; \boldsymbol{\mu}_{m,k}, \mathrm{diag}(\boldsymbol{\sigma}_{m,k}^2)\right)\right\}_{k=1}^{K_m}. \tag{1}$$

Latent target distribution warm-up. We first warm up the probabilistic latent target distributions for each modality through within-modality self-supervision. For modality m, the complete token sequence $\mathbf{H}_m$ is encoded by an EMA target network to produce P_m^T, while an online network predicts the same latent distribution P_m^S from a masked view $\widetilde{\mathbf{H}}_m$. We align the two distributions using

$$\mathcal{L}_{\text{within}} = \mathbb{E}_{m \in \mathcal{M}} D_{\mathrm{KL}}\left(P_m^T(\mathbf{Z}_m \mid \mathbf{H}_m) \,\|\, P_m^S(\mathbf{Z}_m \mid \widetilde{\mathbf{H}}_m)\right). \tag{2}$$

The target network is updated by exponential moving average during warm-up and then frozen to define the latent target distributions for subsequent cross-modal recovery.

Probabilistic Cross-Modal Predictive Learning. Inspired by joint-embedding predictive learning (Assran et al., 2023; Lei et al., 2025), MIRI treats observed modalities as context and a missing modality as an unobserved prediction target. It exploits shared clinical information across modalities to predict the target's probabilistic latent representation, while accounting for uncertainty in cross-modal correspondence.

With the target space fixed, let $\mathcal{O} \subseteq \mathcal{M} \setminus \{m\}$ denote the observed modalities. From their token embeddings $\mathbf{H}_{\mathcal{O}}$, the cross-modal predictor predicts the probabilistic latent representation of the

missing modality m as

$$Q_m(\mathbf{Z}_m \mid \mathbf{H}_{\mathcal{O}}) = \left\{\mathcal{N}\big(\mathbf{z}_{m,k}; \widehat{\boldsymbol{\mu}}_{m,k}, \mathrm{diag}(\widehat{\boldsymbol{\sigma}}^2_{m,k})\big)\right\}_{k=1}^{K_m}. \tag{3}$$

We train the cross-modal predictor to match the fixed target representation using

$$\mathcal{L}_{\mathrm{cross}}(\mathcal{C}) = \mathbb{E}_{(\mathcal{O},m)\sim\mathcal{C}} D_{\mathrm{KL}}\big(P_m^T(\mathbf{Z}_m \mid \mathbf{H}_m) \,\|\, Q_m(\mathbf{Z}_m \mid \mathbf{H}_{\mathcal{O}})\big). \tag{4}$$

We first train all single-source directions with $|\mathcal{O}| = 1$, and then extend to multi-source prediction with $\mathcal{O} = \mathcal{M} \setminus \{m\}$. During training, the target modality is used only to construct P_m^T and is never provided to the prediction pathway. At inference time, prediction uses only the observed modalities, and the predicted mean $\widehat{\boldsymbol{\mu}}_m$ is passed as the recovered latent tokens to the integration stage.

4.2 Integrating Recovered Information

Recovering missing-modality representations is only the first step; these representations must also be made usable for downstream clinical prediction. MIRI addresses this through teacher-guided integration. For each missing modality $m \notin \mathcal{O}$, a lightweight modality-specific bridge B_m maps the predicted latent means $\widehat{\boldsymbol{\mu}}_m$ into the LLM's input embedding space. Full-modality teacher supervision then guides the student to integrate the bridged tokens $B_m(\widehat{\boldsymbol{\mu}}_m)$ with the observed inputs for downstream prediction. This design preserves the original input pathways for observed modalities and selectively supplements only the missing ones.

We then adapt the Medical MLLM to integrate the recovered information with the observed inputs for downstream prediction. Following prior work that transfers knowledge from complete- to incomplete-modality models (Wang et al., 2025a), we instantiate a frozen full-modality teacher with parameters $\bar{\phi}$ and a trainable student with parameters ϕ, both initialized from the same model fine-tuned on complete multimodal inputs. Let $\mathbf{c}^T$ denote the teacher context containing the question and all modality inputs. The student context combines the same question, the original tokens of the observed modalities, and the bridged tokens, i.e., $\mathbf{c}^S_{\mathcal{O}} = [\mathbf{q}, \{\mathbf{H}_o\}_{o\in\mathcal{O}}, \{B_m(\widehat{\boldsymbol{\mu}}_m)\}_{m\notin\mathcal{O}}]$, where $\mathbf{q}$ denotes the question tokens. We further define $\mathbf{c}^S_{\mathcal{O},t} = (\mathbf{c}^S_{\mathcal{O}}, \mathbf{y}_{<t})$ and $\mathbf{c}^T_t = (\mathbf{c}^T, \mathbf{y}_{<t})$. Let $\mathcal{A}_\tau$ denote the answer positions associated with task $\tau \in \mathcal{T}$, where $\mathcal{T} = \{\mathrm{DX}, \mathrm{LOS}, \mathrm{M30}\}$. Let $x \sim \mathcal{D}$ denote a training example and $\mathcal{O} \sim \mathcal{P}$ an observed-modality configuration. We optimize

$$\mathcal{L}_{\mathrm{int}} = \mathbb{E}_{\substack{x\sim\mathcal{D}\\ \mathcal{O}\sim\mathcal{P}}}\left[\frac{1}{|\mathcal{T}|}\sum_{\tau\in\mathcal{T}}\frac{1}{|\mathcal{A}_\tau|}\sum_{t\in\mathcal{A}_\tau}\left[-\log p_\phi(y_t \mid \mathbf{c}^S_{\mathcal{O},t}) + D_{\mathrm{KL}}\big(p_{\bar{\phi}}(\cdot \mid \mathbf{c}^T_t) \,\|\, p_\phi(\cdot \mid \mathbf{c}^S_{\mathcal{O},t})\big)\right]\right]. \tag{5}$$

The first term provides supervision from the ground-truth answers, while the second distills the full-modality teacher into the student at the answer positions. During this stage, the recovery module and teacher remain frozen; only the bridges and student LoRA parameters are optimized.

5 Evaluation and Analysis

5.1 Experimental Setup

We evaluate seven MLLMs in the zero-shot setting: GPT-5.2 (OpenAI, 2025), InternVL3.5-8B (Wang et al., 2025c), Qwen3-VL-8B (Bai et al., 2025), HealthGPT-Pro-8B (Lin et al., 2025), Hulu-Med-7B (Jiang et al., 2025), Lingshu-7B (Xu et al., 2025), and QoQ-Med-Omni-7B (Dai et al., 2025). All models are evaluated across the seven modality-availability settings and the modality-paired diagnosis evaluation, using shortened names in the result tables. To evaluate our method, we apply MIRI to Hulu-Med-4B, Hulu-Med-7B, and Qwen3-VL-8B and compare it with SFT and MissRAG (Pipoli et al., 2025), using the same SFT backbone for controlled comparison. We report micro F1 (MiF1) and macro F1 (MaF1) for diagnosis, accuracy and MaF1 for length of stay, and F1 and balanced accuracy (BAcc) for 30-day mortality; paired diagnosis is evaluated with MiF1 and MaF1 over the diseases assigned to each modality pair.

5.2 Performance across Modality Availability

Table 3 shows that ***more available modalities do not automatically translate into better performance.*** Indeed, in some settings, removing a modality improves performance over complete multimodal input. For example, Hulu-Med improves from 26.55 to 34.76 MiF1 on diagnosis after removing Lab; GPT-5.2 attains higher LOS Acc with CL than with all modalities; and HealthGPT

Table 3: Zero-shot performance across different modality settings (C: CXR, E: ECG, L: Lab).

Diagnosis	CEL		CE		CL		EL		C		E		L	
Model	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1
GPT-5.2	**45.65**	**43.25**	36.39	17.16	**42.27**	**39.77**	**42.24**	**34.74**	27.32	12.88	22.81	3.83	**36.69**	**29.64**
InternVL3.5	29.66	24.07	32.03	13.91	27.19	23.70	25.55	17.85	28.08	10.94	**26.46**	**7.34**	21.81	19.72
Qwen3-VL	35.87	26.00	30.29	9.75	33.44	23.47	33.99	22.84	24.39	8.34	22.67	3.80	29.72	20.91
HealthGPT	30.78	25.86	**36.46**	15.20	29.27	24.39	27.09	20.09	**30.40**	12.20	22.38	4.24	25.93	20.81
Hulu-Med	26.55	24.26	34.76	**18.40**	25.47	23.42	29.06	22.99	27.94	**15.17**	19.39	4.70	24.41	22.34
Lingshu	33.24	21.74	27.73	12.41	33.27	22.41	34.95	22.77	27.90	11.57	21.74	4.65	30.99	21.88
QoQ-Med	31.40	21.91	29.10	9.32	29.58	20.68	29.53	16.41	19.94	7.29	22.77	4.37	25.36	14.84
Length of Stay	CEL		CE		CL		EL		C		E		L	
Model	Acc	MaF1	Acc	MaF1	Acc	MaF1	Acc	MaF1	Acc	MaF1	Acc	MaF1	Acc	MaF1
GPT-5.2	**41.21**	**37.69**	33.51	**26.91**	**41.77**	**38.01**	**39.77**	**37.35**	34.70	**27.25**	22.58	15.11	**39.12**	**35.67**
InternVL3.5	37.67	22.97	**36.40**	22.03	38.47	22.79	37.65	22.37	35.38	21.22	29.92	14.42	33.84	21.29
Qwen3-VL	36.28	28.17	30.25	22.51	36.30	28.29	36.83	28.86	34.21	25.35	28.81	11.69	37.20	29.36
HealthGPT	24.68	14.19	25.54	14.79	27.46	14.57	29.88	14.51	25.50	14.14	28.79	11.38	33.74	21.60
Hulu-Med	29.98	14.62	29.10	11.73	31.49	16.97	29.04	11.94	28.88	11.39	28.81	11.18	31.33	16.39
Lingshu	36.28	22.09	35.89	24.40	36.34	20.40	36.16	21.59	**35.85**	22.44	**31.55**	**21.77**	32.07	22.51
QoQ-Med	36.40	23.38	32.07	19.53	36.93	22.98	32.74	23.31	32.49	19.70	28.34	12.10	32.45	18.49
30-Day Mortality	CEL		CE		CL		EL		C		E		L	
Model	F1	BAcc	F1	BAcc	F1	BAcc	F1	BAcc	F1	BAcc	F1	BAcc	F1	BAcc
GPT-5.2	**31.01**	**66.04**	13.49	53.18	**31.56**	**65.82**	**29.59**	**64.80**	14.43	53.09	2.21	50.23	**27.26**	**63.44**
InternVL3.5	22.17	60.85	23.24	61.24	22.08	60.78	22.00	60.40	**22.72**	**60.70**	20.44	56.36	17.53	50.13
Qwen3-VL	20.37	57.18	20.14	56.53	20.81	58.15	20.67	57.77	21.68	59.43	20.51	56.90	21.46	59.37
HealthGPT	19.14	54.20	**24.57**	**62.50**	20.65	57.91	20.19	56.72	22.57	59.93	**20.94**	**57.68**	20.35	57.19
Hulu-Med	19.75	55.72	19.69	55.56	21.28	59.35	18.54	52.33	21.28	59.42	18.10	50.78	18.01	50.41
Lingshu	18.11	50.75	18.58	52.29	18.12	50.82	17.64	49.33	19.24	54.29	18.32	51.50	15.57	43.24
QoQ-Med	20.74	57.09	19.88	55.82	20.97	57.31	17.01	51.15	20.32	56.33	14.63	52.68	23.28	58.53

improves its M30 BAcc from 54.20 to 62.50 after removing Lab. These results show that the number of available modalities alone is insufficient to explain performance under missing modalities.

The effect of modality availability also varies across models: ***the same change in available modalities can affect different models differently.*** For example, removing Lab from CEL in diagnosis decreases GPT-5.2 MiF1 from 45.65 to 36.39, but increases Hulu-Med from 26.55 to 34.76. Such differences may arise from variations in model architecture, multimodal training data, and training objectives, which can lead to different patterns of modality utilization. This provides a model-side explanation for part of the variation observed above, but leaves a target-side question unresolved: whether the effect of a missing modality also depends on its relevance to the prediction target.

5.3 Modality-Paired Diagnosis Evaluation

To further exclude the influence of modalities irrelevant to the target diagnosis, the modality-paired diagnosis evaluation fixes the diagnosis targets and compares each relevant modality pair with its two single-modality settings. Table 4 shows that ***the impact of missing a modality depends on its relevance to the diagnosis target.*** Notably, paired input achieves higher MaF1 than either single modality in all 21 model–pair comparisons, indicating a consistent benefit from using both relevant modalities together. Across all seven models, C consistently outperforms E within CE, while L consistently outperforms C within CL and E within EL in MaF1. This consistent asymmetry shows that equal modality counts can still yield different performance depending on which modality remains.

Table 4: Modality-paired diagnosis evaluation. Each entry reports MiF1/MaF1.

	CXR–ECG			CXR–Lab			ECG–Lab		
Model	CE	C	E	CL	C	L	EL	E	L
GPT-5.2	**47.49**/36.56	37.20/26.88	30.27/9.00	**58.04/49.56**	**37.92**/27.47	**49.84/30.05**	**59.64/57.48**	31.06/11.48	**51.88/44.78**
InternVL3.5	37.70/33.05	34.53/25.45	**32.09/17.61**	38.63/33.20	34.03/21.72	31.30/24.95	36.49/36.79	**32.59/22.02**	33.41/33.82
Qwen3-VL	37.56/23.23	30.88/19.85	30.15/9.11	47.51/32.17	30.89/17.70	48.41/25.48	50.65/49.06	30.89/11.39	48.34/41.08
HealthGPT	44.61/35.93	37.64/27.44	29.45/10.16	40.08/35.23	37.52/25.42	40.39/25.74	39.88/37.37	30.14/12.71	42.99/36.90
Hulu-Med	43.08/**38.71**	**37.71/29.73**	23.72/11.28	32.37/30.74	37.07/**28.31**	33.71/29.20	38.11/37.38	24.10/14.10	37.23/36.36
Lingshu	34.29/26.39	35.72/25.11	28.90/11.15	40.41/31.23	35.93/23.08	39.33/28.63	41.16/38.34	29.60/13.94	38.63/30.65
QoQ-Med	35.53/22.05	27.06/15.39	30.28/10.50	39.55/27.33	27.61/14.70	33.63/15.86	38.67/33.59	31.02/13.12	33.81/26.04

The per-diagnosis results in Appendix C show this relationship more directly. Across the seven models, performance drops most when C is removed for pneumonia, E for arrhythmia, and L for electrolyte disorders. These disease-level patterns further support the role of target-specific modality relevance, providing the target-side explanation left unresolved above. Alongside the model-side variation in Table 3, this shows that performance under missing modalities depends not simply on

modality count, but on both how models utilize the available modalities and how relevant those modalities are to the prediction target.

Table 5: Method comparison on Hulu-Med-7B across missing-modality settings.

Diagnosis	CE		CL		EL		C		E		L		**Avg.**	
Method	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1	MiF1	MaF1
Zero-shot	34.76	18.40	25.47	23.42	29.06	22.99	27.94	15.17	19.39	4.70	24.41	22.34	26.84	17.84
SFT	**40.81**	20.52	53.81	48.96	50.21	40.00	28.92	14.27	22.74	3.60	41.59	34.96	39.68	27.05
MissRAG	30.89	16.90	55.25	50.44	46.34	39.80	26.13	13.79	23.46	7.78	41.81	36.96	37.31	27.61
MIRI	40.27	**28.67**	**55.95**	**50.59**	**54.19**	**45.49**	**33.17**	**22.43**	**34.90**	**17.80**	**49.20**	**41.76**	**44.61**	**34.46**
Length of Stay	CE		CL		EL		C		E		L		**Avg.**	
Method	Acc	MaF1	Acc	MaF1	Acc	MaF1	Acc	MaF1	Acc	MaF1	Acc	MaF1	Acc	MaF1
Zero-shot	29.10	11.73	31.49	16.97	29.04	11.94	28.88	11.39	28.81	11.18	31.33	16.39	29.78	13.27
SFT	29.10	27.00	**41.51**	**39.49**	**39.24**	**33.89**	27.71	25.77	21.08	15.17	**38.67**	33.50	32.88	29.14
MissRAG	37.65	25.12	40.80	38.85	36.73	24.11	37.65	24.46	34.29	15.57	33.17	21.20	36.72	24.89
MIRI	**40.47**	**33.58**	41.06	38.23	38.45	33.63	**40.04**	**30.67**	**34.42**	**19.04**	38.32	**34.14**	**38.79**	**31.55**
30-Day Mortality	CE		CL		EL		C		E		L		**Avg.**	
Method	F1	BAcc	F1	BAcc	F1	BAcc	F1	BAcc	F1	BAcc	F1	BAcc	F1	BAcc
Zero-shot	19.69	55.56	21.28	59.35	18.54	52.33	21.28	59.42	18.10	50.78	18.01	50.41	19.48	54.64
SFT	33.37	**70.21**	32.62	72.14	29.37	66.60	32.78	**69.92**	18.07	54.45	28.88	**66.61**	29.18	66.65
MissRAG	**34.41**	69.05	32.26	72.24	22.25	60.94	**34.19**	68.32	20.05	55.91	19.60	55.33	27.12	63.63
MIRI	30.46	70.11	**34.44**	**72.67**	**31.07**	**67.50**	29.71	69.70	**21.82**	**59.43**	**29.77**	66.34	**29.54**	**67.62**

Table 6: Method comparison on Hulu-Med-7B for modality-paired diagnosis.

	CXR–ECG			CXR–Lab			ECG–Lab		
Method	CE	C	E	CL	C	L	EL	E	L
Zero-shot	43.08/38.71	37.71/29.73	23.72/11.28	32.37/30.74	37.07/28.31	33.71/29.20	38.11/37.38	24.10/14.10	37.23/36.36
SFT	**54.57/44.95**	39.68/30.72	30.24/8.64	65.69/57.76	40.30/29.64	52.46/34.13	62.59/58.67	30.98/10.80	52.52/48.20
MissRAG	42.85/37.36	35.26/29.73	31.81/17.25	65.26/**58.27**	36.19/28.33	43.05/35.17	54.49/56.68	32.34/20.50	49.74/51.73
MIRI	52.54/44.14	**43.22/32.30**	**39.46/29.78**	**66.26**/57.98	**45.22/38.71**	**57.55/41.27**	**66.61/64.23**	**43.11/38.24**	**60.17/56.86**

5.4 Improving Performance under Missing Modalities

Table 5 compares MIRI with zero-shot inference, SFT, and MissRAG on Hulu-Med-7B across six incomplete-modality settings. We focus on these settings because MIRI is applied only when a modality is missing; when all modalities are available, the teacher model is used directly. Although MIRI does not perform best in every individual setting, it achieves the best average performance on all six metrics, improving over SFT by 4.93/7.41 MiF1/MaF1 for diagnosis, 5.91/2.41 Acc/MaF1 for LOS, and 0.36/0.97 F1/BAcc for M30. The gains are most consistent for diagnosis, where MIRI is best in 11 of 12 setting–metric comparisons, while LOS and M30 vary more across settings.

Table 6 shows that this advantage becomes more consistent when a relevant modality is missing. When one modality from a relevant pair is unavailable, MIRI achieves the best result in all 12 singleton setting–metric comparisons. When both relevant modalities are available, it remains competitive, ranking first in three comparisons and second in the other three. The per-diagnosis results in Appendix C show similar improvements across targets: electrolyte-disorder F1 increases from 0.00 to 38.46 when L is missing, while arrhythmia increases from 4.01 to 31.47 when E is missing.

Table 7: Missing-modality recovery evaluation on Hulu-Med-7B.

	AUROC			AUPRC		
Direction	Obs.	Pred.	Δ	Obs.	Pred.	Δ
ECG → CXR	62.81	**69.39**	+6.59	21.63	**24.35**	+2.72
Lab → CXR	67.22	**71.29**	+4.07	22.68	**26.53**	+3.85
ECG+Lab → CXR	66.12	**74.42**	+8.30	24.03	**29.94**	+5.91
CXR → ECG	62.98	**65.51**	+2.53	15.09	**16.17**	+1.07
Lab → ECG	60.05	**60.40**	+0.35	12.00	**12.16**	+0.16
CXR+Lab → ECG	63.29	**68.33**	+5.04	15.65	**17.62**	+1.98
CXR → Lab	63.70	**65.62**	+1.92	60.90	**62.73**	+1.83
ECG → Lab	60.68	**61.25**	+0.57	57.92	**58.20**	+0.27
CXR+ECG → Lab	64.15	**67.01**	+2.85	61.38	**64.32**	+2.94

5.5 Ablation and Analysis

Before examining whether recovery drives downstream gains, we first ask whether the recovered representations capture information associated with the missing modality. For each recovery direction $\mathcal{O} \to m$, with observed set $\mathcal{O}$ and missing target m, we compare two representations: the observed

modality tokens from $\mathcal{O}$ (Obs.) and the predicted mean tokens $\widehat{\boldsymbol{\mu}}_m$ (Pred.). Both are mean-pooled before probing, with pooled Obs. features concatenated when two modalities are observed. We train linear probes on the same modality-specific clinical labels for m: CXR findings, ECG diagnostic abnormalities, or Lab abnormality groups. As shown in Table 7, Pred. outperforms Obs. across all nine directions on both AUROC and AUPRC, with AUROC gains ranging from 0.35 to 8.30 points. This provides representation-level evidence that MIRI makes information associated with the missing modality more accessible. A complementary UMAP analysis in Appendix D further illustrates cross-modal recovery, with predicted representations aligned with their modality-specific target representations. Label construction and probe details are provided in Appendix B.1.

Table 8: Component ablation on Hulu-Med-7B, averaged over six incomplete-modality settings.

	Components		Diagnosis		Length of Stay		30-Day Mortality	
Variant	**Recovery**	**KD**	MiF1	MaF1	Acc	MaF1	F1	BAcc
w/o Recovery	×	✓	40.20	27.78	33.62	29.40	**29.98**	66.79
w/o KD	✓	×	<u>42.61</u>	<u>32.79</u>	<u>38.78</u>	<u>29.59</u>	29.48	<u>67.44</u>
MIRI	✓	✓	**44.61**	**34.46**	**38.79**	**31.55**	<u>29.54</u>	**67.62**

Table 8 examines whether the recovered information contributes to downstream performance. Removing recovery while retaining KD causes substantial drops in diagnosis and LOS, reducing diagnosis from 44.61 to 40.20 MiF1 and from 34.46 to 27.78 MaF1, and LOS Acc from 38.79 to 33.62. Removing KD while retaining recovery leads to smaller declines, indicating that recovery is the primary contributor to MIRI's downstream gains, while KD provides a complementary benefit. We further examine how this contribution changes with missingness severity. Across the three evaluated backbones, adding recovery improves diagnosis by 1.49 MiF1 and 3.01 MaF1 on average when one modality is missing, compared with 6.33 MiF1 and 6.05 MaF1 when two modalities are missing. Thus, the benefit of recovery increases as less observed modality information remains.

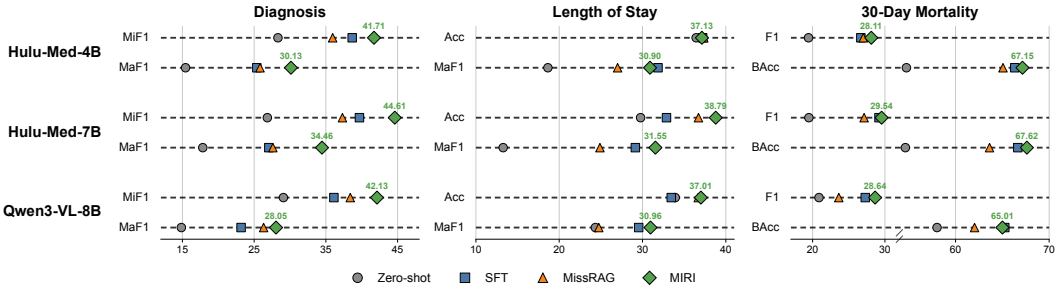

Figure 4: Generalization across scales and backbones over six incomplete-modality settings.

Finally, we examine whether these gains generalize beyond the primary Hulu-Med-7B setting. As shown in Figure 4, MIRI achieves the strongest average diagnosis performance on Hulu-Med-4B, Hulu-Med-7B, and Qwen3-VL-8B in both MiF1 and MaF1. The improvement also extends to the prognostic tasks: relative to SFT, MIRI improves 9 of the 12 LOS and M30 metrics across the three backbones, while the remaining differences are all within 0.96 points. These results show that MIRI's gains generalize across model scales and backbone architectures.

6 Conclusion

We present MedGap, an early-ED benchmark for systematically evaluating Medical MLLMs across missing-modality settings involving CXR, ECG, and Lab, and MIRI, a two-stage framework for improving performance under missing modalities. Our evaluation shows that performance does not improve monotonically with modality availability, and that missing-modality effects vary across models and diagnosis targets. These findings indicate that missing-modality performance depends on both how models use available modalities and disease-specific modality relevance, rather than on modality count alone. MIRI achieves the best average performance across incomplete-modality settings, with the most consistent gains when a disease-relevant modality is missing. Further analysis shows that recovery improves access to missing-modality information, while integration enables its effective use, with recovery becoming more beneficial under more severe missingness. MIRI's gains further generalize across model scales and backbones.

AI USE STATEMENT

In this work, we used generative AI tools to assist with research ideation, code development and debugging, literature retrieval and discovery, and writing and editing portions of the manuscript. All AI-assisted content was reviewed and verified by the authors, who take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the benchmark construction, data preprocessing and splits, model architecture, training procedures, hyperparameters, and evaluation protocols in the main text and Appendix.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023. doi: 10.1109/CVPR52729.2023.01499.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8415–8424, 2021.

Wei Dai, Peilin Chen, Chanakya Ekbote, and Paul Pu Liang. QoQ-Med: Building multimodal clinical foundation models with domain-aware GRPO training. In *Advances in Neural Information Processing Systems*, 2025.

Yuan Gao, Sangwook Kim, Jianzhong You, and Chris McIntosh. ProbMED: A probabilistic framework for medical multimodal binding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20157–20167, 2025.

Brian Gow, Tom Pollard, Larry A. Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W. Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECG: Diagnostic electrocardiogram matched subset. *PhysioNet*, 2023. doi: 10.13026/4nqg-sb35. Version 1.0.

Nasir Hayat, Krzysztof J. Geras, and Farah E. Shamout. Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 479–503. PMLR, 2022.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Christopher Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 590–597, 2019. doi: 10.1609/aaai.v33i01.3301590.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 2021.

Sajid Javed, Arif Mahmood, Iyyakutti Iyappan Ganapathi, Fayaz Ali Dharejo, Naoufel Werghi, and Mohammed Bennamoun. CPLIP: Zero-shot learning for histopathology with comprehensive vision-language alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11450–11459, 2024. doi: 10.1109/CVPR52733.2024.01088.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, Jin Hao, Zijian Chen, Ruijia Wu, Tao Tang, Junhui Lv, Hongxia Xu, Hongwei Wang, Jun Xiao, Bin Feng, Fudong Zhu, Kenli Li, Weidi Xie, Jimeng Sun, Jian Wu, and Zuozhu Liu. Hulu-Med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

Alistair Johnson, Matt Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. MIMIC-CXR-JPG: Chest radiographs with structured labels. *PhysioNet*, 2019a. doi: 10.13026/8360-t248. Version 2.0.0.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and Steven Horng. MIMIC-IV-ED. *PhysioNet*, 2023a. doi: 10.13026/5ntk-km72. Version 2.2.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019b. doi: 10.1038/s41597-019-0322-0.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023b. doi: 10.1038/s41597-022-01899-x.

Lien P. Le, Thu Nguyen, Michael A. Riegler, Pål Halvorsen, and Binh T. Nguyen. Multimodal missing data in healthcare: A comprehensive review and future directions. *Computer Science Review*, 56:100720, 2025. doi: 10.1016/j.cosrev.2024.100720.

Hongyang Lei, Xiaolong Cheng, Qi Qin, Dan Wang, Huazhen Huang, Qingqing Gu, Yetao Wu, and Luo Ji. M3-jepa: Multimodal alignment via multi-gate moe based on the joint-embedding predictive architecture. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 33902–33917. PMLR, 2025.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, Siliang Tang, Jun Xiao, Hui Lin, Yueting Zhuang, and Beng Chin Ooi. HealthGPT: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37975–37995, 2025.

Tianling Liu, Lequan Yu, Tong Han, and Liang Wan. Context-driven missing-modality learning for robust medical diagnosis with image-tabular data. *arXiv preprint arXiv:2605.25968*, 2026.

Juan Miguel Lopez Alcaraz, Hjalmar Bouma, and Nils Strodthoff. Enhancing clinical decision support with physiological waveforms—a multimodal benchmark in emergency care. *Computers in Biology and Medicine*, 192:110196, 2025. doi: 10.1016/j.compbiomed.2025.110196.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 353–367. PMLR, 2023.

OpenAI. GPT-5.2, December 2025. OpenAI product release.

Caroline M. Patterson and Michael R. Loebinger. Community acquired pneumonia: assessment and treatment. *Clinical Medicine*, 12(3):283–286, 2012. doi: 10.7861/clinmedicine.12-3-283.

Vittorio Pipoli, Alessia Saporita, Federico Bolelli, Marcella Cornia, Lorenzo Baraldi, Costantino Grana, Rita Cucchiara, and Elisa Ficarra. MissRAG: Addressing the missing modality challenge in multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3215–3224, 2025. doi: 10.1109/ICCV51701.2025.00308.

Liangce Qi, Yusi Liu, Yuqin Li, Weili Shi, Guanyuan Feng, and Zhengang Jiang. A unified missing modality imputation model with inter-modality contrastive and consistent learning. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15967 of *Lecture Notes in Computer Science*, pp. 44–53. Springer Nature Switzerland, 2025.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Muyu Wang, Shiyu Fan, Yichen Li, Binyu Gao, Zhongrang Xie, and Hui Chen. Robust multi-modal fusion architecture for medical data with knowledge distillation. *Computer Methods and Programs in Biomedicine*, 260:108568, 2025a. doi: 10.1016/j.cmpb.2024.108568.

Muyu Wang, Shiyu Fan, Yichen Li, Zhongrang Xie, and Hui Chen. Missing-modality enabled multi-modal fusion architecture for medical data. *Journal of Biomedical Informatics*, 164:104796, 2025b. doi: 10.1016/j.jbi.2025.104796.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025c.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *Transactions on Machine Learning Research*, 2026.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

Wenfang Yao, Kejing Yin, William K. Cheung, Jia Liu, and Jing Qin. DrFuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2024. doi: 10.1609/aaai.v38i15.29578.

Kejing Yin, Haizhou Xu, Wenfang Yao, Chen Liu, Zijie Chen, Yui Haang Cheung, William K. Cheung, and Jing Qin. When does multimodal learning help in healthcare? a benchmark on EHR and chest x-ray fusion. *arXiv preprint arXiv:2602.23614*, 2026.

Yue Zhang, Chengtao Peng, Qiuli Wang, Dan Song, Kaiyan Li, and S. Kevin Zhou. Unified multi-modal image synthesis for missing modality imputation. *IEEE Transactions on Medical Imaging*, 44(1):4–18, 2025. doi: 10.1109/TMI.2024.3424785.

Yuting Zhang, Kaishen Yuan, Hao Lu, Yutao Yue, Jintai Chen, and Kaishun Wu. MedTVT-R1: A multimodal LLM empowering medical reasoning and diagnosis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

A Additional Experimental Details

All training experiments are conducted on 8 NVIDIA H800 GPUs using bfloat16 precision and a random seed of 42. For both training and evaluation, we retain at most 3 CXR images and 3 ECG images per encounter. We use greedy decoding for all generative evaluations and keep the input preprocessing of each backbone unchanged.

For supervised fine-tuning, we use 4-bit NF4 QLoRA with a LoRA rank of 16, scaling factor of 32, and dropout of 0.05. Models are trained with a learning rate of 2e-4 and an effective batch size of 32. For MIRI's missing-modality recovery, we use a hidden dimension of 1024 and 128, 128, and 64 latent tokens for CXR, ECG, and Lab, respectively. Within-modality training is performed for 300 steps, followed by 1,536 steps each for single-source and multi-source recovery, using a per-device batch size of 16, two gradient-accumulation steps, and a learning rate of 1e-4.

For downstream integration, each model is initialized from its corresponding full-modality SFT checkpoint and trained for one epoch. We use a per-device batch size of 1 with four gradient-accumulation steps, giving an effective batch size of 32 across 8 GPUs. The LoRA and modality-specific bridge learning rates are 5e-6 and 5e-5, respectively, and the bridge hidden dimension is 1024. Training uses 4-bit NF4 quantization with bfloat16 computation. MissRAG is applied on top of the same full-modality SFT checkpoint used by MIRI, with all methods evaluated under the same downstream protocol and generation settings.

B Missing-Modality Recovery Evaluation

B.1 Data and Modality-Specific Probe Labels

The recovery data are constructed from the training portion of the original patient-level split, without using any patients from the benchmark test set. These patients are further partitioned into recovery training, validation, and test sets with an 80/10/10 patient-level split and no patient overlap. Patients with incomplete modality-specific probe annotations are restricted to the recovery training set, ensuring that all validation and test encounters used for probing have evaluable labels. Importantly, these labels are used only for evaluating the learned and recovered representations and are not used as supervision for missing-modality recovery.

For CXR, we select 12 findings from the structured CheXpert labels provided with MIMIC-CXR-JPG (Johnson et al., 2019a), which are generated using the CheXpert labeler (Irvin et al., 2019): Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pleural Other, Pneumonia, and Pneumothorax. A finding is labeled positive when its CheXpert value is 1 and negative when it is 0. A missing finding is also treated as negative when *No Finding*=1; uncertain findings and otherwise unlabeled findings are excluded through the label mask.

For ECG, we construct 21 labels from the machine-generated diagnostic statements provided in MIMIC-IV-ECG (Gow et al., 2023): sinus tachycardia, sinus bradycardia, atrial fibrillation, atrial flutter, paced rhythm, PAC, PVC, RBBB, LBBB, LAFB, IVCD, first-degree AV block, short PR, left-axis deviation, right-axis deviation, low voltage, LVH, prolonged QT, ST elevation, ST depression, and ST–T abnormality. We use rule-based matching while excluding negated or historical mentions. Definite mentions are treated as positive, uncertain mentions are masked, and findings not mentioned in an otherwise available ECG report are treated as negative.

For Lab, we derive 14 abnormality groups from MIMIC-IV measurements (Johnson et al., 2023b) within the 12-hour encounter window, using the provided reference-range status to classify measurements as high, low, or normal and group clinically related abnormalities into physician-reviewed labels. The groups are renal dysfunction (high creatinine or BUN); electrolyte derangement (high or low sodium, potassium, chloride, calcium, magnesium, or phosphate); acid–base disorder (high or low bicarbonate, pH, pCO_2, or base excess); oxygenation abnormality (low arterial pO_2 or oxygen saturation); leukocyte abnormality (high or low WBC or absolute neutrophil count); anemia (low hemoglobin or hematocrit); thrombocytopenia (low platelet count); coagulation abnormality (high INR, PT, or PTT); hepatobiliary abnormality (high ALT, AST, total bilirubin, or alkaline phosphatase); myocardial injury (high troponin T); cardiac stress (high NT-proBNP); hyperlactatemia (high lactate); hyperglycemia (high glucose); and muscle injury (high creatine kinase). A label is evaluated only when at least one constituent measurement has a valid reference-range status.

B.2 PROBE PROTOCOL

For each recovery direction $X \rightarrow Y$, we freeze the backbone and MIRI and compare two representations using the same Y-specific probe labels. Obs. is formed from the original tokens of the observed modality or modalities X, whereas Rec. consists of the latent tokens of the missing modality Y recovered by MIRI; both are mean-pooled before probing, and pooled Obs. representations are concatenated when two modalities are observed. We train a linear probe on the recovery validation split and evaluate it on the patient-disjoint recovery test split. Each probe is trained for 30 epochs with AdamW using a learning rate of 10^{-3}, weight decay of 10^{-4}, and batch size 1024. We report macro-averaged AUROC and AUPRC over valid target labels.

Table 9: Additional missing-modality recovery results on Hulu-Med-4B and Qwen3-VL-8B.

		AUROC			AUPRC		
Model	Direction	Obs.	Rec.	Δ	Obs.	Rec.	Δ
Hulu-Med-4B	ECG→CXR	63.78	**66.96**	+3.18	21.33	**23.03**	+1.70
	Lab→CXR	67.35	**72.06**	+4.71	22.83	**27.78**	+4.96
	ECG+Lab→CXR	67.76	**74.32**	+6.56	24.51	**28.53**	+4.02
	CXR→ECG	62.79	**64.54**	+1.75	14.78	**15.65**	+0.87
	Lab→ECG	60.89	**61.68**	+0.79	12.21	**12.57**	+0.36
	CXR+Lab→ECG	64.05	**67.66**	+3.61	15.40	**17.09**	+1.69
	CXR→Lab	63.60	**65.59**	+2.00	60.83	**61.63**	+0.80
	ECG→Lab	**61.37**	60.74	-0.63	**58.50**	58.23	-0.28
	CXR+ECG→Lab	64.51	**67.51**	+3.00	61.66	**63.80**	+2.13
Qwen3-VL-8B	ECG→CXR	64.25	**66.85**	+2.60	21.04	**23.40**	+2.37
	Lab→CXR	66.89	**71.18**	+4.29	22.67	**26.16**	+3.49
	ECG+Lab→CXR	67.77	**73.86**	+6.09	23.40	**27.58**	+4.18
	CXR→ECG	61.31	**63.68**	+2.38	14.47	**15.42**	+0.95
	Lab→ECG	59.71	**59.94**	+0.23	11.94	**12.47**	+0.53
	CXR+Lab→ECG	62.59	**66.78**	+4.20	15.13	**17.23**	+2.10
	CXR→Lab	**63.58**	62.06	-1.53	**60.05**	59.12	-0.93
	ECG→Lab	59.98	**60.03**	+0.05	**57.46**	56.67	-0.79
	CXR+ECG→Lab	63.89	**64.73**	+0.84	60.48	**61.45**	+0.97

B.3 ADDITIONAL RECOVERY RESULTS

As shown in Table 9, the additional results exhibit a similar overall pattern across the two backbones. On Hulu-Med-4B, Rec. outperforms Obs. on both AUROC and AUPRC in eight of the nine recovery directions. On Qwen3-VL-8B, Rec. improves AUROC in eight of nine directions and AUPRC in seven of nine directions. The gains are particularly consistent for CXR and ECG recovery, while Lab recovery is less uniform across backbones. Together with the Hulu-Med-7B results in Table 7, these results show that the improved linear accessibility of missing-modality information is observed across all three evaluated backbones.

C PER-DIAGNOSIS MISSING-MODALITY ANALYSIS

Table 10 provides a finer-grained view of missing-modality degradation by decomposing the modality-paired evaluation across individual diagnosis targets. The effect of removing the same modality varies substantially across diagnoses and is often consistent across models. For example, removing CXR produces large F1 reductions for pneumonia in the CXR–ECG scope, whereas removing ECG has a much smaller effect on the same target. In contrast, arrhythmia is substantially more affected by removing ECG, while electrolyte disorders exhibit large degradation when Lab is removed in both the CXR–Lab and ECG–Lab scopes. These results show that the impact of modality loss cannot be fully characterized by the missing modality alone: it also depends on the clinical target being predicted. This target-level heterogeneity motivates the modality-paired design of MedGap, which complements aggregate availability evaluation by identifying where missing-modality degradation concentrates across clinical targets.

Table 11 further evaluates whether performance improvements persist across these target–missingness combinations. Averaged over the diagnosis-specific missing conditions within each modality pair, MIRI achieves 31.04 F1 for CXR–ECG, 39.99 for CXR–Lab, and 47.55 for ECG–Lab, exceeding the strongest corresponding baselines by 7.55, 8.11, and 11.44 points, respectively. The target-level results further illustrate these gains: when Lab is removed in the ECG–Lab scope, MIRI improves electrolyte-disorder F1 from 0.00 with SFT to 38.46; when ECG is removed, arrhythmia improves from 4.01 to 31.47. Similarly, when CXR is removed in the CXR–Lab scope,

Table 10: Disease-level zero-shot F1 sensitivity to modality removal. Values denote F1-point changes relative to paired input.

Model	*CXR–ECG*									
	Pneumonia		Heart Failure		Arrhythmia		Acute MI		Pleural Disease	
	−CXR	−ECG	−CXR	−ECG	−CXR	−ECG	−CXR	−ECG	−CXR	−ECG
GPT-5.2	-41.46	+2.22	-55.89	-1.52	-3.87	-43.23	-1.78	-3.57	-44.81	-12.32
InternVL3.5	-42.66	-2.96	-1.90	+5.36	-0.53	-26.62	+4.63	-4.41	-36.71	-9.33
Qwen3-VL	-32.65	+4.48	-39.31	-0.91	+0.40	-31.97	+2.42	0.00	-1.42	+11.54
HealthGPT	-42.21	-3.30	-51.04	-3.11	-0.45	-29.26	+7.67	0.00	-42.78	-6.76
Hulu-Med	-40.42	-5.43	-50.42	-1.07	+0.33	-21.63	-4.16	-13.04	-42.50	-3.71
Lingshu	-42.46	-1.26	-42.10	+4.77	+26.09	-6.59	+7.52	-5.41	-25.26	+2.07
QoQ-Med	-33.35	+0.63	-25.88	-1.63	+1.47	-37.44	0.00	0.00	0.00	+5.17

Model	*CXR–Lab*									
	Pneumonia		Heart Failure		Acute RF		Pleural Disease		Electrolyte	
	−CXR	−Lab	−CXR	−Lab	−CXR	−Lab	−CXR	−Lab	−CXR	−Lab
GPT-5.2	-49.95	-6.46	-2.86	-17.51	+0.92	-11.19	-43.73	-11.24	-1.96	-64.07
InternVL3.5	-16.05	+0.78	-5.52	-9.44	-3.37	-6.20	-15.37	-1.86	-0.93	-40.70
Qwen3-VL	-32.13	-0.50	+5.01	-16.83	+1.43	0.00	-5.43	+7.05	-2.33	-62.09
HealthGPT	-25.01	+2.56	+3.33	-6.32	-0.76	-0.63	-25.03	+8.34	+0.02	-53.05
Hulu-Med	-4.07	-3.03	-1.19	+2.65	+1.73	+9.32	-7.96	+20.01	+3.79	-41.10
Lingshu	-2.97	+3.71	-0.18	+1.37	+1.03	-2.04	-10.31	+5.80	-0.52	-49.57
QoQ-Med	-25.41	-1.98	-19.37	-6.75	-0.16	-1.16	-7.81	-2.64	-4.61	-50.63

Model	*ECG–Lab*							
	Heart Failure		Arrhythmia		Acute MI		Electrolyte	
	−ECG	−Lab	−ECG	−Lab	−ECG	−Lab	−ECG	−Lab
GPT-5.2	-1.06	-70.08	-52.58	-10.79	+2.33	-42.47	+0.52	-60.63
InternVL3.5	-7.03	-18.21	-3.34	-1.35	-0.73	+1.02	-0.78	-40.55
Qwen3-VL	-3.94	-64.18	-29.28	-0.35	+3.69	-23.97	-2.40	-62.16
HealthGPT	+19.73	-37.85	-32.25	-0.40	+9.97	-8.02	+0.67	-52.40
Hulu-Med	+2.62	-42.89	-8.35	+0.44	+3.51	-2.98	-1.85	-47.70
Lingshu	+4.46	-40.86	-36.18	+0.87	+3.39	-6.14	-2.41	-51.46
QoQ-Med	-2.01	-13.64	-32.82	+0.53	+5.50	-21.88	-0.84	-46.86

pneumonia and acute respiratory failure increase from 20.67 to 37.49 and from 14.29 to 23.29, respectively. Together, these results show that MedGap can characterize robustness at the level of specific modality–target relationships, while MIRI provides consistent aggregate improvements across all three clinically relevant modality pairs.

Table 11: Disease-level F1 under modality removal on Hulu-Med-7B. Columns indicate the removed modality; Avg. is the mean across diagnosis–missingness combinations within each modality pair.

Method	*CXR–ECG*										Avg.
	Pneumonia		Heart Failure		Arrhythmia		Acute MI		Pleural Disease		
	−CXR	−ECG	−CXR	−ECG	−CXR	−ECG	−CXR	−ECG	−CXR	−ECG	
Zero-shot	0.00	34.99	3.90	53.25	43.61	**21.65**	8.88	0.00	0.00	38.79	20.51
SFT	0.00	44.30	0.00	53.51	43.20	6.58	0.00	0.91	0.00	48.30	19.68
MissRAG	3.76	43.24	35.30	39.52	46.70	6.11	0.00	0.90	**0.48**	**58.87**	23.49
MIRI	**34.38**	**47.30**	**36.97**	**57.33**	**62.82**	3.31	**14.71**	**1.56**	0.00	52.01	**31.04**

Method	*CXR–Lab*										Avg.
	Pneumonia		Heart Failure		Acute RF		Pleural Disease		Electrolyte		
	−CXR	−Lab	−CXR	−Lab	−CXR	−Lab	−CXR	−Lab	−CXR	−Lab	
Zero-shot	33.95	34.99	49.41	53.25	5.97	13.56	10.82	38.79	45.85	0.96	28.76
SFT	20.67	44.30	64.44	53.51	14.29	2.08	0.00	48.30	71.26	0.00	31.88
MissRAG	35.42	43.24	40.88	39.52	10.22	0.00	**13.16**	**58.87**	**76.17**	0.00	31.75
MIRI	**37.49**	**47.30**	**68.33**	**57.33**	**23.29**	**22.38**	2.66	52.01	74.59	**14.51**	**39.99**

Method	*ECG–Lab*								Avg.
	Heart Failure		Arrhythmia		Acute MI		Electrolyte		
	−ECG	−Lab	−ECG	−Lab	−ECG	−Lab	−ECG	−Lab	
Zero-shot	49.41	3.90	34.82	43.61	15.37	8.88	45.85	0.00	25.23
SFT	64.44	0.00	4.01	43.20	**53.08**	0.00	71.26	0.00	29.50
MissRAG	40.88	35.30	**41.20**	46.70	48.65	0.00	**76.17**	0.00	36.11
MIRI	**68.33**	**36.97**	31.47	**62.82**	53.06	**14.71**	74.59	**38.46**	**47.55**

D REPRESENTATION VISUALIZATION

To qualitatively examine cross-modal recovery in MIRI, Figure 5 visualizes the target and predicted latent representations for CXR, ECG, and Lab. *Target* denotes the mean μ_m of the latent target distribution produced by the frozen EMA target encoder from the true modality, while *Predicted* denotes the mean $\hat{\mu}_m$ of the latent distribution predicted from the other two modalities without access to the target modality. Since UMAP operates on point representations, we visualize the distribution means while omitting the corresponding variances. Before UMAP projection, each set of latent tokens is mean-pooled into a single representation. Across all three modalities, the predicted representations substantially overlap with their corresponding target representations while retaining modality-specific structure. This qualitative alignment indicates that cross-modal prediction recovers the missing modality within its learned latent target space.

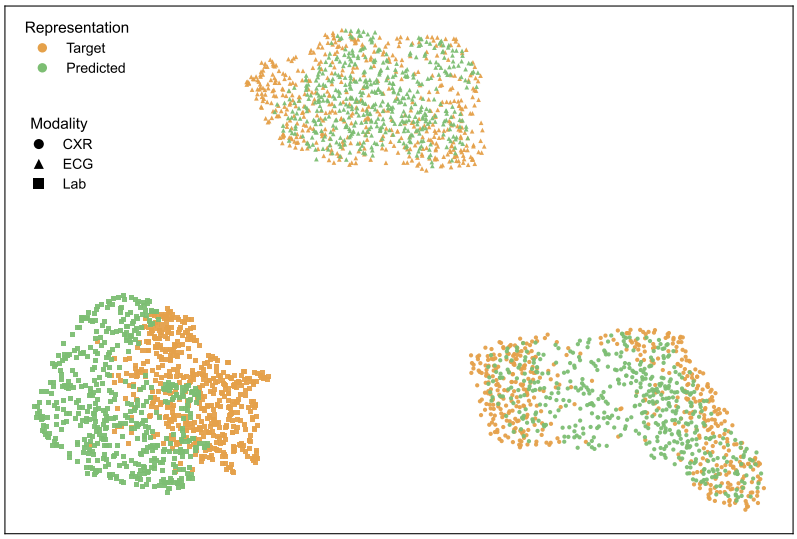

Figure 5: UMAP visualization of latent representations across CXR, ECG, and Lab.