MedMIX: Modality-Internal Expert Fusion for Multimodal Medical Diagnosis

Seungik Cho [*1] **Anqi Li** [*2] **Wei Qiu** [2]

Abstract

Multimodal clinical prediction faces three challenges: multiple foundation models (FMs) with complementary strengths per modality, pervasive missing modalities at training and test time, and sample-specific variation in modality contributions. We introduce **MedMIX**, a multimodal framework that combines intra-modality expert fusion, learned inter-modality fusion, and training-only large–small model collaboration for robust medical prediction under incomplete modalities. Within each modality, MedMIX aggregates complementary embeddings from multiple small expert models; across modalities, it performs learned fusion over available modalities; and during training, it leverages large teacher models to improve deployed representations without additional inference cost. Across three heterogeneous benchmarks (OpenI, MIMIC-IV-MM, and MMIST-ccRCC), MedMIX achieves consistently strong performance while remaining robust under controlled missing-modality perturbations, and further demonstrates sustained robustness under cross-cohort shift on MIMIC-III. These results highlight MedMIX as a practical framework that unifies within-modality expert collaboration, sample-specific cross-modality fusion, and efficient large–small model collaboration while remaining robust to incomplete modalities.

1. Introduction

Multimodal clinical prediction requires integrating heterogeneous sources—medical images, free-text reports, and structured time-series—each of which can be encoded by multiple foundation models (FMs) with complementary inductive biases (Acosta et al., 2022). Existing approaches either rely on a single large backbone per modality, which is brittle to distributional shift and costly at inference, or concatenate all modality embeddings without accounting for inter-expert complementarity within a modality (Chen et al., 2025; Soenksen et al., 2022). Two additional challenges compound this difficulty. First, clinical data are frequently incomplete: a given patient may lack a radiology report or a lateral view, so any fusion module must gracefully degrade under arbitrary missing patterns (Zhang et al., 2022; Wu et al., 2024). Second, even when all modalities are present, their predictive reliability varies per sample; fusing an uncertain modality's signal equally with a confident one degrades predictions (Han et al., 2022).

We propose **MedMIX**, a framework that addresses both challenges through three components: **(i)** a two-stage multimodal fusion architecture in which each modality is represented by a committee of small model experts (domain-specific, general-purpose, and retrieval-augmented), integrated by a shared intra-modality router, and **(ii)** a learned inter-modality fusion that assigns sample-specific weights over available modality logits via a learned scorer, with masked softmax re-normalization that explicitly excludes unavailable modalities and **(iii)** a training-only teacher–student distillation component that aligns per-modality student representations with Teacher Large Model (LM) encoder embeddings, improving sample efficiency without additional inference cost.

We evaluate on four benchmark datasets: **OpenI** (Demner-Fushman et al., 2016) for chest pathology classification, **MIMIC-IV-MM** (Johnson et al., 2024) built from MIMIC-IV resources for ICU outcome prediction, **MMIST ccRCC** (Mota et al., 2024) for cancer survival prediction, and **MIMIC-III** (Johnson et al., 2016) for external validation. Across four evaluation metrics, MedMIX outperforms **AdaCoMed** (Chen et al., 2025), **MeanAvg**, and **OneLLM** (Han et al., 2024). Our main contributions are:

- **MedMIX**, a multimodal clinical prediction framework combining modality-internal expert fusion—where complementary FM embeddings within each modality are aggregated via a learned MoE router—with learned inter-modality fusion that assigns sample-specific weights over available modality logits while handling arbitrary missing patterns via masked softmax, augmented by a training-only teacher–student

[*]Equal contribution [1]Department of Physics and Astronomy, Rice University, Texas, USA [2]Department of Electrical and Computer Engineering, Rice University, Texas, USA. Correspondence to: Wei Qiu <wq8@rice.edu>.

Proceedings of the ICML 2026 Workshop on Structured Data for Health, Seoul, Korea. 2026.

distillation objective.

- **Comprehensive robustness evaluation** under clinically realistic incomplete-modality settings, spanning train-time missing, test-time random missing-rate sweeps, and external cross-cohort validation on MIMIC-III, demonstrating that MedMIX maintains reliable performance across all conditions.
- **Inference efficiency analysis** shows that MedMIX achieves the highest macro-averaged predictive performance across three benchmarks while maintaining the best efficiency score in terms of trainable parameters and FLOPs relative to all baselines.

2. Method

Overview. Given an input sample with M modalities, each modality m is represented by K_m pre-extracted expert embeddings $\{e_k^{(m)}\}_{k=1}^{K_m}$ from lightweight FMs (biomedical, general-purpose, and retrieval-augmented). As illustrated in Figure 1, our model first aggregates experts within each modality via *Intra-Modality MoE*, then produces per-modality logits using independent classifier heads, and finally combines them through *learned inter-modality fusion* with masking of unavailable modalities. During training only, an optional teacher–student distillation objective aligns modality representations with Teacher LM encoders.

2.1. Intra-Modality Expert Aggregation

Each expert k in modality m is first refined by a lightweight residual adapter:

$$\begin{aligned} \tilde{e}_k^{(m)} &= e_k^{(m)} + \mathrm{Adapter}_k^{(m)}\big(e_k^{(m)}\big), \\ \mathrm{Adapter}_k^{(m)} &: \; d_k^{(m)} \to r_k^{(m)} \to d_k^{(m)}. \end{aligned} \tag{1}$$

where $r_k^{(m)} = \min(128, \max(32, d_k^{(m)}/16))$. We then project each refined embedding to a shared space as $z_k^{(m)} = \mathrm{Proj}_k^{(m)}\big(\tilde{e}_k^{(m)}\big) \in \mathbb{R}^d$ using a Linear–LayerNorm–GELU–Dropout block. Missing experts are masked by zeroing: $z_k^{(m)} \leftarrow z_k^{(m)} \cdot \mathbf{1}[\mathrm{mask}_k^{(m)}]$.

Routing uses a modality-specific shared per-expert scorer $s_k^{(m)} = \mathrm{Router}^{(m)}(z_k^{(m)}) \in \mathbb{R}$. Unavailable experts are masked to -10^4 before the softmax:

$$\begin{aligned} \boldsymbol{g}^{(m)} &= \mathrm{softmax}\big(\mathbf{s}^{(m)}\big), \\ z^{(m)} &= \Big(\sum_k g_k^{(m)} z_k^{(m)}\Big) \cdot \mathbf{1}\Big[\textstyle\sum_k \mathrm{mask}_k^{(m)} > 0\Big]. \end{aligned} \tag{2}$$

2.2. Learned Inter-Modality Fusion

Each modality representation $z^{(m)}$ is decoded into per-modality logits $\ell^{(m)} \in \mathbb{R}^C$ through an independent classifier head. A learned fusion score $s^{(m)} = \mathrm{Scorer}^{(m)}(z^{(m)})$ provides a sample-specific weighting signal. Unavailable modalities are assigned $\tilde{s}^{(m)} = -10^4$ before the softmax, so fusion weights are re-normalized over the observed modality set:

$$w^{(m)} = \mathrm{softmax}(\tilde{\mathbf{s}})^{(m)}, \qquad \ell_{\mathrm{fused}} = \sum_m w^{(m)} \ell^{(m)}. \tag{3}$$

The scorer is trained jointly with the task objective, so fusion weights adapt end-to-end to both modality content and availability.

2.3. Teacher–Student Distillation

To strengthen representations without increasing inference cost, we add a training-only distillation objective. For each modality m, a frozen teacher encoder produces embedding $z_T^{(m)}$, projected to dimension d via a learned linear head. Two alignment terms are used for each modality: cosine distillation, $\mathcal{L}_{\cos}^{(m)} = 1 - \cos\big(z^{(m)}, z_T^{(m)}\big)$, and relational knowledge distillation (RKD), which aligns pairwise distance structures. We define the per-modality distillation objective as $\mathcal{L}_{\mathrm{distill}}^{(m)} = \mathcal{L}_{\cos}^{(m)} + \lambda_{\mathrm{RKD}}\, \mathcal{L}_{\mathrm{RKD}}^{(m)}$, and the total distillation loss as $\mathcal{L}_{\mathrm{distill}} = \sum_m a^{(m)} \mathcal{L}_{\mathrm{distill}}^{(m)}$, where $a^{(m)} \in \{0, 1\}$ indicates whether modality m is available for the current sample. The distillation weight ramps from 0 to $\lambda_D^{\max}$ over T_D warmup epochs. The total loss is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{task}} + \lambda_D(t)\, \mathcal{L}_{\mathrm{distill}}. \tag{4}$$

The router uses a separate learning rate $\eta_{\mathrm{router}} = 0.3\, \eta_{\mathrm{base}}$ to stabilize early training.

3. Results

3.1. Experimental Setup

Datasets. We evaluate our framework on four representative multi-modal medical datasets spanning distinct clinical prediction settings. (1) **OpenI** is a public dataset containing 7,470 radiographs paired with 3,955 de-identified reports. We use chest X-ray images and radiology reports with a 0.65–0.15–0.20 train/validation/test split, and perform multi-label disease classification using labels derived from report text under a unified schema. (2) **MIMIC-IV-MM** integrates MIMIC-CXR-JPG, MIMIC-IV-Note, and MIMIC-IV, combining chest X-rays, radiology reports, and time-series electronic health record (EHR) data with 46 clinical variables. We use a 0.72–0.13–0.15 train/validation/test split. (3) **MMIST-ccRCC** focuses on clear cell renal cell carcinoma and includes computed tomography (CT), whole-slide pathology images (WSI), and clinical variables from 618 patients. We follow the official predefined split, and the downstream task is 12-month vital status prediction. (4) **MIMIC-III** serves as an external validation set to assess

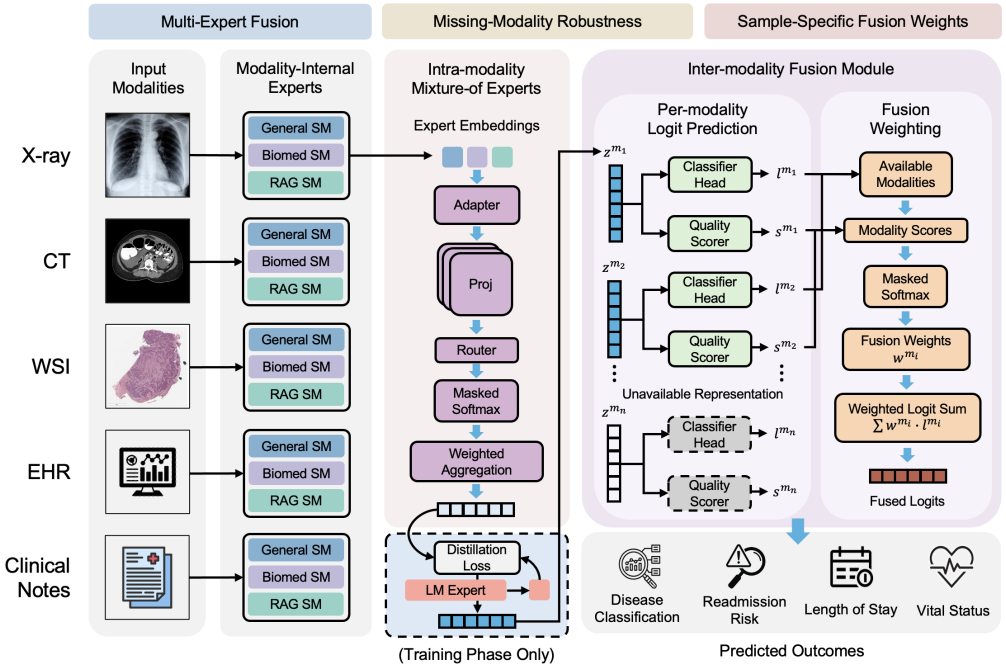

Figure 1. Overview of MedMIX. The framework combines modality-internal expert aggregation, learned inter-modality fusion with sample-specific modality weighting, and training-only teacher–student distillation for robust multimodal prediction under missing modalities.

cross-cohort generalizability, comprising ICU patients with radiology reports and time-series EHR data.

Baselines. We compare against: (1) **MeanAvg**, which averages all expert embeddings per modality before a shared multilayer perceptron (MLP); (2) **AdaCoMed**, the current LM large-small collaboration framework; (3) **OneLLM**, a pre-trained unified multimodal encoder fine-tuned on each task; and (4) modality-specific **Teacher LM** encoders used as strong reference models.

Expert Embeddings. For each modality we extract three complementary expert embeddings with small models (SM): a general-purpose vision or language SM (*General SM Expert*), a domain-specific biomedical SM (*Biomed SM Expert*), and a retrieval-augmented embedding (*RAG SM Expert*). The specific expert and teacher models used for each modality and dataset are listed in Appendix Table 1. Teacher LM embeddings are used for distillation only.

Implementation. All models are trained on NVIDIA A100 GPUs in PyTorch. We set the base learning rate to 1×10^{-5} and use a separate router learning-rate group with ratio 0.3 ($\eta_{\text{router}} = 0.3\,\eta_{\text{base}}$). Training runs for 200 epochs with early stopping on validation loss (patience 20). We report Accuracy (Acc), AUROC, AUPRC, and macro-F1 (mF1) as mean $\pm$ standard deviation over 5 seeds.

3.2. Main Results

Table 1 presents the main comparison on OpenI, MIMIC-IV-MM, and MMIST-ccRCC. Overall, our modality-internal expert fusion consistently outperforms strong baselines, including AdaCoMed and OneLLM, across AUROC, AUPRC, mF1, and Acc. Relative to the strongest baseline for each metric, the model achieves dataset-level relative gains of 0.6%–4.0% on OpenI, 1.8%–9.1% on MMIST-ccRCC, and 0.34%–3.47% on MIMIC-IV-MM. We therefore interpret the main comparison as evidence that modality-internal expert collaboration is beneficial on these three tasks, rather than as proof of universal superiority across all multimodal settings.

Table 1. Main results on OpenI, MIMIC-IV-MM, and MMIST-ccRCC. Entries labeled `*-Large` denote modality-specific teacher large models used as single-modality reference baselines.

Dataset	Method	AUROC	AUPRC	mF1	Acc
OpenI	Front-Large	0.7482 ± 0.0014	0.3155 ± 0.0037	0.3427 ± 0.0043	0.7626 ± 0.0211
	Side-Large	0.7104 ± 0.0059	0.2752 ± 0.0071	0.2977 ± 0.0058	0.7466 ± 0.0469
	Note-Large	0.9385 ± 0.0056	0.7481 ± 0.0064	0.6980 ± 0.0098	0.9363 ± 0.0034
	MeanAvg	0.8904 ± 0.0015	0.5485 ± 0.0091	0.7201 ± 0.0040	0.8764 ± 0.0083
	AdaCoMed	0.9260 ± 0.0027	0.6772 ± 0.0102	0.6297 ± 0.0151	0.9193 ± 0.0092
	OneLLM	0.8784 ± 0.0076	0.5635 ± 0.0230	0.7124 ± 0.0127	0.8732 ± 0.0173
	MedMIX	**0.9570 ± 0.0006**	**0.7780 ± 0.0047**	**0.7246 ± 0.0061**	**0.9463 ± 0.0019**
MIMIC-IV-MM	XRay-Large	0.6487 ± 0.0034	0.3713 ± 0.0041	0.4508 ± 0.0033	0.5393 ± 0.0140
	Text-Large	0.6634 ± 0.0031	0.3948 ± 0.0046	0.4524 ± 0.0030	0.5536 ± 0.0332
	TimeSeries-Large	0.5349 ± 0.0018	0.2738 ± 0.0018	0.4145 ± 0.0002	0.3009 ± 0.0007
	MeanAvg	0.7004 ± 0.0006	0.4147 ± 0.0017	0.6208 ± 0.0015	0.7111 ± 0.0120
	AdaCoMed	0.7059 ± 0.0013	0.4405 ± 0.0024	0.6275 ± 0.0025	0.7221 ± 0.0043
	OneLLM	0.5921 ± 0.0036	0.3344 ± 0.0074	0.5453 ± 0.0152	0.6521 ± 0.0337
	MedMIX	**0.7168 ± 0.0019**	**0.4586 ± 0.0019**	**0.6375 ± 0.0035**	**0.7352 ± 0.0106**
MMIST-ccRCC	CT-Large	0.6245 ± 0.0098	0.8811 ± 0.0086	0.4880 ± 0.0024	0.5521 ± 0.0033
	WSI-Large	0.5797 ± 0.0646	0.8945 ± 0.0201	0.5263 ± 0.0520	0.8331 ± 0.0500
	Note-Large	0.7372 ± 0.0077	0.9251 ± 0.0038	0.7127 ± 0.0193	0.8843 ± 0.0174
	MeanAvg	0.7075 ± 0.0064	0.9290 ± 0.0048	0.5867 ± 0.0814	0.7521 ± 0.1625
	AdaCoMed	0.7407 ± 0.0094	0.9476 ± 0.0026	0.6813 ± 0.0097	0.8711 ± 0.0132
	OneLLM	0.7445 ± 0.0131	0.9398 ± 0.0057	0.6983 ± 0.0042	0.8709 ± 0.0213
	MedMIX	**0.8121 ± 0.0068**	**0.9673 ± 0.0019**	**0.7257 ± 0.0146**	**0.9025 ± 0.0096**

3.3. Missing Modality Experiments

We evaluate MedMIX under clinically realistic missing-modality conditions to assess robustness when modality availability is incomplete.

Training-time missing. We assume training data contain missing modalities and apply a fixed missing rate $p \in \{0.3, 0.5, 0.7\}$. For each run, we drop one modality at a time during training and evaluate on the standard test split. The MIMIC-IV-MM one-modality-at-a-time results are summarized in Table 2, while full cross-dataset results are provided in Appendix Table 2. On MIMIC-IV-MM, performance remains stable across missing rates and dropped modalities, with AUROC ranging from 0.7109 to 0.7170.

Table 2. MIMIC-IV-MM results under train-time one-modality-at-a-time missingness.

Drop p	Modality	AUROC	AUPRC	mF1	Acc
0.3	Front CXR	0.7146±0.0020	0.4514±0.0037	0.6205±0.0112	0.6958±0.0133
	Lateral CXR	0.7152±0.0016	0.4522±0.0020	0.6313±0.0027	0.7201±0.0092
	EHR	0.7142±0.0026	0.4498±0.0025	0.6286±0.0023	0.7147±0.0073
	Note	0.7156±0.0011	0.4511±0.0023	0.6313±0.0033	0.7175±0.0077
0.5	Front CXR	0.7150±0.0014	0.4530±0.0016	0.6263±0.0105	0.7043±0.0117
	Lateral CXR	0.7154±0.0019	0.4523±0.0032	0.6303±0.0040	0.7134±0.0113
	EHR	0.7125±0.0015	0.4451±0.0023	0.6285±0.0029	0.7201±0.0087
	Note	0.7170±0.0014	0.4533±0.0010	0.6307±0.0063	0.7233±0.0136
0.7	Front CXR	0.7123±0.0014	0.4506±0.0024	0.6259±0.0052	0.7097±0.0101
	Lateral CXR	0.7155±0.0019	0.4524±0.0030	0.6301±0.0039	0.7137±0.0105
	EHR	0.7109±0.0021	0.4438±0.0027	0.6281±0.0022	0.7141±0.0113
	Note	0.7165±0.0016	0.4540±0.0032	0.6343±0.0042	0.7200±0.0044

Test-time missing. We evaluate robustness under random test-time modality dropout with missing rates $r \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$. For each sample, one randomly selected modality is dropped with probability r. The MIMIC-IV-MM sweep is summarized in Table 3, while full test-missing-rate results are reported in Appendix Table 3. Performance degrades as expected with increasing missingness, but the decline remains controlled: on OpenI, AUROC drops from 0.9396 ($r = 0.1$) to 0.8630 ($r = 1.0$), whereas on MIMIC-IV-MM it changes only from 0.7119 to 0.6914 over the same range.

Table 3. MIMIC-IV-MM results under test-time random missing-rate sweep.

Missing Rate	AUROC	AUPRC	mF1	Acc
0.1	0.7119±0.0031	0.4457±0.0033	0.6230±0.0089	0.7061±0.0120
0.3	0.7069±0.0031	0.4386±0.0027	0.6195±0.0090	0.7020±0.0121
0.5	0.7024±0.0024	0.4348±0.0033	0.6166±0.0080	0.6985±0.0107
0.7	0.6977±0.0043	0.4275±0.0041	0.6140±0.0070	0.6946±0.0110
1.0	0.6914±0.0021	0.4222±0.0030	0.6090±0.0061	0.6895±0.0103

3.4. Structural Ablation

To isolate the contribution of each major architectural component, we perform a structural ablation across OpenI, MIMIC-IV-MM, and MMIST-ccRCC. Appendix Table 4 compares MedMIX against variants without distillation, without intra-modality fusion, and without inter-modality fusion while retaining a single modality. Across all three datasets, MedMIX achieves the strongest or near-strongest performance on most metrics, indicating that the gains arise from the complementary effect of distillation, intra-modality expert aggregation, and inter-modality fusion rather than from any single component alone. Notably, removing intra-modality fusion consistently reduces AUROC and mF1, while collapsing to a single retained modality causes larger and dataset-dependent drops, underscoring the importance of adaptive multimodal fusion.

3.5. External Validation

To assess cross-cohort generalization, we train the model on MIMIC-IV-MM and directly test on MIMIC-III using the same test protocol. As shown in Appendix Table 5, MedMIX outperforms MeanAvg, OneLLM, and AdaCoMed across all four metrics (AUROC, AUPRC, mF1, and Acc), indicating improved robustness under distribution shift.

3.6. Efficiency Analysis

We evaluate inference efficiency using trainable parameter count, FLOPs, and peak GPU memory. As summarized in Appendix Table 6, we define $\mathrm{Perf} = (\mathrm{AUROC} + \mathrm{AUPRC} + \mathrm{mF1} + \mathrm{Acc})/4$, Cost as the geometric mean of resource ratios relative to MedMIX, and $\mathrm{EffScore} = (\mathrm{Perf}/\mathrm{Perf}_0)/\mathrm{Cost}$. MedMIX achieves the strongest predictive performance on OpenI ($\mathrm{Perf} = 0.851$), MIMIC-IV-MM ($\mathrm{Perf} = 0.629$), and MMIST-ccRCC ($\mathrm{Perf} = 0.852$), while retaining the highest efficiency score ($\mathrm{EffScore} = 1.000$) in all three settings. The Average block further shows the best macro-averaged performance ($\mathrm{Perf} = 0.777$) and efficiency score, supporting MedMIX as the most balanced choice for practical multimodal deployment.

4. Conclusion

In this paper, we mainly focus on robust multimodal medical diagnosis under heterogeneous expert models and incomplete clinical modalities. This goal is achieved by the proposed MedMIX framework with three core components, i.e., modality-internal expert fusion, missing-aware inter-modality fusion, and training-only large–small model distillation. The experimental results demonstrate the effectiveness and efficiency of MedMIX across diverse clinical prediction benchmarks. In addition, the validity of the core components designed in MedMIX is shown by the ablation studies and missing-modality experiments. In the future, we will put effort into extending MedMIX to more structured real-world missingness patterns and broader clinical deployment settings.

References

Acosta, J. N., Falcone, G. J., Rajpurkar, P., and Topol, E. J. Multimodal biomedical AI. *Nature Medicine*, 28(9):1773–1784, 2022. doi: 10.1038/s41591-022-01981-2.

Chen, W., Zhao, Z., Yao, J., Zhang, Y., Bu, J., and Wang, H. Multi-modal medical diagnosis via large-small model collaboration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30763–30773, 2025.

Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., and McDonald, C. J. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2): 304–310, 2016.

Han, J., Gong, K., Zhang, Y., Wang, J., Zhang, K., Lin, D., Qiao, Y., Gao, P., and Yue, X. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26584–26595, 2024.

Han, Z., Zhang, C., Fu, H., and Zhou, J. T. Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2551–2566, 2022. doi: 10.1109/TPAMI.2022.3171983.

Johnson, A., Pollard, T., and Mark, R. MIMIC-III Clinical Database. *PhysioNet*, September 2016. doi: 10.13026/C2XW26. URL https://doi.org/10.13026/C2XW26. Version 1.4.

Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L. A., and Mark, R. MIMIC-IV. *PhysioNet*, October 2024. doi: 10.13026/kpb9-mt58. URL https://doi.org/10.13026/kpb9-mt58. Version 3.1.

Mota, T., Verdelho, M. R., Araújo, D. J., Bissoto, A., Santiago, C., and Barata, C. Mmist-ccrcc: A real world medical dataset for the development of multi-modal systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2395–2403, 2024.

Soenksen, L. R., Ma, Y., Zeng, C., Boussioux, L., Villalobos Carballo, K., Na, L., Wiberg, H. M., Li, M. L., Fuentes, I., and Bertsimas, D. Integrated multimodal artificial intelligence framework for healthcare applications. *npj Digital Medicine*, 5(1):149, 2022. doi: 10.1038/s41746-022-00689-4.

Wu, Z., Dadu, A., Tustison, N., Avants, B., Nalls, M., Sun, J., and Faghri, F. Multimodal patient representation learning with missing modalities and labels. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Je5SHCKpPa.

Zhang, C., Chu, X., Ma, L., Zhu, Y., Wang, Y., Wang, J., and Zhao, J. M^{3}Care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2418–2428, 2022. doi: 10.1145/3534678.3539388.

Appendix A. Additional Result Tables

Appendix Table 1. Expert and teacher FMs per modality and dataset.

Dataset	Modality	Biomed	General	RAG	Teacher FM
OpenI	Frontal CXR	BiomedCLIP	DINOv2-B	MAIRA-2+MedCPT+Qwen2-0.5B	DINOv3 ViT-L
	Lateral CXR	BiomedCLIP	DINOv2-B	MAIRA-2+MedCPT+Qwen2-0.5B	DINOv3 ViT-L
	Radiology Note	GatorTron-B	Qwen2-0.5B	MedCPT+Qwen2-0.5B	Qwen2-7B
MIMIC-IV-MM	Frontal CXR	BiomedCLIP	DINOv2-B	MAIRA-2+MedCPT+Qwen2-0.5B	DINOv3 ViT-L
	Lateral CXR	BiomedCLIP	DINOv2-B	MAIRA-2+MedCPT+Qwen2-0.5B	DINOv3 ViT-L
	EHR	MIRA	Chronos-S	MedCPT+Qwen2-0.5B	Chronos-L
	Radiology Note	GatorTron-B	Qwen2-0.5B	MedCPT+Qwen2-0.5B	Qwen2-7B
MMIST-ccRCC	CT	Merlin	DINOv2-B	Merlin+MedCPT+Qwen2-0.5B	RadFM
	WSI	BiomedCLIP	DINOv2-B	HistGen+MedCPT+Qwen2-0.5B	DINOv3 ViT-L
	Clinical Note	GatorTron-B	Qwen2-0.5B	MedCPT+Qwen2-0.5B	Qwen2-7B
MIMIC-III	EHR	MIRA	Chronos-S	MedCPT+Qwen2-0.5B	Chronos-L
	Radiology Note	GatorTron-B	Qwen2-0.5B	MedCPT+Qwen2-0.5B	Qwen2-7B

Appendix Table 2. Train-time missing-modality results under one-modality-at-a-time ablation across datasets and drop rates.

Drop p	Dataset	Modality	AUROC	AUPRC	mF1	Acc
0.3	OpenI	Front CXR	0.9458±0.0017	0.7739±0.0083	0.7121±0.0129	0.9412±0.0091
		Side CXR	0.9465±0.0019	0.7770±0.0095	0.7194±0.0127	0.9453±0.0045
		Note	0.9548±0.0005	0.7911±0.0038	0.7161±0.0056	0.9416±0.0016
	MIMIC-IV-MM	Front CXR	0.7146±0.0020	0.4514±0.0037	0.6205±0.0112	0.6958±0.0133
		Lateral CXR	0.7152±0.0016	0.4522±0.0020	0.6313±0.0027	0.7201±0.0092
		EHR	0.7142±0.0026	0.4498±0.0025	0.6286±0.0023	0.7147±0.0073
		Note	0.7156±0.0011	0.4511±0.0023	0.6313±0.0033	0.7175±0.0077
	MMIST-ccRCC	CT	0.7930±0.0104	0.9622±0.0029	0.7278±0.0065	0.9041±0.0041
		WSI	0.7941±0.0139	0.9628±0.0033	0.7067±0.0194	0.8959±0.0112
		Note	0.7778±0.0118	0.9594±0.0026	0.7055±0.0089	0.8926±0.0000
0.5	OpenI	Front CXR	0.9464±0.0018	0.7766±0.0091	0.7144±0.0156	0.9419±0.0098
		Side CXR	0.9466±0.0018	0.7782±0.0067	0.7142±0.0063	0.9426±0.0051
		Note	0.9542±0.0007	0.7870±0.0028	0.7138±0.0089	0.9404±0.0032
	MIMIC-IV-MM	Front CXR	0.7150±0.0014	0.4530±0.0016	0.6263±0.0105	0.7043±0.0117
		Lateral CXR	0.7154±0.0019	0.4523±0.0032	0.6303±0.0040	0.7134±0.0113
		EHR	0.7125±0.0015	0.4451±0.0023	0.6285±0.0029	0.7201±0.0087
		Note	0.7170±0.0014	0.4533±0.0010	0.6307±0.0063	0.7233±0.0136
	MMIST-ccRCC	CT	0.7608±0.0152	0.9520±0.0056	0.7291±0.0059	0.9025±0.0062
		WSI	0.7993±0.0154	0.9640±0.0035	0.7082±0.0136	0.8942±0.0062
		Note	0.7604±0.0135	0.9561±0.0031	0.6854±0.0082	0.8843±0.0091
0.7	OpenI	Front CXR	0.9457±0.0021	0.7738±0.0102	0.7045±0.0175	0.9358±0.0101
		Side CXR	0.9461±0.0020	0.7742±0.0094	0.7165±0.0126	0.9442±0.0040
		Note	0.9513±0.0015	0.7732±0.0035	0.6984±0.0016	0.9329±0.0028
	MIMIC-IV-MM	Front CXR	0.7123±0.0014	0.4506±0.0024	0.6259±0.0052	0.7097±0.0101
		Lateral CXR	0.7155±0.0019	0.4524±0.0030	0.6301±0.0039	0.7137±0.0105
		EHR	0.7109±0.0021	0.4438±0.0027	0.6281±0.0022	0.7141±0.0113
		Note	0.7165±0.0016	0.4540±0.0032	0.6343±0.0042	0.7200±0.0044
	MMIST-ccRCC	CT	0.7531±0.0084	0.9465±0.0051	0.7035±0.0425	0.8909±0.0160
		WSI	0.7972±0.0175	0.9641±0.0038	0.7137±0.0217	0.8942±0.0151
		Note	0.7350±0.0130	0.9473±0.0061	0.6383±0.0749	0.8215±0.1341

Appendix Table 3. Test-time missing-modality results under random missing-rate sweep, grouped by dataset.

Dataset	Missing Rate	AUROC	AUPRC	mF1	Acc
OpenI	0.1	0.9396±0.0024	0.7593±0.0104	0.7004±0.0134	0.9407±0.0049
	0.3	0.9258±0.0061	0.7299±0.0127	0.6753±0.0142	0.9357±0.0051
	0.5	0.9063±0.0034	0.6871±0.0106	0.6364±0.0111	0.9303±0.0042
	0.7	0.8899±0.0045	0.6599±0.0091	0.6108±0.0063	0.9261±0.0028
	1.0	0.8630±0.0057	0.6136±0.0103	0.5669±0.0147	0.9195±0.0029
MIMIC-IV-MM	0.1	0.7119±0.0031	0.4457±0.0033	0.6230±0.0089	0.7061±0.0120
	0.3	0.7069±0.0031	0.4386±0.0027	0.6195±0.0090	0.7020±0.0121
	0.5	0.7024±0.0024	0.4348±0.0033	0.6166±0.0080	0.6985±0.0107
	0.7	0.6977±0.0043	0.4275±0.0041	0.6140±0.0070	0.6946±0.0110
	1.0	0.6914±0.0021	0.4222±0.0030	0.6090±0.0061	0.6895±0.0103
MMIST-ccRCC	0.1	0.7895±0.0119	0.9623±0.0035	0.7019±0.0205	0.8926±0.0117
	0.3	0.7440±0.0117	0.9510±0.0032	0.6805±0.0214	0.8876±0.0112
	0.5	0.7373±0.0293	0.9477±0.0082	0.6785±0.0500	0.8827±0.0184
	0.7	0.7420±0.0465	0.9498±0.0105	0.6574±0.0519	0.8661±0.0224
	1.0	0.6910±0.0542	0.9358±0.0150	0.6312±0.0829	0.8628±0.0259

Appendix Table 4. Structural ablation results on OpenI, MIMIC-IV-MM, and MMIST-ccRCC.

Dataset	Setting	AUROC	AUPRC	mF1	Acc
OpenI	MedMIX	**0.9570±0.0006**	**0.7780±0.0047**	**0.7246±0.0061**	**0.9463±0.0019**
	w/o distillation	0.9456±0.0005	0.7705±0.0049	0.7101±0.0093	0.9414±0.0023
	w/o intra-modality fusion	0.9478±0.0015	0.7571±0.0046	0.6872±0.0052	0.9328±0.0058
	w/o inter-modality fusion (keep Front CXR)	0.7752±0.0056	0.3599±0.0107	0.3480±0.0135	0.7286±0.0372
	w/o inter-modality fusion (keep Side CXR)	0.7464±0.0057	0.3106±0.0121	0.3332±0.0103	0.7508±0.0248
	w/o inter-modality fusion (keep Note)	0.9358±0.0049	0.7287±0.0149	0.6709±0.0126	0.9264±0.0058
MIMIC-IV-MM	MedMIX	**0.7168±0.0019**	**0.4586±0.0019**	**0.6375±0.0035**	**0.7352±0.0106**
	w/o distillation	0.7095±0.0049	0.4420±0.0063	0.6205±0.0057	0.6894±0.0091
	w/o intra-modality fusion	0.6757±0.0017	0.4006±0.0022	0.6053±0.0020	0.6950±0.0167
	w/o inter-modality fusion (keep Front CXR)	0.6811±0.0011	0.4026±0.0019	0.5976±0.0042	0.6769±0.0170
	w/o inter-modality fusion (keep Lateral CXR)	0.5305±0.0007	0.2685±0.0004	0.4326±0.0018	**0.7253±0.0271**
	w/o inter-modality fusion (keep EHR)	0.5905±0.0046	0.3167±0.0060	0.5036±0.0063	0.7039±0.0020
	w/o inter-modality fusion (keep Note)	0.6737±0.0010	0.4000±0.0012	0.6012±0.0021	0.6999±0.0134
MMIST-ccRCC	MedMIX	**0.8121±0.0068**	**0.9673±0.0019**	0.7257±0.0146	**0.9025±0.0096**
	w/o distillation	0.7615±0.0167	0.9526±0.0051	0.6407±0.0270	0.8099±0.0193
	w/o intra-modality fusion	0.7607±0.0122	0.9548±0.0050	0.6903±0.0189	0.8942±0.0033
	w/o inter-modality fusion (keep CT)	0.6505±0.0062	0.9177±0.0029	0.4770±0.0103	0.6761±0.0166
	w/o inter-modality fusion (keep WSI)	0.5056±0.0107	0.8744±0.0115	0.4727±0.0162	0.8364±0.0128
	w/o inter-modality fusion (keep Note)	0.7527±0.0085	0.9394±0.0052	**0.7320±0.0107**	0.8926±0.0074

Appendix Table 5. External-validation results on MIMIC-III.

Method	AUROC	AUPRC	mF1	Acc
MeanAvg	0.5387 ± 0.0034	0.1324 ± 0.0019	0.3517 ± 0.0159	0.5058 ± 0.0348
AdaCoMed	0.5240 ± 0.0114	0.1359 ± 0.0064	0.3916 ± 0.0243	0.6114 ± 0.0433
OneLLM	0.5323 ± 0.0093	0.1291 ± 0.0056	0.3662 ± 0.0152	0.6073 ± 0.0566
MedMIX	**0.5718 ± 0.0049**	**0.1508 ± 0.0034**	**0.4560 ± 0.0100**	**0.7371 ± 0.0257**

Appendix Table 6. Efficiency comparison across OpenI, MIMIC-IV-MM, and MMIST-ccRCC.

Dataset	Method	Params (B)	FLOPs (G)	Peak GPU Mem. (GB)	Perf	EffScore
OpenI	MeanAvg	2.85	773.74	14.44	0.759	0.891
	AdaCoMed	8.20	909.39	14.58	0.788	0.615
	OneLLM	8.63	1493.23	14.80	0.757	0.489
	MedMIX	**2.85**	**773.74**	**14.44**	**0.851**	**1.000**
MIMIC-IV-MM	MeanAvg	2.87	1751.50	14.21	0.612	0.959
	AdaCoMed	9.37	3608.26	15.20	0.621	0.504
	OneLLM	15.24	5132.74	18.20	0.536	0.310
	MedMIX	**2.86**	**1751.50**	**14.21**	**0.637**	**1.000**
MMIST-ccRCC	MeanAvg	3.47	43340.61	1.78	0.744	0.872
	AdaCoMed	9.88	41664.49	14.41	0.810	0.338
	OneLLM	8.64	59466.65	14.70	0.813	0.313
	MedMIX	**3.46**	**43340.61**	**1.78**	**0.852**	**1.000**
Average	MeanAvg	3.06	15288.62	10.14	0.705	0.903
	AdaCoMed	9.15	15394.05	14.73	0.740	0.580
	OneLLM	10.84	22030.87	15.90	0.702	0.450
	MedMIX	**3.06**	**15288.62**	**10.14**	**0.780**	**1.000**