Class-Dependent Miscalibration Severely Degrades Selective Prediction in Multimodal Clinical Prediction Models

L. Julián Lechuga López[1,2] LEOPOLDO.LECHUGA@NYU.EDU
Farah E. Shamout[1,2] FARAH.SHAMOUT@NYU.EDU
Tim G. J. Rudner[3] TIM.RUDNER@UTORONTO.CA

[1] *NYU Abu Dhabi, Abu Dhabi, UAE*
[2] *New York University, New York, NY, USA*
[3] *University of Toronto, Toronto, ON, Canada*

Abstract

As artificial intelligence systems transition from research to clinical deployment, ensuring their reliability becomes critical for clinical decision-making tasks, as incorrect predictions can have serious consequences. Deploying AI in healthcare therefore requires prediction systems with robust safeguards against error, such as selective prediction, where uncertain predictions are deferred to human experts for review. In this study, we carefully evaluate the reliability of uncertainty-based selective prediction for multilabel clinical condition classification using multimodal data. Our findings show that models often exhibit severe class-dependent miscalibration causing predictive performance to *degrade* under uncertainty-guided selective prediction—attributing high uncertainty to correct predictions and low uncertainty to incorrect predictions. Our evaluation highlights fundamental shortcomings of commonly used evaluation metrics for clinical AI. To address these shortcomings, we propose practical recommendations for calibration-aware model assessment and selective prediction design, offering a pathway to safer, more reliable AI systems that clinicians and patients can trust.

Keywords: Calibration; Multimodal learning; Selective prediction.

Data and Code Availability This study uses publicly available benchmark datasets: MIMIC-IV Johnson et al. (2021) and MIMIC-CXR Johnson et al. (2019). Both datasets are available to researchers through the PhysioNet platform, subject to credentialed access and data use agreements. We plan to make our code available in a future release.

Institutional Review Board (IRB) This research uses only publicly available, de-identified datasets (MIMIC-IV and MIMIC-CXR) that do not constitute human subjects research. Therefore, IRB approval was not required.

1. Introduction

Machine learning is increasingly embedded into the healthcare sector to support clinical decision-making, improve diagnostic performance, accelerate drug discovery, and optimize patient management (Paul et al., 2025; Hanna et al., 2025). From enhancing disease detection (Mall et al., 2022) to personalizing treatment plans (Agarwal, 2024), machine learning-driven systems have demonstrated transformative potential. However, in high-stakes clinical environments, exhibiting good performance, such as high accuracy, is not sufficient. Models in healthcare must provide fail-safe mechanisms to ensure they are reliable and interpretable, to allow clinicians to recognize when predictions are unreliable so they can intervene accordingly (Javed et al., 2024).

One such mechanism is selective prediction, where a model can abstain from making a prediction and request a review by a human expert if its predictive uncertainty is high (Geifman and El-Yaniv, 2017). By abstaining in cases of high uncertainty, selective prediction provides a fail-safe mechanism against critical errors that can endanger patients. However, for selective prediction to improve model safety and robustness, models must be able to provide reliable uncertainty estimates. If uncertainty estimates are systematically miscalibrated, selective prediction may fail to identify risky cases, undermining its value in clinical decision-making.

Unfortunately, evidence from unimodal domains such as computer vision and time-series analysis shows that many models systematically misestimate uncertainty, often being overconfident when wrong (Guo et al., 2017; Morey et al., 2025; Deng et al., 2025). In healthcare settings, these challenges may be further compounded when having to deal with multimodal data, where heterogeneous sources such as electronic health records (EHR), imaging, and text are integrated to provide a holistic view of a patient's state (Simon et al., 2025). While multimodal methods frequently improve discrimination metrics, their calibration properties, and thus their reliability for selective prediction, remain poorly understood (Zhao et al., 2024).

In this paper, we study whether state-of-the-art multimodal clinical condition classification models provide reliable uncertainty estimates that enable effective selective prediction. We find that although multimodal models outperform unimodal baselines on standard evaluation metrics, multimodal fusion substantially degrades selective prediction, calling into question its suitability as a fail-safe mechanism in safety-critical healthcare settings (Table 1).

To understand the roots of this failure, we perform a careful analysis across unimodal and multimodal models trained on one of the largest publicly-available multimodal datasets, consisting of EHR and chest X-ray data. This comparison reveals that multimodal fusion does not consistently improve calibration and, in many cases, exacerbates class-dependent miscalibration. In particular, we consider class-dependent expected calibration errors and find that positive-class predictions are systematically overconfident, especially for conditions underrepresented in the data, undermining the effectiveness of selective prediction.

Finally, we investigate whether simple loss-upweight training strategies can mitigate these issues and find that they offer only modest improvements, falling short of resolving class-dependent miscalibration in multimodal models. Our findings indicate that state-of-the-art multimodal models cannot deliver selective prediction as a reliable fail-safe mechanism, underscoring calibration-aware selective prediction as a key open challenge for safe deployment.

To summarize, our key contributions are:

1. We show that selective prediction in multimodal clinical condition classification models degrades performance across a wide range of clinical conditions.

2. We investigate the causes of this failure by comparing unimodal and multimodal models and show that the degradation in selective prediction is driven by class-dependent miscalibration.

3. We introduce a simple label-dependent loss upweighting scheme that partially corrects class-dependent miscalibration and improves selective prediction.

4. We quantify correlation changes in selective AUC with stratified ECE AUC revealing that, substantial miscalibration remains across architectures, underscoring the need for more targeted approaches for safe multimodal clinical condition classification.

We find that high predictive performance in multimodal clinical condition classification often coincides with severe miscalibration, especially for rare conditions. In safety-critical settings, a model that looks accurate but fails to flag its own errors cannot be trusted. Although selective prediction could enable safety guardrails, current multimodal models systematically undermine this promise. Closing this gap demands models that couple strong discrimination with reliable selective prediction, making calibration-aware design essential for trustworthy clinical AI.

2. Related Work

Multimodal learning is increasingly used in clinical machine learning for integrating heterogeneous data sources such as structured EHR data, medical imaging, and free-text clinical reports (Warner et al., 2024; Simon et al., 2025). By combining complementary modalities, these models aim to emulate how clinicians synthesize diverse information for diagnosis and treatment planning (Ng et al., 2024; Lin et al., 2025).

Multimodal learning has been used in varying contexts: in some cases referring to multiple imaging views of the same modality (e.g., frontal and lateral chest X-rays) (Warner et al., 2024), in others describing fusion across fundamentally different data types such as imaging and tabular or textual inputs (Niu et al., 2024; Lee et al., 2024). While earlier work often focused on combining multiple imaging modalities for classification or segmentation (Sangeetha et al., 2024; Adebiyi et al., 2024), recent advances in large language models and vision–language models have broadened multimodal clinical systems to include combinations of radiology images, structured

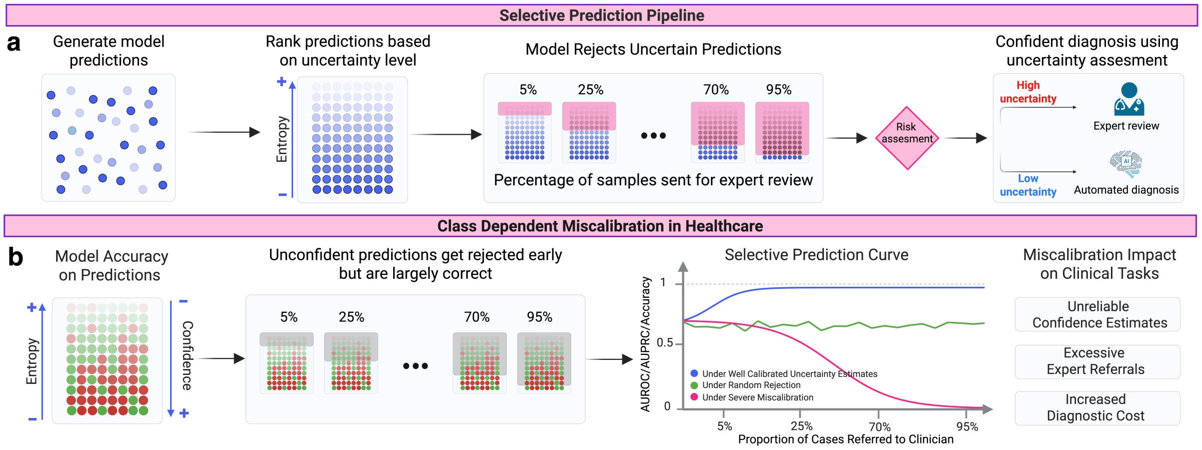
Selective Prediction Pipeline
a
Generate model predictions
Rank predictions based on uncertainty level
+
Entropy
−
Model Rejects Uncertain Predictions
5%
25%
70%
95%
Percentage of samples sent for expert review
Risk assesment
Confident diagnosis using uncertainty assesment
High uncertainty
Expert review
Low uncertainty
AI
Automated diagnosis
Class Dependent Miscalibration in Healthcare
b
Model Accuracy on Predictions
+
Entropy
−
−
Confidence
+
Unconfident predictions get rejected early but are largely correct
5%
25%
70%
95%
Selective Prediction Curve
AUROC/AUPRC/Accuracy
1
0.5
0
Under Well Calibrated Uncertainty Estimates
Under Random Rejection
Under Severe Miscalibration
5%
25%
70%
95%
Proportion of Cases Referred to Clinician
Miscalibration Impact on Clinical Tasks
Unreliable Confidence Estimates
Excessive Expert Referrals
Increased Diagnostic Cost

Figure 1: **Selective Prediction Can Serve as a Fail-Safe Mechanism in Safety-Critical Healthcare Settings.** **a)** Selective prediction pipeline: a model outputs prediction confidence/uncertainty scores; low-confidence cases are rejected and referred to an expert, yielding coverage-controlled evaluation and diagnosis. **b)** Calibration drives selective behavior: miscalibration causes predictions that are mostly correct to have high uncertainty (low confidence) and incorrect predictions to have high confidence. Theoretical selective prediction curves show the difference between a well-calibrated model, random selection, and a severely miscalibrated model, showing improving, flat, and degrading performance with increasing coverage. Clinically, overconfidence risks missed or delayed diagnoses, while underconfidence triggers unnecessary rejections and shifts excess workload to clinicians.

vitals, free-text notes, and even patient–provider conversations (Mahesh et al., 2024; Zambrano Chaves et al., 2025). This highlights the growing importance of multimodal integration in clinical AI, for tasks such as clinical question answering (Shoham and Rappoport, 2024), automated report generation (Chen et al., 2024), and medical documentation (Khan et al., 2025), underscoring the need to assess not only discrimination performance but also calibration and reliability when multiple heterogeneous sources are fused. Despite this progress, prior work in multimodal clinical ML has paid little attention to calibration, leaving open questions about whether improvements in predictive performance translate into reliable probability estimates for safe deployment (Zhao et al., 2024; Deng et al., 2025).

To investigate these challenges in a concrete and clinically meaningful setting, we focus on multimodal fusion of EHR time-series and chest X-ray (CXR) imaging, which is particularly relevant for critical care tasks such as diagnosis, mortality prediction, and intervention planning (Schilcher et al., 2024). Fusion architectures, such as MedFuse (Hayat et al., 2022), DrFuse (Yao et al., 2024) or Metra (Khader et al., 2023), have shown that even relatively simple strategies like early or late concatenation can yield performance gains on benchmarks such as MIMIC (Johnson et al., 2021). This paired EHR–CXR setting provides a practical and widely used benchmark for investigating how multimodal learning impacts selective prediction, calibration, and reliability in a clinical setting (Figure 1.b). In Appendix 3 we provide the theoretical details of our selective prediction pipeline and uncertainty quantification procedure.

3. Methods

Clinical Condition Classification. The objective of this multilabel classification task is to predict whether a patient presents any of 25 chronic, acute, or mixed clinical conditions during a given stay in the intensive care unit (ICU) (Hayat et al., 2022). Derived from one of the largest publicly-available multimodal datasets, this task has become a standard benchmark for multimodal predictive modeling in healthcare. The clinical conditions and their prevalence are detailed in Appendix 1.

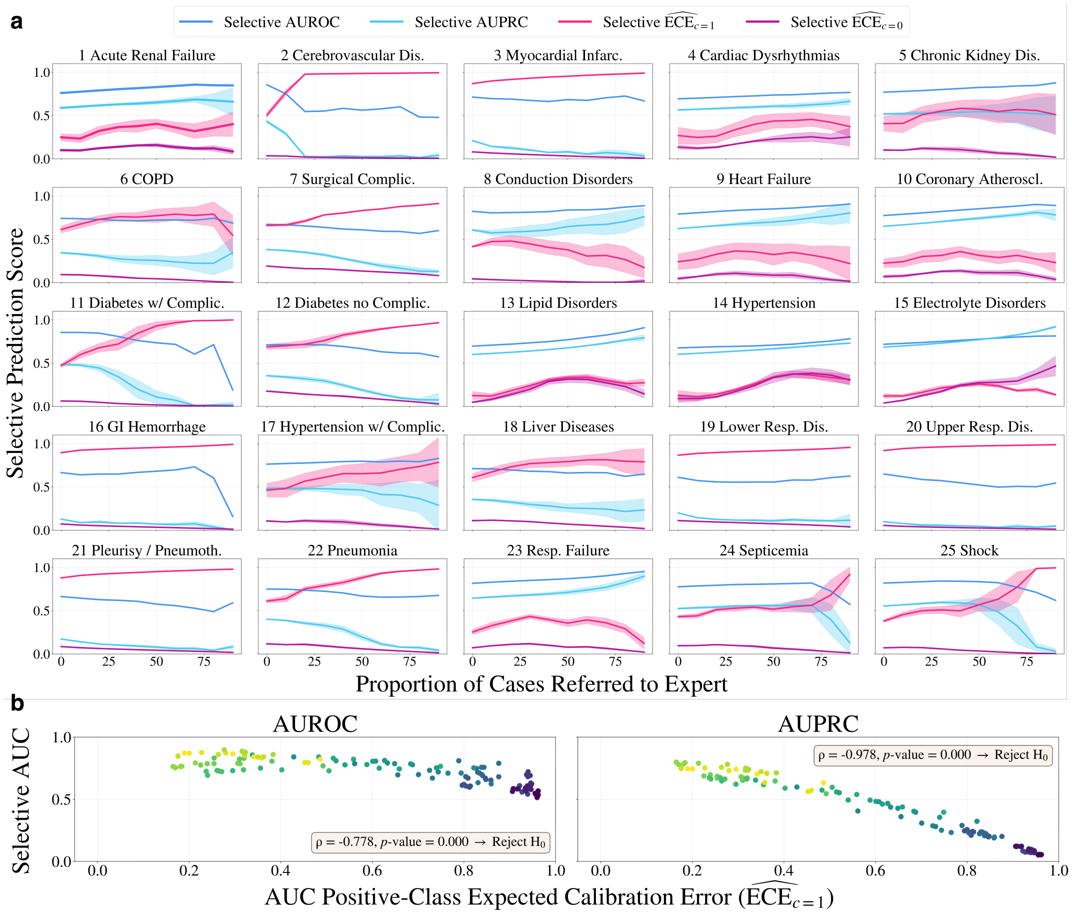

Figure 2: **How Does Calibration Shape Selective Prediction? a)** Selective AUROC, AUPRC, and class-stratified ECE across 25 conditions showing how performance evolves as low-confidence cases are rejected highlighting the imbalance between calibration and discrimination. Shaded regions show variance over five seeds. **b)** Higher positive-class ECE AUC consistently predicts lower selective AUC across all conditions, showing that overconfidence in positive predictions drives instability in selective evaluation. The null hypothesis of no non-negative correlation between ECE and Selective AUC across conditions, is rejected for both metrics (Spearman's rank H_0, $p < 0.05$).

Multimodal Dataset. We construct paired multimodal samples from MIMIC-IV (Johnson et al., 2021) (structured EHR time-series) and MIMIC-CXR (Johnson et al., 2019) (frontal-view chest X-rays), such that each patient sample contains both modalities: x^{ehr}, x^{cxr} and multilabel ground truth y_{ehr}. We follow standard preprocessing steps and do not allow overlap of patients across training and test splits, as in previous work (Hayat et al., 2022).

3.1. Model Variants

Architecture. We adopt the MedFuse architecture (Hayat et al., 2022) to train a multimodal classifier $f(\cdot)$ on paired EHR and CXR data $\mathcal{D} = \{(x_n^{ehr}, x_n^{cxr}, y_n)\}_{n=1}^{N}$. EHR time-series are encoded using a two-layer LSTM (Φ_{ehr}) (Hochreiter and Schmidhuber, 1997), and CXR images using a ResNet-34 (Φ_{cxr}) (He et al., 2015), with latent representations concatenated and passed to a classification head $g(\cdot)$ with sigmoid outputs. Models are trained with the standard binary cross-entropy loss, i.e. $\log p(y|\hat{y}) = -(y \log \hat{y} + (1-y)\log(1-\hat{y}))$.

Unimodal Baselines. We train unimodal models using only Φ_{ehr} (unimodal EHR) and Φ_{cxr} (unimodal CXR) to establish reference points for performance and calibration. The CXR encoder is initial-

Table 1: **Performance Across Conditions.** Aggregate metrics suggest that multimodal architectures outperform unimodal baselines. However, they mask substantial condition-specific variability, particularly in calibration. (**Best**, Second Best).

Model	**AUROC**	**AUPRC**	**Selective AUROC**	**Selective AUPRC**	$\widehat{\textbf{ECE}}\downarrow$
EHR (LSTM) (Hochreiter and Schmidhuber, 1997)	$0.681_{\pm 0.007}$	$0.371_{\pm 0.010}$	$0.625_{\pm 0.115}$	$0.292_{\pm 0.091}$	$\mathbf{1.42}_{\pm \mathbf{0.335}}$
CXR (ResNet-34) (He et al., 2015)	$0.654_{\pm 0.013}$	$0.350_{\pm 0.012}$	$0.643_{\pm 0.095}$	$0.342_{\pm 0.077}$	$9.41_{\pm 0.826}$
MedFuse (Hayat et al., 2022)	$\underline{0.739_{\pm 0.005}}$	$\underline{0.449_{\pm 0.008}}$	$\underline{0.705_{\pm 0.136}}$	$\underline{0.396_{\pm 0.112}}$	$\underline{2.08_{\pm 0.725}}$
MedFuse (Loss Upweight) (Hayat et al., 2022)	$\mathbf{0.742}_{\pm \mathbf{0.006}}$	$\mathbf{0.447}_{\pm \mathbf{0.010}}$	$\mathbf{0.764}_{\pm \mathbf{0.098}}$	$\mathbf{0.532}_{\pm \mathbf{0.093}}$	$4.88_{\pm 1.470}$
DrFuse (Yao et al., 2024)	$0.726_{\pm 0.007}$	$0.418_{\pm 0.010}$	$0.677_{\pm 0.139}$	$0.352_{\pm 0.116}$	$2.54_{\pm 1.15}$
MeTra (Khader et al., 2023)	$0.707_{\pm 0.010}$	$0.399_{\pm 0.015}$	$0.661_{\pm 0.125}$	$0.333_{\pm 0.104}$	$2.32_{\pm 0.83}$
Δ MedFuse vs EHR	+0.058 (8.5%)	+0.078 (21.0%)	+0.080 (12.8%)	+0.104 (35.6%)	+0.66 (46%)
Δ MedFuse (Loss Upweight) vs EHR	+0.061 (9.0%)	+0.076 (20.5%)	+0.139 (22.2%)	+0.240 (82.2%)	+3.46 (243.7%)

ized with ImageNet weights, and the EHR encoder is trained from scratch.

Multimodal Architectures. The original MedFuse fusion model serves as our deterministic multimodal baseline, allowing us to measure the calibration impact of modality fusion and providing a reference for uncertainty-aware variants. To assess the robustness of our findings to the choice of architecture, we additionally evaluate two alternative multimodal backbones: DrFuse (Yao et al., 2024), which learns EHR and CXR representations with divergence-based alignment, and MeTra (Khader et al., 2023), which encodes images and clinical variables using a transformer-based cross-modal fusion encoder.

4. How Does Calibration Shape Selective Prediction?

Calibration Drives Selective Performance. We next examine selective prediction thresholds for all 25 conditions using MedFuse, plotting selective AUROC and AUPRC alongside stratified ECE (Figure 2.a). Ideally, selective metrics should increase as uncertain predictions are correctly rejected, but we observe that many conditions deviate from this expectation. At extreme thresholds, selective AUROC and AUPRC collapse when the class distribution becomes particularly imbalanced. The mid-threshold regimes reveal more concerning behaviors where selective metrics stagnate or degrade, indicating that the model does not consistently reject the "right" cases.

Linking these failures to calibration reveals clear patterns. Conditions with low, balanced positive and negative ECE exhibit stable and improving selective AUROC/AUPRC (e.g., Labels 13 to 15). By contrast, high positive-class ECE produces noisy, unstable selective curves: AUROC hovers near chance and AUPRC declines sharply (e.g., Labels 2, 22, 24 and 25). Variance across random seeds follows the same trend, with poor calibration associated with wide, unstable spreads.

Calibration as a Predictor of Evaluation Stability. Across conditions (Figure 2.b), positive-class ECE AUC and Selective AUC are consistently negatively correlated: higher positive-class ECE predicts smaller or even negative changes in selective prediction metrics. The effect is strongest for AUPRC, where degradation is nearly monotonic, while AUROC shows greater variability but follows the same trend. We fomally quantify this with the null hypothesis of no non-negative association between ECE and Selective AUC across conditions, which is rejected for both metrics (Spearman's rank correlation, $p < 0.05$).

Overall, these findings establish positive-class calibration as a reliable predictor of selective prediction stability. High positive-class ECE serves as an early warning that selective prediction will fail to yield meaningful improvements over standard evaluation. Since this holds across all conditions, it reflects a generalizable phenomenon rather than a label-specific artifact.

> **Takeaway.** Positive-class calibration is a leading indicator of selective reliability: when $\text{ECE}_{c=1}$ is high, selective AUROC/AUPRC fail to improve and can deteriorate.

Extended Results. In Appendix 4 we complement this analysis with full quantitative comparisons on individual modalities (CXR, EHR) and multimodal architectures (MedFuse, DrFuse and MeTra) across different evaluation settings.

5. Conclusion

Our empirical findings highlight a central challenge: *multimodal fusion improves discrimination but undermines calibration, making selective prediction unreliable.* Addressing this gap requires moving beyond aggregate performance and developing calibration-aware selective prediction methods that can provide trustworthy fail-safe mechanisms. Without this, the deployment of multimodal clinical AI risks offering the illusion of safety while leaving critical vulnerabilities unaddressed.

References

Abdulmateen Adebiyi, Nader Abdalnabi, Emily Hoffman Smith, Jesse Hirner, Eduardo J Simoes, Mirna Becevic, and Praveen Rao. Accurate skin lesion classification using multimodal learning on the ham10000 dataset. *MedRxiv*, pages 2024–05, 2024.

Shashank Agarwal. Machine learning based personalized treatment plans for chronic conditions. In *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, pages 1127–1132. IEEE, 2024.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

Jiawen Deng, Mohamed E Elghobashy, Kathleen Zang, Shubh K Patel, Eddie Guo, and Kiyan Heybati. So you've got a high auc, now what? an overview of important considerations when bringing machine-learning models from computer to bedside. *Medical Decision Making*, page 0272989X251343082, 2025.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, pages 1184–1193. PMLR, 2018.

Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Matthew G Hanna, Liron Pantanowitz, Rajesh Dash, James H Harrison, Mustafa Deebajah, Joshua Pantanowitz, and Hooman H Rashidi. Future of artificial intelligence (ai)-machine learning (ml) trends in pathology and medicine. *Modern Pathology*, page 100705, 2025.

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*, pages 479–503. PMLR, August 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Haseeb Javed, Shaker El-Sappagh, and Tamer Abuhmed. Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust ai applications. *Artificial Intelligence Review*, 58(1):12, 2024.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and S Horng IV. Mimic-iv-ed. *PhysioNet*, 2021.

Alistair EW 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, 2019.

Firas Khader, Jakob Nikolas Kather, Gustav Müller-Franzes, Tianci Wang, Tianyu Han, Soroosh Tayebi Arasteh, Karim Hamesch, Keno Bressem, Christoph Haarburger, Johannes Stegmaier, et al. Medical transformer for multimodal survival prediction in intensive care: integration of imaging and non-imaging data. *Scientific Reports*, 13(1): 10666, 2023.

Wasif Khan, Seowung Leem, Kyle B See, Joshua K Wong, Shaoting Zhang, and Ruogu Fang. A comprehensive survey of foundation models in medicine. *IEEE Reviews in Biomedical Engineering*, 2025.

Chih-Kuo Lee, Ting-Li Chen, Jeng-En Wu, Min-Tsun Liao, Chiehhung Wang, Weichung Wang, and Cheng-Ying Chou. Multimodal deep learning models utilizing chest x-ray and electronic health record data for predictive screening of acute heart failure in emergency department. *Computer Methods and Programs in Biomedicine*, 255:108357, 2024.

Qika Lin, Yifan Zhu, Xin Mei, Ling Huang, Jingying Ma, Kai He, Zhen Peng, Erik Cambria, and Mengling Feng. Has multimodal learning delivered universal intelligence in healthcare? a comprehensive survey. *Information Fusion*, 116:102795, 2025.

Nandhini Mahesh, Chitralekha S Devishamani, Keerthana Raghu, Maanasi Mahalingam, Pragathi Bysani, Arjun V Chakravarthy, and Rajiv Raman. Advancing healthcare: the role and impact of ai and foundation models. *American Journal of Translational Research*, 16(6):2166, 2024.

Shachi Mall, Ashutosh Srivastava, Bireshwar Dass Mazumdar, Manmohan Mishra, Sunil L Bangare, and A Deepak. Implementation of machine learning techniques for disease diagnosis. *Materials Today: Proceedings*, 51:2198–2201, 2022.

Dane A Morey, Michael F Rayo, and David D Woods. Empirically derived evaluation requirements for responsible deployments of ai in safety-critical settings. *npj Digital Medicine*, 8(1):374, 2025.

Jeremy Y Ng, Holger Cramer, Myeong Soo Lee, and David Moher. Traditional, complementary, and integrative medicine and artificial intelligence: Novel opportunities in healthcare. *Integrative medicine research*, 13(1):101024, 2024.

Shuai Niu, Jing Ma, Liang Bai, Zhihua Wang, Li Guo, and Xian Yang. Ehr-knowgen: Knowledge-enhanced multimodal learning for disease diagnosis generation. *Information Fusion*, 102:102069, 2024.

Jibon Kumar Paul, Mahir Azmal, Omar Faruk Talukder, ANM Shah Newaz Been Haque, Meghla Meem, and Ajit Ghosh. Harnessing machine learning for improved diagnosis, drug discovery, and patient care. *Computational and Structural Biotechnology Reports*, page 100051, 2025.

Stephan Rabanser and Nicolas Papernot. What does it take to build a performant selective classifier? *arXiv preprint arXiv:2510.20242*, 2025.

Stephan Rabanser, Anvith Thudi, Kimia Hamidieh, Adam Dziedzic, and Nicolas Papernot. Selective classification via neural network training dynamics. *arXiv preprint arXiv:2205.13532*, 2022.

Tim G. J. Rudner, Sanyam Kapoor, Shikai Qiu, and Andrew Gordon Wilson. Function-space regularization in neural networks: A probabilistic perspective. In *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2023a.

Tim G. J. Rudner, Sanyam Kapoor, Shikai Qiu, and Andrew Gordon Wilson. Should we learn most likely functions or parameters? In *Advances in Neural Information Processing Systems 36*, 2023b.

SKB Sangeetha, Sandeep Kumar Mathivanan, P Karthikeyan, Hariharan Rajadurai, Basu Dev Shivahare, Saurav Mallik, and Hong Qin. An enhanced multimodal fusion deep learning neural network for lung cancer classification. *Systems and Soft Computing*, 6:200068, 2024.

Jörg Schilcher, Alva Nilsson, Oliver Andlid, and Anders Eklund. Fusion of electronic health records and radiographic images for a multimodal deep learning prediction model of atypical femur fractures. *Computers in Biology and Medicine*, 168: 107704, 2024.

Ofir Ben Shoham and Nadav Rappoport. Medconceptsqa: Open source medical concepts qa benchmark. *Computers in Biology and Medicine*, 182: 109089, 2024.

Benjamin D Simon, Kutsev Bengisu Ozyoruk, David G Gelikman, Stephanie A Harmon, and Barış Türkbey. The future of multimodal artificial intelligence models for integrating imaging and clinical metadata: a narrative review. *Diagnostic and Interventional Radiology*, 31(4):303, 2025.

Elisa Warner, Joonsang Lee, William Hsu, Tanveer Syeda-Mahmood, Charles E Kahn Jr, Olivier Gevaert, and Arvind Rao. Multimodal machine learning in image-based and clinical biomedicine: Survey and prospects. *International journal of computer vision*, 132(9):3753–3769, 2024.

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, pages 16416–16424, 2024.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng

Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16 (1):3108, 2025.

Tianyi Zhao, Liangliang Zhang, Yao Ma, and Lu Cheng. A survey on safe multi-modal learning systems. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6655–6665, 2024.

Appendix

Table of Contents

Reproducibility

We have made a significant effort to ensure the reproducibility of our results. An anonymized implementation of our method is provided at anonymous.4open.science/r/medcalibration-B187, which includes training, evaluation, and analysis scripts. The experimental setup—including hyperparameters, model configurations, and sampling parameters are described in Appendix 2. All datasets used in our experiments are publicly available, and we additionally provide scripts for data preparation. We provide model checkpoints for all models used in this paper with instructions for reproducing results in the `README.md` file in the released code.

1. Clinical Condition Dataset Distribution

Table 2: **Clinical Conditions and Dataset Characteristics.** Overview of the 25 conditions analyzed in this study, including their associated organ/system, primary diagnostic modality (CXR or EHR), and the prevalence of positive and negative samples in the training set.

		Modality		Train Set Prevalence	
Clinical Condition	**Organ**	**CXR**	**EHR**	**Positive**	**Negative**
1 Acute and unspecified renal failure	Kidneys	-	✓	32.32% (2,532)	67.68% (5,309)
2 Acute cerebrovascular disease	Brain (Cerebrovascular system)	-	✓	7.99% (625)	92.01% (7,216)
3 Acute myocardial infarction	Heart (Coronary arteries)	-	✓	8.57% (675)	91.43% (7,166)
4 Cardiac dysrhythmias	Heart (Conduction system)	-	✓	37.36% (2,942)	62.64% (4,899)
5 Chronic kidney disease	Kidneys	✓	✓	23.34% (1,824)	76.66% (6,017)
6 Chronic obstructive pulmonary disease	Lungs (Airways)	✓	✓	16.40% (1,269)	83.60% (6,572)
7 Complications of surgical procedures	Multiple systems (context-specific)	-	✓	21.75% (1,703)	78.25% (6,138)
8 Conduction disorders	Heart (Electrical system)	✓	✓	10.79% (838)	89.21% (7,003)
9 Congestive heart failure; nonhypertensive	Heart	-	✓	31.03% (2,451)	68.97% (5,390)
10 Coronary atherosclerosis and other heart disease	Heart (Coronary arteries)	-	✓	31.00% (2,452)	69.00% (5,389)
11 Diabetes mellitus with complications	Endocrine system (Pancreas) + Target organs	-	✓	11.82% (909)	88.18% (6,932)
12 Diabetes mellitus without complication	Endocrine system (Pancreas)	-	✓	20.58% (1,634)	79.42% (6,207)
13 Disorders of lipid metabolism	Liver and Circulatory system	-	✓	41.05% (3,207)	58.95% (4,634)
14 Essential hypertension	Cardiovascular system	-	✓	44.14% (3,467)	55.86% (4,374)
15 Fluid and electrolyte disorders	Kidneys and Endocrine system	-	✓	45.48% (3,547)	54.52% (4,294)
16 Gastrointestinal hemorrhage	GI Tract	-	✓	7.08% (562)	92.92% (7,279)
17 Hypertension with complications	Cardiovascular system	-	✓	21.17% (1,649)	78.83% (6,192)
18 Other liver diseases	Liver	✓	✓	16.12% (1,248)	83.88% (6,593)
19 Other lower respiratory disease	Lungs (Lower airways)	✓	✓	12.93% (1,001)	87.07% (6,840)
20 Other upper respiratory disease	Lungs/Nasal passages (Upper airways)	✓	✓	6.29% (500)	93.71% (7,341)
21 Pleurisy; pneumothorax	Lungs/Pleura	✓	✓	9.93% (784)	90.07% (7,057)
22 Pneumonia (except caused by tuberculosis or std)	Lungs (Alveoli)	✓	✓	18.84% (1,489)	81.16% (6,352)
23 Respiratory failure	Lungs and/or Neuromuscular system	-	✓	28.37% (2,229)	71.63% (5,612)
24 Septicemia (except in labor)	Bloodstream (Systemic)	✓	✓	22.21% (1,732)	77.79% (6,109)
25 Shock	Cardiovascular system (Systemic)	-	✓	18.01% (1,413)	81.99% (6,428)

2. Training Details

Hyperparameter Optimization and Model Selection

We conducted 50 experiments corresponding to a randomly sampled hyperparameter set. Every set was evaluated across five runs with different random seeds. The learning rate was sampled uniformly from $[10^{-5}, 10^{-2}]$, and the number of training epochs from $\{5, 10, 15, 20, 30\}$. We used a fixed batch size of 16 and a cosine decay scheduler with $\alpha = 0$. Model checkpoints were saved at the final training epoch, and the optimal hyperparameter set was selected based on the highest mean validation AUROC across seeds, yielding a total of 250 runs per task.

Using the selected configuration, we retrained the models by combining the training and validation splits and ran five seeds with different initializations to obtain the final models for test evaluation. We report the performance results on the test set with means and standard errors computed over five random seeds. This procedure was applied to unimodal CXR, EHR, MedFuse, and MedFuse + GAP models following the original MedFuse protocol to identify the best-performing configuration based on AUROC.

All models, were trained using the Adam optimizer. Our experiments were executed using NVIDIA A100 and V100 80Gb Tensor Core GPUs.

Data Availability

The MIMIC (Medical Information Mart for Intensive Care) dataset is publicly available for research purposes. Access to the data requires completion of a data use agreement, which ensures compliance with the Health Insurance Portability and Accountability Act (HIPAA). Researchers can request access through the PhysioNet platform at https://physionet.org, where detailed instructions and requirements are provided. In our study, we used versions MIMIC-IV and MIMIC-CXR which include de-identified health data, vital signs, laboratory test results, medication records and chest X-ray images from patients admitted to the intensive care units at the Beth Israel Deaconess Medical Center.

3. Preliminaries

Calibration. A core requirement for selective prediction is that model confidence scores are well-calibrated, (i.e., predicted probabilities reflect true likelihoods of correctness), which is commonly measured using the Expected Calibration Error (ECE) (Guo et al., 2017). Formally, we can measure miscalibration as the expected gap between predicted confidence and empirical accuracy:

$$\mathrm{ECE} = \mathbb{E}_{\hat{P}}\big[\,\big|\mathbb{P}(\hat{Y} = Y \mid \hat{P} = p) - p\big|\,\big]. \tag{3.1}$$

To approximate ECE in practice, we can use a finite partition of the probability space into M bins $\{B_m\}_{m=1}^{M}$:

$$\widehat{\mathrm{ECE}} = \sum_{m=1}^{M} \frac{|B_m|}{n} \big|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\big|, \tag{3.2}$$

where $\mathrm{acc}(B_m)$ and $\mathrm{conf}(B_m)$ denote the average accuracy and confidence of samples in bin B_m, respectively, and n is the number of samples. Additionally, we may wish to compute the conditional calibration of a predictor. To do so, we can define a conditional expected calibration error,

$$\mathrm{ECE}_c = \mathbb{E}_{\hat{P}|\hat{Y}=c}\big[\,\big|\mathbb{P}(\hat{Y} = c \mid \hat{P} = p) - p\big|\,\big]. \tag{3.3}$$

which captures class-dependent calibration. We can approximate the conditional calibration error by only considering the subset of points with label c and, again, specifying a finite partition of the probability space into M bins $\{B_m^c\}_{m=1}^{M}$:

$$\widehat{\mathrm{ECE}}_c = \sum_{m=1}^{M} \frac{|B_m^c|}{n_c} \big|\mathrm{acc}(B_m^c) - \mathrm{conf}(B_m^c)\big|, \tag{3.4}$$

where $\text{acc}(B_m^c)$ and $\text{conf}(B_m^c)$ denote the average accuracy and confidence of samples in bin B_m^c, respectively, and n_c is the number of samples whose label is c.

Selective Prediction. Selective prediction modifies the standard prediction pipeline by introducing a "reject option" denoted by $\perp$, using a gating mechanism defined by a selection function $s : \mathcal{X} \to \mathbb{R}$ that determines whether a prediction should be made for a given input point $\mathbf{x} \in \mathcal{X}$ (El-Yaniv et al., 2010). For a rejection threshold τ, with s representing the entropy of $\mathbf{x}$, the prediction model is given by:

$$(p(\mathbf{y} \,|\, \cdot; f), s)(\mathbf{x}) = \begin{cases} p(\mathbf{y} \,|\, \mathbf{x}; f), & \text{if } s \leq \tau \\ \perp, & \text{otherwise,} \end{cases} \tag{3.5}$$

with $p(\mathbf{y} \mid \cdot; f)$ being a model's predictive distribution. A variety of methods have been proposed to find a selection function s (Rabanser et al., 2022; Rabanser and Papernot, 2025), with model's predictive uncertainty being a common choice: $s = \mathcal{H}(p(\mathbf{y} \mid \mathbf{x}; f))$. That is, a point $\mathbf{x} \in \mathcal{X}$ will be placed into the rejection class if the model's predictive uncertainty is above a certain threshold.

To evaluate the predictive performance of a prediction model $(p(\mathbf{y} \mid \cdot; f), s)(\mathbf{x})$, we compute the predictive performance of the classifier $p(\mathbf{y} \mid \mathbf{x}; f)$ over a range of thresholds τ (Figure 1.a). Performance is then assessed with standard metrics (e.g., accuracy, AUROC, AUPRC, or ECE), yielding selective prediction curves that capture the trade-off between coverage—the proportion of cases for which predictions are made—and performance (Geifman and El-Yaniv, 2017).

Predictive Uncertainty Quantification. Furthermore, understanding calibration requires examining how models quantify predictive uncertainty, which is typically decomposed into aleatoric uncertainty, arising from inherent noise in the data, and epistemic uncertainty, arising from limited knowledge about the model or parameters (Depeweg et al., 2018). A common way to quantify predictive uncertainty is via the Shannon entropy of the predictive distribution:

$$\mathcal{H}(p(\mathbf{y} \mid \mathbf{x})) = -\sum_{c} p(\mathbf{y} = c \mid \mathbf{x}) \log p(\mathbf{y} = c \mid \mathbf{x}), \tag{3.6}$$

where higher entropy indicates greater uncertainty (Rudner et al., 2023a,b).

Table 3: **Condition-level Performance of Standard and Selective Prediction Metrics**. MedFuse generally improves AUROC, AUPRC and selective AUROC, but calibration gains are inconsistent compared to EHR. (Dark-bold: $p < 0.05$, Wilcoxon signed-rank test, 5 seeds; Light-bold: highest mean, not significant)

Clinical Condition			EHR					CXR					MedFuse		
	AUROC	AUPRC	Selective AUROC	Selective AUPRC	$\widehat{\text{ECE}}$ ↓	AUROC	AUPRC	Selective AUROC	Selective AUPRC	$\widehat{\text{ECE}}$ ↓	AUROC	AUPRC	Selective AUROC	Selective AUPRC	$\widehat{\text{ECE}}$ ↓
1	0.724±0.003	0.564±0.010	0.772±0.100	0.627±0.100	**1.62**±0.52	0.639±0.009	0.453±0.014	0.684±0.081	0.503±0.075	12.77±0.98	**0.761**±0.005	**0.589**±0.008	**0.804**±0.111	**0.634**±0.109	2.31±0.97
2	**0.873**±0.005	**0.459**±0.010	0.569±0.180	**0.085**±0.138	1.08±0.14	0.650±0.009	0.139±0.009	**0.577**±0.121	0.084±0.057	3.99±0.31	0.856±0.010	0.432±0.019	0.560±0.176	0.075±0.116	**0.88**±0.24
3	0.693±0.014	0.175±0.016	0.617±0.150	0.089±0.070	**0.73**±0.07	0.640±0.015	0.153±0.017	0.567±0.102	**0.092**±0.037	4.35±0.42	**0.713**±0.005	**0.209**±0.010	**0.633**±0.207	0.083±0.049	0.76±0.23
4	0.598±0.004	0.462±0.012	0.618±0.069	0.491±0.065	**1.15**±0.67	0.647±0.009	0.534±0.013	0.683±0.077	0.598±0.077	12.50±0.48	**0.694**±0.006	**0.565**±0.010	**0.727**±0.079	**0.610**±0.071	2.61±1.65
5	0.696±0.006	0.402±0.012	0.642±0.113	0.244±0.113	**2.13**±0.45	0.662±0.021	0.378±0.019	0.696±0.085	0.409±0.078	15.37±0.37	**0.770**±0.003	**0.519**±0.010	**0.787**±0.177	**0.521**±0.169	2.34±2.04
6	0.612±0.005	0.203±0.008	0.564±0.089	0.140±0.045	**1.87**±0.23	0.702±0.012	0.301±0.016	0.676±0.130	0.255±0.139	7.15±1.26	**0.742**±0.004	**0.345**±0.011	**0.717**±0.120	**0.279**±0.127	2.17±0.45
7	0.649±0.007	0.365±0.013	0.578±0.098	0.221±0.083	**1.16**±0.45	0.597±0.015	0.308±0.009	0.555±0.069	**0.263**±0.067	9.15±0.58	**0.673**±0.001	**0.383**±0.003	**0.602**±0.094	0.241±0.100	1.46±0.31
8	0.627±0.010	0.174±0.007	0.558±0.098	0.109±0.054	**1.09**±0.19	0.814±0.010	0.608±0.013	0.784±0.130	0.597±0.198	3.42±0.45	**0.824**±0.004	**0.609**±0.009	**0.826**±0.099	**0.648**±0.119	2.32±0.45
9	0.686±0.005	0.463±0.011	0.690±0.113	0.419±0.128	**1.81**±0.56	0.755±0.008	0.565±0.017	0.808±0.097	0.658±0.100	11.92±1.95	**0.792**±0.004	**0.625**±0.008	**0.842**±0.110	**0.703**±0.123	4.50±1.71
10	0.697±0.004	0.517±0.008	0.658±0.096	0.412±0.144	**1.80**±0.43	0.728±0.013	0.583±0.019	0.780±0.094	0.655±0.093	11.15±1.04	**0.776**±0.002	**0.651**±0.004	**0.835**±0.109	**0.723**±0.115	3.52±1.29
11	0.755±0.015	0.304±0.014	0.623±0.181	0.100±0.070	1.41±0.26	0.608±0.019	0.171±0.013	0.583±0.104	0.133±0.064	5.42±0.83	**0.855**±0.006	**0.485**±0.012	**0.676**±0.269	**0.198**±0.193	**1.15**±0.22
12	0.623±0.012	0.286±0.009	0.556±0.085	0.195±0.061	**1.38**±0.28	0.589±0.022	0.257±0.013	0.590±0.088	**0.214**±0.049	8.45±1.35	**0.711**±0.006	**0.355**±0.010	**0.650**±0.121	0.200±0.110	1.74±0.56
13	0.650±0.003	0.553±0.010	0.642±0.085	0.620±0.070	**0.99**±0.42	0.638±0.003	0.519±0.010	0.670±0.088	0.561±0.075	15.71±0.34	**0.696**±0.002	**0.599**±0.004	**0.766**±0.139	**0.668**±0.125	3.61±0.64
14	0.591±0.011	0.529±0.012	0.572±0.067	0.594±0.068	2.77±0.48	0.611±0.011	0.529±0.009	0.637±0.075	0.565±0.061	18.90±1.04	**0.676**±0.005	**0.603**±0.006	**0.714**±0.087	**0.660**±0.085	**1.86**±0.62
15	0.703±0.005	0.676±0.005	0.715±0.103	0.782±0.083	**1.41**±0.42	0.600±0.011	0.544±0.009	0.626±0.071	0.586±0.062	16.93±1.17	**0.718**±0.003	**0.685**±0.004	**0.755**±0.114	**0.789**±0.082	3.39±0.92
16	0.628±0.007	0.111±0.008	0.590±0.148	0.073±0.033	**0.58**±0.14	0.597±0.021	0.109±0.010	0.558±0.113	**0.077**±0.028	4.36±0.27	**0.665**±0.009	**0.128**±0.003	**0.606**±0.174	0.072±0.043	1.05±0.36
17	0.690±0.005	0.377±0.012	0.618±0.108	0.204±0.093	**1.91**±0.27	0.666±0.017	0.343±0.017	0.674±0.074	0.324±0.056	14.12±0.48	**0.763**±0.004	**0.482**±0.014	**0.746**±0.188	**0.413**±0.175	2.70±1.70
18	0.643±0.008	0.285±0.005	0.580±0.083	0.158±0.075	**0.98**±0.19	0.672±0.013	0.345±0.003	**0.674**±0.081	**0.387**±0.112	6.42±0.82	**0.712**±0.003	**0.356**±0.009	0.658±0.101	0.273±0.097	1.75±0.19
19	0.594±0.008	0.172±0.008	0.543±0.080	0.120±0.032	**0.55**±0.15	0.547±0.022	0.154±0.007	0.520±0.075	**0.150**±0.065	5.55±1.19	**0.611**±0.005	**0.199**±0.008	**0.568**±0.095	0.119±0.036	1.50±0.29
20	0.623±0.010	0.091±0.007	0.531±0.113	0.056±0.031	0.98±0.15	0.615±0.019	**0.103**±0.009	0.523±0.091	**0.060**±0.029	3.03±0.34	**0.649**±0.011	0.098±0.003	**0.532**±0.109	0.054±0.027	**0.68**±0.10
21	0.591±0.009	0.129±0.006	0.573±0.129	0.097±0.037	1.91±0.38	0.658±0.023	**0.179**±0.013	0.569±0.102	**0.104**±0.048	5.11±0.98	**0.663**±0.014	0.172±0.006	**0.574**±0.103	0.089±0.056	**1.43**±0.48
22	0.722±0.007	0.368±0.007	0.607±0.119	0.184±0.114	**1.10**±0.27	0.665±0.009	0.288±0.008	0.624±0.118	0.203±0.070	8.30±1.29	**0.750**±0.004	**0.402**±0.005	**0.669**±0.154	**0.214**±0.140	2.04±0.83
23	0.803±0.009	0.620±0.013	0.804±0.146	0.607±0.177	**1.57**±0.36	0.714±0.008	0.493±0.017	0.725±0.102	0.475±0.095	11.41±1.37	**0.817**±0.005	**0.645**±0.007	**0.872**±0.101	**0.738**±0.115	2.09±0.85
24	0.754±0.006	0.483±0.010	0.691±0.141	0.358±0.177	**1.88**±0.40	0.649±0.005	0.350±0.008	0.626±0.099	0.302±0.097	11.07±1.03	**0.776**±0.004	**0.525**±0.010	**0.752**±0.151	**0.475**±0.178	2.38±0.68
25	0.808±0.006	0.519±0.023	0.704±0.179	0.316±0.214	**1.55**±0.48	0.693±0.008	0.336±0.006	0.659±0.110	0.301±0.106	8.71±0.31	**0.820**±0.003	**0.554**±0.005	**0.748**±0.203	**0.413**±0.242	1.65±0.35
Average	0.681±0.007	0.371±0.010	0.625±0.115	0.292±0.091	**1.42**±0.34	0.654±0.013	0.350±0.012	0.643±0.095	0.342±0.077	9.41±0.83	**0.739**±0.005	**0.449**±0.008	**0.705**±0.136	**0.396**±0.112	2.08±0.73

4. Further Empirical Results

In this section, we present complimentary findings of our study on selective prediction and calibration in multimodal clinical condition classification guided by the following main questions:

1. *Do multimodal models improve or degrade reliable selective prediction?*
2. *Do other multimodal architectures exhibit the same reliability patterns?*
3. *Can loss-upweight training mitigate calibration failures?*

To this end, we compare unimodal and multimodal models on standard discrimination and calibration metrics to establish whether multimodal fusion improves reliability. Additionally, we decompose calibration into positive- and negative-class components to identify which types of predictions drive miscalibration. We also contextualize these calibration properties with selective prediction behavior across coverage thresholds, showing when and why selective performance degrades. This analysis allows us to assess both aggregate and fine-grained metrics, linking calibration quality with selective reliability. We conclude by evaluating a simple label-upweighted loss training strategy, which provides improvements in selective prediction but highlights the remaining challenges in mitigating class-dependent miscalibration.

4.1. *Do Multimodal Models Improve or Degrade Reliable Selective Prediction?*

Overall Performance. Table 3 presents the performance metrics for the unimodal models and the deterministic multimodal baseline (MedFuse). Across the 25 conditions, MedFuse generally achieves higher AUROC, AUPRC, and selective AUROC than both unimodal baselines, often with statistical significance. However, its calibration gains over EHR are less consistent, illustrating a recurrent mismatch between discrimination and probability alignment.

For example, in *Acute and unspecified renal failure (1)*, MedFuse achieves the highest AUROC (0.761) and AUPRC (0.589), reducing ECE relative to CXR (12.77 vs. 2.31), yet EHR remains the best calibrated

Table 4: **Condition-level Performance of ECE stratification.** Positive-class ECE dominates across conditions, highlighting overconfidence in positive predictions. (Dark-bold: $p < 0.05$, Wilcoxon signed-rank test, 5 seeds; Light-bold: highest mean, not significant)

Clinical Condition	EHR			CXR			MedFuse		
	$\widehat{\mathbf{ECE}}\downarrow$	$\widehat{\mathbf{ECE}}_{c=1}\downarrow$	$\widehat{\mathbf{ECE}}_{c=0}\downarrow$	$\widehat{\mathbf{ECE}}\downarrow$	$\widehat{\mathbf{ECE}}_{c=1}\downarrow$	$\widehat{\mathbf{ECE}}_{c=0}\downarrow$	$\widehat{\mathbf{ECE}}\downarrow$	$\widehat{\mathbf{ECE}}_{c=1}\downarrow$	$\widehat{\mathbf{ECE}}_{c=0}\downarrow$
1	$\mathbf{1.62}_{\pm\mathbf{0.52}}$	$30.76_{\pm 1.82}$	$14.32_{\pm 1.14}$	$12.77_{\pm 0.98}$	$35.26_{\pm 5.07}$	$\mathbf{3.63}_{\pm\mathbf{1.74}}$	$2.31_{\pm 0.97}$	$\mathbf{24.85}_{\pm\mathbf{3.96}}$	$9.97_{\pm 1.14}$
2	$1.08_{\pm 0.14}$	$\mathbf{47.76}_{\pm\mathbf{3.00}}$	$4.68_{\pm 0.25}$	$3.99_{\pm 0.31}$	$83.21_{\pm 3.28}$	$4.13_{\pm 0.75}$	$\mathbf{0.88}_{\pm\mathbf{0.24}}$	$50.56_{\pm 3.07}$	$\mathbf{3.56}_{\pm\mathbf{0.19}}$
3	$\mathbf{0.73}_{\pm\mathbf{0.07}}$	$88.27_{\pm 0.33}$	$8.22_{\pm 0.06}$	$4.35_{\pm 0.42}$	$\mathbf{80.59}_{\pm\mathbf{5.85}}$	$\mathbf{4.91}_{\pm\mathbf{0.70}}$	$0.76_{\pm 0.23}$	$86.93_{\pm 0.14}$	$8.05_{\pm 0.25}$
4	$\mathbf{1.15}_{\pm\mathbf{0.67}}$	$49.64_{\pm 1.14}$	$28.93_{\pm 0.65}$	$12.50_{\pm 0.48}$	$29.02_{\pm 5.95}$	$\mathbf{4.34}_{\pm\mathbf{3.01}}$	$2.61_{\pm 1.65}$	$\mathbf{26.90}_{\pm\mathbf{9.52}}$	$13.57_{\pm 2.63}$
5	$\mathbf{2.13}_{\pm\mathbf{0.45}}$	$67.12_{\pm 0.76}$	$19.71_{\pm 0.41}$	$15.37_{\pm 0.37}$	$45.19_{\pm 2.51}$	$\mathbf{5.92}_{\pm\mathbf{0.92}}$	$2.34_{\pm 2.04}$	$\mathbf{40.71}_{\pm\mathbf{8.56}}$	$10.21_{\pm 1.06}$
6	$\mathbf{1.87}_{\pm\mathbf{0.23}}$	$81.50_{\pm 0.22}$	$16.00_{\pm 0.07}$	$7.15_{\pm 1.26}$	$64.69_{\pm 14.12}$	$\mathbf{3.32}_{\pm\mathbf{1.04}}$	$2.17_{\pm 0.45}$	$\mathbf{61.34}_{\pm\mathbf{3.97}}$	$9.27_{\pm 0.68}$
7	$\mathbf{1.16}_{\pm\mathbf{0.45}}$	$70.71_{\pm 1.80}$	$20.18_{\pm 0.20}$	$9.15_{\pm 0.58}$	$\mathbf{66.07}_{\pm\mathbf{3.97}}$	$\mathbf{7.68}_{\pm\mathbf{1.09}}$	$1.46_{\pm 0.31}$	$66.13_{\pm 1.23}$	$19.12_{\pm 0.40}$
8	$\mathbf{1.09}_{\pm\mathbf{0.19}}$	$87.20_{\pm 0.28}$	$10.42_{\pm 0.04}$	$3.42_{\pm 0.45}$	$\mathbf{38.99}_{\pm\mathbf{6.15}}$	$\mathbf{2.77}_{\pm\mathbf{0.79}}$	$2.32_{\pm 0.45}$	$41.64_{\pm 0.88}$	$4.38_{\pm 0.89}$
9	$\mathbf{1.81}_{\pm\mathbf{0.56}}$	$41.15_{\pm 1.10}$	$17.36_{\pm 0.89}$	$11.92_{\pm 1.95}$	$\mathbf{20.38}_{\pm\mathbf{5.59}}$	$8.40_{\pm 4.88}$	$4.50_{\pm 1.71}$	$24.13_{\pm 9.07}$	$\mathbf{4.69}_{\pm\mathbf{1.96}}$
10	$\mathbf{1.80}_{\pm\mathbf{0.43}}$	$38.09_{\pm 0.54}$	$17.50_{\pm 0.76}$	$11.15_{\pm 1.04}$	$30.34_{\pm 5.39}$	$\mathbf{3.44}_{\pm\mathbf{2.02}}$	$3.52_{\pm 1.29}$	$\mathbf{22.44}_{\pm\mathbf{4.96}}$	$7.08_{\pm 1.88}$
11	$1.41_{\pm 0.26}$	$80.37_{\pm 1.11}$	$10.92_{\pm 0.15}$	$5.42_{\pm 0.83}$	$74.92_{\pm 6.19}$	$\mathbf{5.25}_{\pm\mathbf{0.72}}$	$\mathbf{1.15}_{\pm\mathbf{0.22}}$	$\mathbf{47.13}_{\pm\mathbf{2.20}}$	$6.30_{\pm 0.34}$
12	$\mathbf{1.38}_{\pm\mathbf{0.28}}$	$77.14_{\pm 0.38}$	$19.78_{\pm 0.07}$	$8.45_{\pm 1.35}$	$\mathbf{62.23}_{\pm\mathbf{7.64}}$	$\mathbf{6.34}_{\pm\mathbf{2.18}}$	$1.74_{\pm 0.56}$	$68.84_{\pm 2.85}$	$17.64_{\pm 0.79}$
13	$\mathbf{0.99}_{\pm\mathbf{0.42}}$	$22.03_{\pm 1.04}$	$15.95_{\pm 0.73}$	$15.71_{\pm 0.34}$	$\mathbf{12.50}_{\pm\mathbf{9.56}}$	$18.16_{\pm 6.99}$	$3.61_{\pm 0.64}$	$12.51_{\pm 4.83}$	$\mathbf{4.88}_{\pm\mathbf{0.85}}$
14	$2.77_{\pm 0.48}$	$26.80_{\pm 2.53}$	$25.28_{\pm 1.99}$	$18.90_{\pm 1.04}$	$24.94_{\pm 8.12}$	$14.45_{\pm 7.33}$	$\mathbf{1.86}_{\pm\mathbf{0.62}}$	$\mathbf{12.64}_{\pm\mathbf{5.92}}$	$\mathbf{9.09}_{\pm\mathbf{4.73}}$
15	$\mathbf{1.41}_{\pm\mathbf{0.42}}$	$\mathbf{10.36}_{\pm\mathbf{0.73}}$	$8.89_{\pm 0.74}$	$16.93_{\pm 1.17}$	$26.86_{\pm 3.30}$	$8.69_{\pm 1.27}$	$3.39_{\pm 0.92}$	$12.05_{\pm 2.66}$	$\mathbf{3.95}_{\pm\mathbf{0.52}}$
16	$\mathbf{0.58}_{\pm\mathbf{0.14}}$	$91.64_{\pm 0.15}$	$6.89_{\pm 0.09}$	$4.36_{\pm 0.27}$	$\mathbf{87.04}_{\pm\mathbf{3.51}}$	$\mathbf{5.26}_{\pm\mathbf{0.63}}$	$1.05_{\pm 0.36}$	$89.64_{\pm 0.80}$	$7.21_{\pm 0.42}$
17	$\mathbf{1.91}_{\pm\mathbf{0.27}}$	$72.75_{\pm 0.85}$	$18.83_{\pm 0.48}$	$14.12_{\pm 0.48}$	$50.40_{\pm 1.45}$	$\mathbf{4.05}_{\pm\mathbf{0.56}}$	$2.70_{\pm 1.70}$	$\mathbf{46.22}_{\pm\mathbf{7.91}}$	$10.70_{\pm 0.96}$
18	$\mathbf{0.98}_{\pm\mathbf{0.19}}$	$80.92_{\pm 0.15}$	$15.41_{\pm 0.14}$	$6.42_{\pm 0.82}$	$\mathbf{59.33}_{\pm\mathbf{5.58}}$	$\mathbf{5.30}_{\pm\mathbf{0.43}}$	$1.75_{\pm 0.19}$	$60.84_{\pm 4.55}$	$11.05_{\pm 0.92}$
19	$\mathbf{0.55}_{\pm\mathbf{0.15}}$	$86.08_{\pm 0.27}$	$12.65_{\pm 0.12}$	$5.55_{\pm 1.19}$	$\mathbf{78.99}_{\pm\mathbf{4.96}}$	$\mathbf{7.68}_{\pm\mathbf{0.63}}$	$1.50_{\pm 0.29}$	$86.71_{\pm 0.43}$	$11.15_{\pm 0.47}$
20	$0.98_{\pm 0.15}$	$91.94_{\pm 0.23}$	$6.23_{\pm 0.09}$	$3.03_{\pm 0.34}$	$\mathbf{86.07}_{\pm\mathbf{1.83}}$	$\mathbf{3.03}_{\pm\mathbf{0.70}}$	$\mathbf{0.68}_{\pm\mathbf{0.10}}$	$92.06_{\pm 0.39}$	$5.53_{\pm 0.33}$
21	$1.91_{\pm 0.38}$	$88.61_{\pm 0.18}$	$9.83_{\pm 0.14}$	$5.11_{\pm 0.98}$	$\mathbf{81.78}_{\pm\mathbf{4.69}}$	$\mathbf{3.49}_{\pm\mathbf{0.52}}$	$\mathbf{1.43}_{\pm\mathbf{0.48}}$	$87.93_{\pm 0.48}$	$8.56_{\pm 0.60}$
22	$\mathbf{1.10}_{\pm\mathbf{0.27}}$	$63.95_{\pm 0.75}$	$14.87_{\pm 0.30}$	$8.30_{\pm 1.29}$	$64.87_{\pm 8.07}$	$\mathbf{4.72}_{\pm\mathbf{0.81}}$	$2.04_{\pm 0.83}$	$\mathbf{61.13}_{\pm\mathbf{2.02}}$	$11.72_{\pm 0.84}$
23	$\mathbf{1.57}_{\pm\mathbf{0.36}}$	$\mathbf{24.30}_{\pm\mathbf{1.74}}$	$10.96_{\pm 0.99}$	$11.41_{\pm 1.37}$	$43.39_{\pm 7.84}$	$\mathbf{2.39}_{\pm\mathbf{0.73}}$	$2.09_{\pm 0.85}$	$25.38_{\pm 3.17}$	$7.17_{\pm 0.15}$
24	$\mathbf{1.88}_{\pm\mathbf{0.40}}$	$47.22_{\pm 1.96}$	$13.14_{\pm 0.80}$	$11.07_{\pm 1.03}$	$59.06_{\pm 5.83}$	$\mathbf{3.53}_{\pm\mathbf{1.01}}$	$2.38_{\pm 0.68}$	$\mathbf{43.16}_{\pm\mathbf{1.67}}$	$9.60_{\pm 0.84}$
25	$\mathbf{1.55}_{\pm\mathbf{0.48}}$	$42.32_{\pm 3.98}$	$9.78_{\pm 0.65}$	$8.71_{\pm 0.31}$	$60.56_{\pm 4.27}$	$\mathbf{2.58}_{\pm\mathbf{1.21}}$	$1.65_{\pm 0.35}$	$\mathbf{38.14}_{\pm\mathbf{1.48}}$	$7.12_{\pm 0.45}$
Average	$\mathbf{1.42}_{\pm\mathbf{0.34}}$	$60.35_{\pm 1.08}$	$14.27_{\pm 0.48}$	$9.41_{\pm 0.83}$	$54.67_{\pm 5.63}$	$\mathbf{5.74}_{\pm\mathbf{1.71}}$	$2.08_{\pm 0.73}$	$\mathbf{49.20}_{\pm\mathbf{3.47}}$	$8.86_{\pm 0.97}$

model (1.62). In *Coronary atherosclerosis (10)*, MedFuse nearly doubles EHR's calibration error (3.52 vs. 1.80) but presents the best performance across all other metrics. In fact, the only condition where MedFuse outperforms both unimodal variants in calibration at a statistically significant level is *Other upper respiratory disease (20)* (MedFuse = 0.68 vs. CXR = 3.03, EHR = 0.98). Yet in this case, the improvements do not translate into statistically superior AUROC, AUPRC, or their selective prediction counterparts.

Overall, multimodal fusion reliably boosts discrimination but fails to improve, and sometimes worsens, calibration. This highlights the risk of relying solely on standard metrics when assessing multimodal models in safety-critical settings.

Class-level Calibration Stratification. To probe this behavior, we decompose ECE into positive- and negative-class components as defined by Equation (3.3) (i.e., $\text{ECE}_{c=1}$ and $\text{ECE}_{c=0}$). As shown in Table 4, we observe that positive-class ECE is consistently larger than negative-class ECE, showing that overconfidence in positive predictions is the dominant calibration failure. This imbalance is hidden when reporting only standard ECE and can therefore mislead conclusions about robustness.

We also observe that no single model dominates across all conditions. EHR and MedFuse generally achieve lower positive-class ECE than CXR, while CXR often achieves the lowest negative-class ECE. This inversion illustrates how aggregate calibration masks strengths and weaknesses in specific components and why positive-class miscalibration, in particular, undermines selective prediction. It also suggests that the multimodal improvements in positive-class calibration likely originate from the EHR modality rather than from fusion per se.

Table 5: **Condition-level Performance of Multimodal Architectures**. Although the performance of all three models is comparable, MedFuse reliably outperforms both DrFuse and MeTra across different conditions and evaluation metrics, with the downside that all models present high ECE scores. This further confirms that complex architectures do not necessarily improve predictive performance and do not solve the reliability issues observed in this clinical task. (Dark-bold: $p < 0.05$, Wilcoxon signed-rank test, 5 seeds; Light-bold: highest mean, not significant)

Clinical Condition			**MedFuse**					**DrFuse**					**MeTra**		
	AUROC	**AUPRC**	**Selective AUROC**	**Selective AUPRC**	$\widehat{\textbf{ECE}}\downarrow$	**AUROC**	**AUPRC**	**Selective AUROC**	**Selective AUPRC**	$\widehat{\textbf{ECE}}\downarrow$	**AUROC**	**AUPRC**	**Selective AUROC**	**Selective AUPRC**	$\widehat{\textbf{ECE}}\downarrow$
1	**0.761**±0.005	**0.589**±0.008	**0.804**±0.111	**0.634**±0.109	**2.31**±0.97	0.732±0.004	0.551±0.008	0.767±0.104	0.585±0.113	4.06±1.21	0.722±0.009	0.541±0.011	0.759±0.083	0.573±0.071	2.73±0.39
2	0.856±0.010	0.432±0.019	0.560±0.176	0.075±0.116	**0.88**±0.24	0.899±0.004	0.460±0.013	0.584±0.272	**0.118**±0.166	1.06±0.44	**0.900**±0.004	**0.490**±0.022	**0.602**±0.253	0.112±0.172	1.13±0.32
3	0.713±0.005	**0.209**±0.010	0.633±0.207	0.083±0.049	**0.76**±0.23	0.697±0.013	0.166±0.010	**0.637**±0.169	**0.090**±0.069	2.21±0.58	**0.717**±0.010	0.197±0.014	0.610±0.158	0.085±0.058	1.98±1.12
4	**0.694**±0.006	**0.565**±0.010	**0.727**±0.079	**0.610**±0.071	**2.61**±1.65	0.650±0.012	0.520±0.013	0.692±0.084	0.547±0.089	3.42±1.66	0.628±0.011	0.500±0.008	0.648±0.076	0.529±0.079	3.59±1.65
5	**0.770**±0.003	**0.519**±0.010	**0.787**±0.177	**0.521**±0.169	**2.34**±2.04	0.726±0.006	0.448±0.009	0.693±0.127	0.334±0.152	3.59±2.23	0.709±0.017	0.414±0.019	0.668±0.101	0.299±0.105	2.68±1.31
6	**0.742**±0.004	**0.345**±0.011	**0.717**±0.120	**0.279**±0.127	2.17±0.45	0.670±0.013	0.274±0.015	0.573±0.099	0.137±0.060	**1.94**±0.66	0.618±0.010	0.219±0.007	0.557±0.086	0.140±0.042	2.52±0.61
7	0.673±0.001	0.383±0.003	0.602±0.094	0.241±0.100	**1.46**±0.31	0.693±0.008	0.409±0.008	0.662±0.097	0.349±0.097	1.90±0.93	**0.701**±0.003	**0.427**±0.005	**0.673**±0.091	**0.380**±0.104	1.76±0.51
8	**0.824**±0.004	**0.609**±0.009	**0.826**±0.099	**0.648**±0.119	2.32±0.45	0.691±0.007	0.234±0.006	0.611±0.121	0.104±0.066	**2.30**±0.70	0.674±0.007	0.220±0.011	0.592±0.111	0.107±0.051	2.61±0.72
9	**0.792**±0.004	**0.625**±0.008	**0.842**±0.110	**0.703**±0.123	4.50±1.71	0.732±0.006	0.544±0.009	0.751±0.137	0.527±0.169	3.95±2.06	0.705±0.013	0.503±0.019	0.723±0.103	0.515±0.120	**2.02**±0.62
10	**0.776**±0.002	**0.651**±0.004	**0.835**±0.109	**0.723**±0.115	3.52±1.29	0.735±0.005	0.575±0.005	0.764±0.123	0.574±0.154	4.17±3.96	0.705±0.008	0.545±0.009	0.734±0.093	0.560±0.102	**3.28**±0.71
11	0.855±0.006	0.485±0.012	0.676±0.269	0.198±0.193	**1.15**±0.22	**0.867**±0.005	**0.502**±0.015	0.775±0.267	0.464±0.226	1.41±0.45	0.838±0.008	0.450±0.027	**0.822**±0.192	**0.487**±0.234	2.22±1.11
12	0.711±0.006	0.355±0.010	0.650±0.121	0.200±0.110	**1.74**±0.56	**0.726**±0.007	**0.377**±0.007	**0.673**±0.131	**0.262**±0.132	3.52±1.24	0.700±0.020	0.361±0.024	0.620±0.116	0.223±0.117	2.22±0.45
13	**0.696**±0.002	**0.599**±0.004	**0.766**±0.139	0.668±0.125	3.61±0.64	0.678±0.005	0.596±0.010	0.706±0.089	**0.670**±0.094	3.72±1.65	0.650±0.008	0.568±0.011	0.682±0.083	0.647±0.085	**2.57**±1.67
14	**0.676**±0.005	**0.603**±0.006	**0.714**±0.087	**0.660**±0.085	**1.86**±0.62	0.647±0.008	0.580±0.013	0.639±0.086	0.650±0.074	2.02±1.20	0.627±0.011	0.560±0.009	0.623±0.087	0.620±0.073	2.58±1.81
15	0.718±0.003	**0.685**±0.004	**0.755**±0.114	**0.789**±0.082	**3.39**±0.92	**0.720**±0.005	0.676±0.005	0.738±0.099	0.773±0.074	3.78±0.92	0.698±0.007	0.660±0.017	0.742±0.096	0.754±0.085	3.54±0.48
16	0.665±0.009	0.128±0.003	0.606±0.174	**0.072**±0.043	1.05±0.36	**0.711**±0.006	**0.171**±0.009	0.625±0.221	0.065±0.044	**1.01**±0.19	0.690±0.017	0.159±0.023	**0.625**±0.168	0.069±0.055	1.34±0.49
17	**0.763**±0.004	**0.482**±0.014	**0.746**±0.188	**0.413**±0.175	2.70±1.70	0.714±0.006	0.413±0.008	0.665±0.123	0.281±0.138	3.71±2.60	0.701±0.017	0.386±0.018	0.635±0.103	0.247±0.098	**2.52**±0.89
18	**0.712**±0.003	**0.356**±0.009	**0.658**±0.101	**0.273**±0.097	**1.75**±0.19	0.700±0.005	0.315±0.008	0.625±0.115	0.167±0.094	2.10±0.48	0.674±0.010	0.288±0.016	0.612±0.113	0.157±0.073	2.37±1.06
19	**0.611**±0.005	**0.199**±0.008	**0.568**±0.095	0.119±0.036	**1.50**±0.29	0.598±0.013	0.171±0.006	0.556±0.082	0.120±0.032	1.57±0.58	0.575±0.016	0.157±0.006	0.538±0.080	**0.124**±0.036	1.67±0.44
20	0.649±0.011	0.098±0.003	0.532±0.109	0.054±0.027	**0.68**±0.10	**0.694**±0.015	**0.139**±0.016	**0.565**±0.137	0.053±0.055	1.23±0.58	0.671±0.016	0.113±0.013	0.554±0.130	**0.057**±0.056	1.29±0.44
21	**0.663**±0.014	**0.172**±0.006	0.574±0.103	0.089±0.056	1.43±0.48	0.649±0.003	0.166±0.007	0.569±0.108	0.093±0.046	**1.37**±0.77	0.606±0.010	0.131±0.007	**0.596**±0.114	**0.097**±0.036	2.26±0.22
22	0.750±0.004	0.402±0.005	**0.669**±0.154	**0.214**±0.140	2.04±0.83	**0.762**±0.002	**0.423**±0.008	0.668±0.152	0.209±0.138	2.20±1.12	0.740±0.008	0.388±0.019	0.649±0.162	0.210±0.124	**2.00**±1.18
23	0.817±0.005	**0.645**±0.007	**0.872**±0.101	**0.738**±0.115	**2.09**±0.85	**0.819**±0.003	0.621±0.008	0.829±0.151	0.630±0.169	2.94±1.00	0.818±0.005	0.620±0.019	0.800±0.170	0.543±0.209	2.36±0.82
24	0.776±0.004	0.525±0.010	0.752±0.151	0.475±0.178	**2.38**±0.68	**0.797**±0.005	**0.545**±0.012	**0.776**±0.155	**0.513**±0.197	2.76±0.80	0.783±0.008	0.528±0.015	0.747±0.155	0.436±0.190	2.57±0.80
25	0.820±0.003	0.554±0.005	0.748±0.203	0.413±0.242	**1.65**±0.35	**0.847**±0.004	**0.571**±0.014	**0.772**±0.233	**0.484**±0.253	1.71±0.67	0.830±0.008	0.552±0.026	0.719±0.205	0.367±0.236	2.18±1.00
Average	0.739±0.005	0.449±0.008	0.705±0.136	0.396±0.112	2.08±0.73	0.726±0.007	0.418±0.010	0.677±0.139	0.352±0.116	2.54±1.15	0.707±0.010	0.399±0.015	0.661±0.125	0.333±0.104	2.32±0.83

Takeaway 1. Multimodal fusion improves discrimination but does not reliably improve calibration; positive-class overconfidence is the dominant source of error.

4.2. *Do Other Multimodal Architectures Exhibit the Same Reliability Patterns?*

Extending to DrFuse and MeTRA. To evaluate whether our findings generalize beyond MedFuse, we additionally assess two multimodal architectures, DrFuse (Yao et al., 2024) and MeTRA (Khader et al., 2023), which also integrate EHR and CXR modalities for clinical condition prediction. Despite distinct and more complex design choices, both models continue to exhibit reliability trends consistent with MedFuse. In particular, Table 5 shows that MedFuse is consistently the best performing model in terms of evaluation metrics, without any single architecture meaningfully being best in terms of calibration error across conditions.

Extending our correlation analysis between positive-class ECE AUC and Selective AUC, in Figure C1 we show that all three multimodal architectures exhibit the same behaviour for both AUROC and AUPRC, highlighting the strong positive-class ECE, maintaining the same imbalance of overconfidence in positive predictions affecting selective AUC.

This is further demonstrated by the ECE stratification results across models (Table 6), showing that complex fusion architectures do not tackle the reliability issues of class-dependent miscalibration. In addition, the selective prediction curves (Appendix 4.5, Figures C8 and C9) follow nearly identical patterns to MedFuse: moderate gains at intermediate rejection thresholds followed by collapse at extreme coverage levels, suggesting that the underlying calibration dynamics, not the specific fusion mechanism, drive the observed reliability outcomes. Finally, in Appendix 6 (Figure D1, Figure D2, Figure D3), we visualize three representative clinical conditions (2, 8, 25) comparing all multimodal models, confirming that while DrFuse and MeTRA

Table 6: **Condition-level Performance of ECE stratification Across Multimodal Architectures.** Positive-class ECE dominates across conditions, highlighting overconfidence for in positive predictions for all multimodal architectures regardless of the complexity of the fusion strategy. (Dark-bold: $p < 0.05$, Wilcoxon signed-rank test, 5 seeds; Light-bold: highest mean, not significant)

Clinical Condition	**MedFuse**			**DrFuse**			**MeTra**		
	$\widehat{\textbf{ECE}}\downarrow$	$\widehat{\textbf{ECE}}_{c=1}\downarrow$	$\widehat{\textbf{ECE}}_{c=0}\downarrow$	$\widehat{\textbf{ECE}}\downarrow$	$\widehat{\textbf{ECE}}_{c=1}\downarrow$	$\widehat{\textbf{ECE}}_{c=0}\downarrow$	$\widehat{\textbf{ECE}}\downarrow$	$\widehat{\textbf{ECE}}_{c=1}\downarrow$	$\widehat{\textbf{ECE}}_{c=0}\downarrow$
1	**2.31**$_{\pm0.97}$	**24.85**$_{\pm3.96}$	9.97$_{\pm1.14}$	4.06$_{\pm1.21}$	27.74$_{\pm5.78}$	**7.01**$_{\pm2.40}$	2.73$_{\pm0.39}$	35.27$_{\pm2.04}$	13.65$_{\pm1.70}$
2	**0.88**$_{\pm0.24}$	50.56$_{\pm3.07}$	3.56$_{\pm0.19}$	1.06$_{\pm0.44}$	37.08$_{\pm7.82}$	**2.98**$_{\pm0.46}$	1.13$_{\pm0.32}$	**33.53**$_{\pm10.49}$	3.52$_{\pm0.15}$
3	**0.76**$_{\pm0.23}$	86.93$_{\pm0.14}$	8.05$_{\pm0.25}$	2.21$_{\pm0.58}$	88.15$_{\pm2.27}$	**7.83**$_{\pm1.71}$	1.98$_{\pm1.12}$	**84.65**$_{\pm2.99}$	9.32$_{\pm1.95}$
4	**2.61**$_{\pm1.65}$	**26.90**$_{\pm9.52}$	**13.57**$_{\pm2.63}$	3.42$_{\pm1.66}$	35.79$_{\pm13.68}$	17.54$_{\pm7.67}$	3.59$_{\pm1.65}$	39.75$_{\pm13.97}$	19.53$_{\pm5.75}$
5	**2.34**$_{\pm2.04}$	**40.71**$_{\pm8.56}$	**10.21**$_{\pm1.06}$	3.59$_{\pm2.23}$	49.62$_{\pm9.38}$	12.31$_{\pm1.96}$	2.68$_{\pm1.31}$	60.85$_{\pm4.41}$	16.17$_{\pm0.66}$
6	2.17$_{\pm0.45}$	**61.34**$_{\pm3.97}$	**9.27**$_{\pm0.68}$	**1.94**$_{\pm0.66}$	79.53$_{\pm1.91}$	13.44$_{\pm1.82}$	2.52$_{\pm0.61}$	80.43$_{\pm0.66}$	15.52$_{\pm1.17}$
7	**1.46**$_{\pm0.31}$	66.13$_{\pm1.23}$	19.12$_{\pm0.40}$	1.90$_{\pm0.93}$	65.82$_{\pm3.70}$	18.38$_{\pm1.68}$	1.76$_{\pm0.51}$	**56.89**$_{\pm6.12}$	**16.28**$_{\pm0.59}$
8	2.32$_{\pm0.45}$	**41.64**$_{\pm0.88}$	**4.38**$_{\pm0.89}$	**2.30**$_{\pm0.70}$	82.47$_{\pm3.45}$	9.75$_{\pm1.62}$	2.61$_{\pm0.72}$	82.51$_{\pm4.15}$	10.35$_{\pm1.64}$
9	4.50$_{\pm1.71}$	**24.13**$_{\pm9.07}$	**4.69**$_{\pm1.96}$	3.95$_{\pm2.06}$	34.55$_{\pm11.33}$	10.07$_{\pm2.26}$	**2.02**$_{\pm0.62}$	38.30$_{\pm9.77}$	14.65$_{\pm3.27}$
10	3.52$_{\pm1.29}$	**22.44**$_{\pm4.96}$	**7.08**$_{\pm1.88}$	4.17$_{\pm3.96}$	34.09$_{\pm13.14}$	11.16$_{\pm2.60}$	**3.28**$_{\pm0.71}$	37.68$_{\pm7.23}$	16.13$_{\pm1.71}$
11	**1.15**$_{\pm0.22}$	47.13$_{\pm2.20}$	6.30$_{\pm0.34}$	1.41$_{\pm0.45}$	**35.95**$_{\pm5.61}$	**5.01**$_{\pm0.89}$	2.22$_{\pm1.11}$	55.17$_{\pm11.94}$	6.25$_{\pm1.24}$
12	**1.74**$_{\pm0.56}$	68.84$_{\pm2.85}$	17.64$_{\pm0.79}$	3.52$_{\pm1.24}$	**55.32**$_{\pm11.18}$	**16.14**$_{\pm3.87}$	2.22$_{\pm0.45}$	62.38$_{\pm8.01}$	16.60$_{\pm1.96}$
13	3.61$_{\pm0.64}$	**12.51**$_{\pm4.83}$	**4.88**$_{\pm0.85}$	3.72$_{\pm1.65}$	17.27$_{\pm12.41}$	8.27$_{\pm5.89}$	**2.57**$_{\pm1.67}$	27.10$_{\pm8.21}$	15.39$_{\pm2.24}$
14	**1.86**$_{\pm0.62}$	**12.64**$_{\pm5.92}$	**9.09**$_{\pm4.73}$	2.02$_{\pm1.20}$	16.93$_{\pm11.53}$	12.38$_{\pm6.58}$	2.58$_{\pm1.81}$	14.50$_{\pm10.83}$	10.32$_{\pm7.64}$
15	**3.39**$_{\pm0.92}$	12.05$_{\pm2.66}$	3.95$_{\pm0.52}$	3.78$_{\pm0.92}$	**10.59**$_{\pm3.32}$	**3.01**$_{\pm1.87}$	3.54$_{\pm0.48}$	14.38$_{\pm6.59}$	7.02$_{\pm3.27}$
16	1.05$_{\pm0.36}$	89.64$_{\pm0.80}$	7.21$_{\pm0.42}$	**1.01**$_{\pm0.19}$	88.91$_{\pm0.80}$	**6.84**$_{\pm0.74}$	1.34$_{\pm0.49}$	**88.77**$_{\pm2.63}$	6.89$_{\pm0.98}$
17	2.70$_{\pm1.70}$	**46.22**$_{\pm7.91}$	**10.70**$_{\pm0.96}$	3.71$_{\pm2.60}$	54.74$_{\pm9.45}$	12.52$_{\pm2.57}$	**2.52**$_{\pm0.89}$	65.36$_{\pm3.14}$	15.65$_{\pm0.56}$
18	**1.75**$_{\pm0.19}$	**60.84**$_{\pm4.55}$	**11.05**$_{\pm0.92}$	2.10$_{\pm0.48}$	70.47$_{\pm3.99}$	14.39$_{\pm1.54}$	2.37$_{\pm1.06}$	78.01$_{\pm3.50}$	14.10$_{\pm1.39}$
19	**1.50**$_{\pm0.29}$	86.71$_{\pm0.43}$	**11.15**$_{\pm0.47}$	1.57$_{\pm0.58}$	86.99$_{\pm1.08}$	11.54$_{\pm1.14}$	1.67$_{\pm0.44}$	**86.28**$_{\pm0.82}$	12.37$_{\pm0.55}$
20	**0.68**$_{\pm0.10}$	92.06$_{\pm0.39}$	5.53$_{\pm0.33}$	1.23$_{\pm0.58}$	91.04$_{\pm0.93}$	**4.63**$_{\pm0.81}$	1.29$_{\pm0.44}$	**90.78**$_{\pm2.42}$	5.28$_{\pm1.12}$
21	1.43$_{\pm0.48}$	**87.93**$_{\pm0.48}$	8.56$_{\pm0.60}$	**1.37**$_{\pm0.77}$	89.23$_{\pm0.93}$	**8.20**$_{\pm0.81}$	2.26$_{\pm0.22}$	88.25$_{\pm1.49}$	9.52$_{\pm1.33}$
22	2.04$_{\pm0.83}$	61.13$_{\pm2.02}$	11.72$_{\pm0.84}$	2.20$_{\pm1.12}$	**59.74**$_{\pm7.21}$	**11.26**$_{\pm1.04}$	**2.00**$_{\pm1.18}$	62.64$_{\pm6.49}$	12.51$_{\pm1.24}$
23	**2.09**$_{\pm0.85}$	25.38$_{\pm3.17}$	7.17$_{\pm0.15}$	2.94$_{\pm1.00}$	22.12$_{\pm3.08}$	**4.87**$_{\pm0.58}$	2.36$_{\pm0.82}$	**18.30**$_{\pm5.56}$	5.82$_{\pm1.39}$
24	**2.38**$_{\pm0.68}$	43.16$_{\pm1.67}$	9.60$_{\pm0.84}$	2.76$_{\pm0.80}$	**36.35**$_{\pm4.61}$	**7.43**$_{\pm1.94}$	2.57$_{\pm0.80}$	45.99$_{\pm7.53}$	10.29$_{\pm1.44}$
25	**1.65**$_{\pm0.35}$	38.14$_{\pm1.48}$	7.12$_{\pm0.45}$	1.71$_{\pm0.67}$	**28.90**$_{\pm2.32}$	**4.83**$_{\pm1.20}$	2.18$_{\pm1.00}$	40.13$_{\pm7.60}$	7.04$_{\pm0.57}$
Average	2.08$_{\pm0.73}$	49.20$_{\pm3.47}$	8.86$_{\pm0.97}$	2.54$_{\pm1.15}$	51.98$_{\pm6.04}$	9.67$_{\pm2.23}$	2.32$_{\pm0.83}$	55.52$_{\pm5.94}$	11.61$_{\pm1.82}$

occasionally show local improvements, the overall calibration and selective prediction trends remain virtually unchanged.

Takeaway 2. Advances in multimodal fusion—such as transformer-based models—do not alter the fundamental trend: multimodal architectures enhance discrimination but leave class-dependent miscalibration largely unresolved.

4.3. *Can Loss-Upweight Training Mitigate Calibration Failures?*

Label-Dependent Loss Upweighting. To probe and mitigate class-dependent miscalibration, we implement a simple label-dependent loss reweighting procedure. For each clinical condition c, we assign a weight w_c to its loss term, increasing the contribution of low-prevalence (underrepresented) positive labels during training, which explicitly upweights rare but clinically important conditions. For each multimodal backbone (MedFuse, DrFuse, MeTra), we train a label-upweighted version, enabling us to study how this simple intervention affects class-dependent calibration and selective prediction across architectures and capacities.

Persistent Reliability Gaps. To test whether the observed miscalibration can be mitigated, we utilize a straightforward method using positive prevalence as a scaling loss factor during training to emphasize the model's attention on underrepresented conditions. Table 7 shows the improvement of selective prediction

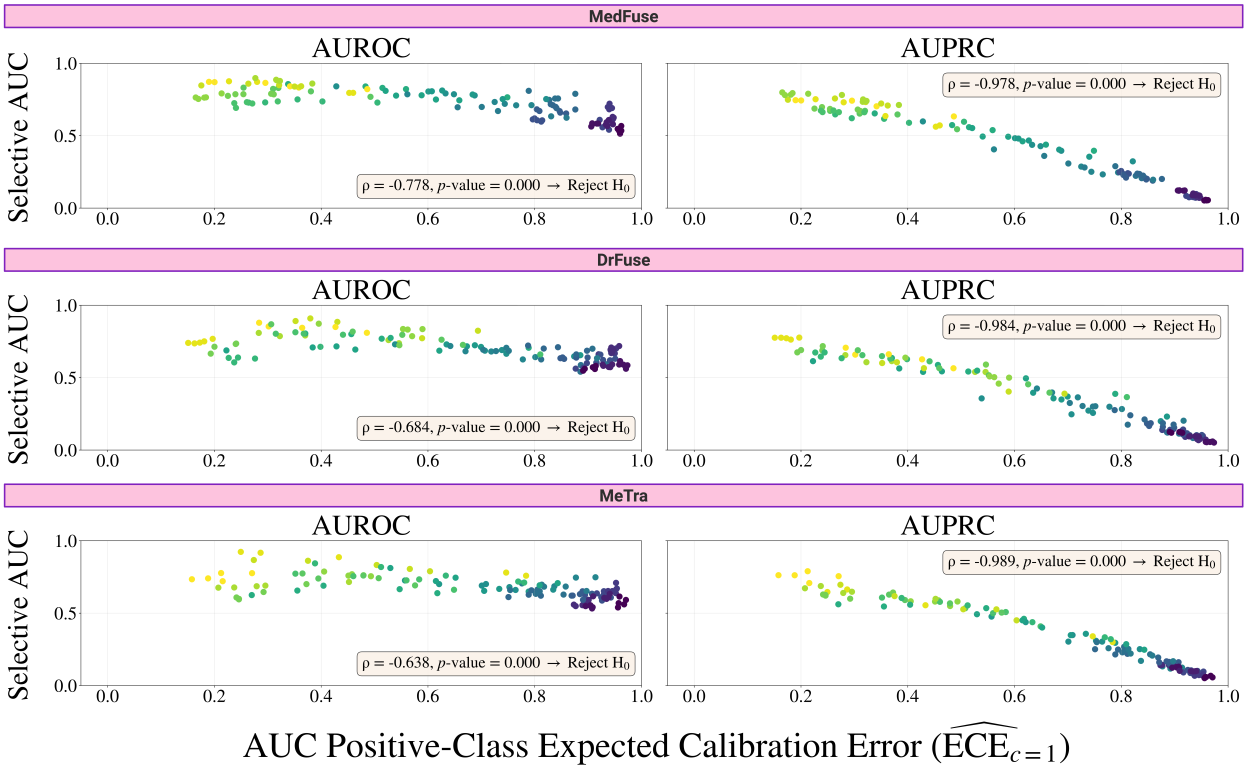

Figure C1: **Do Other Multimodal Architectures Exhibit the Same Reliability Patterns?** For each multimodal architecture, higher positive-class ECE AUC consistently predicts lower selective AUC across conditions for both AUROC and AUPRC, indicating that overconfidence in positive predictions drives instability in selective evaluation regardless of fusion strategy. For all architectures and metrics, we reject the null hypothesis of no non-negative association between stratified ECE and selective AUC (Spearman's rank H_0, $p < 0.05$), showing that increased architectural complexity alone does not resolve performance–reliability trade-offs.

metrics over the standard AUC when using loss upweighting for the MedFuse architecture. We note consistent but modest improvements in selective AUROC and AUPRC, with ΔAUC (Selective–Standard). While average ΔAUPRC values are slightly higher, the improvements remain small and variable across conditions. Still, these improvements fall short of statistical significance, indicating that while selective behaviour trends upward, the underlying reliability gap persists.

Turning to calibration, standard ECE computation shows no global improvement, with only minor and statistically insignificant increases in error for the upweighted variant. However, stratified ECE analysis paints a more nuanced picture: the upweighted model achieves statistically significant positive-class calibration improvements in 23 of 25 conditions, and negative-class improvements in 11 of 25. These findings confirm that loss reweighting effectively reduces overconfidence for rare positive labels, though this does not fully translate into significant enhancements in selective prediction.

Furthermore, when applying the same loss upweight strategy to DrFuse and MeTra we observe similar patterns. In particular, Figure C2 shows that the negative relationship between positive-class ECE AUC and selective AUC has some general improvements as points are clustered closer together in lower ECE AUC values. However, we consistently reject our defined H_0 (no non-negative correlation) with statistical significance which shows that although there are some improvements, this simple loss mechanism does not fully resolve the reliability issues in selective prediction for clinical condition classification.

Table 7: **Loss Upweighting During Training Leads to Inconsistent Improvements.** Improvements are concentrated mainly in stratified ECE but the reliability issue in improvement between standard and selective AUC remains. (Dark-bold: $p < 0.05$, Wilcoxon signed-rank test, 5 seeds)

Clinical Condition	**MedFuse Upscaled**		**Improvement (MedFuse vs MedFuse Upscaled)**				
	Δ AUROC	**Δ AUPRC**	**Δ Selective AUROC**	**Δ Selective AUPRC**	$\Delta\widehat{\textbf{ECE}}\downarrow$	$\Delta\widehat{\textbf{ECE}}_{c=1}\downarrow$	$\Delta\widehat{\textbf{ECE}}_{c=0}\downarrow$
1	-0.004±0.016	**0.084±0.006**	-0.047±0.031	0.038±0.064	0.008±0.013	**-0.210±0.041**	**-0.050±0.033**
2	0.001±0.008	**0.059±0.006**	**0.312±0.015**	**0.415±0.018**	0.015±0.010	**-0.397±0.035**	**-0.015±0.008**
3	**0.028±0.009**	**0.054±0.015**	**0.111±0.005**	**0.176±0.018**	0.088±0.022	**-0.748±0.023**	0.033±0.028
4	-0.042±0.025	**0.068±0.005**	-0.077±0.019	0.021±0.032	0.003±0.005	**-0.194±0.090**	**-0.082±0.046**
5	**0.045±0.018**	**0.103±0.008**	0.018±0.048	0.078±0.101	0.023±0.017	**-0.357±0.099**	-0.048±0.044
6	**0.055±0.033**	**0.097±0.010**	0.085±0.062	**0.165±0.068**	0.026±0.027	**-0.475±0.055**	-0.033±0.042
7	-0.025±0.003	**0.085±0.011**	**0.043±0.009**	**0.235±0.026**	0.044±0.016	**-0.563±0.026**	**-0.087±0.017**
8	**0.041±0.012**	**0.150±0.016**	0.034±0.032	0.086±0.103	0.023±0.008	**-0.215±0.059**	0.020±0.013
9	**0.051±0.028**	**0.112±0.012**	-0.002±0.012	0.023±0.039	0.015±0.012	**-0.175±0.103**	0.032±0.058
10	0.036±0.031	**0.091±0.010**	-0.026±0.023	0.009±0.028	0.023±0.009	**-0.171±0.075**	0.004±0.047
11	**0.040±0.014**	**0.124±0.027**	**0.225±0.029**	**0.410±0.072**	0.035±0.008	**-0.423±0.036**	**-0.013±0.006**
12	0.007±0.013	**0.046±0.018**	**0.070±0.015**	**0.212±0.027**	0.051±0.009	**-0.607±0.048**	**-0.072±0.032**
13	**0.044±0.030**	**0.065±0.014**	-0.032±0.028	-0.013±0.018	0.023±0.007	-0.053±0.058	0.054±0.044
14	0.018±0.036	**0.059±0.008**	-0.030±0.017	-0.001±0.019	0.004±0.010	-0.048±0.083	-0.026±0.085
15	0.012±0.014	**0.096±0.004**	-0.028±0.022	-0.017±0.010	-0.002±0.014	**-0.078±0.022**	0.004±0.022
16	**0.076±0.012**	**0.042±0.004**	**0.125±0.006**	**0.098±0.006**	0.062±0.041	**-0.769±0.068**	0.004±0.041
17	**0.051±0.017**	**0.102±0.006**	0.057±0.049	0.150±0.110	0.024±0.014	**-0.401±0.099**	-0.048±0.044
18	0.005±0.026	**0.101±0.017**	0.065±0.052	**0.195±0.080**	0.040±0.030	**-0.522±0.060**	-0.042±0.044
19	0.011±0.011	**0.052±0.006**	**0.055±0.008**	**0.132±0.006**	0.022±0.011	**-0.807±0.032**	**-0.073±0.019**
20	0.008±0.009	**0.056±0.008**	**0.175±0.016**	**0.146±0.006**	0.038±0.017	**-0.739±0.016**	-0.004±0.021
21	-0.046±0.031	**0.027±0.012**	**0.079±0.049**	**0.116±0.026**	0.051±0.038	**-0.804±0.041**	-0.012±0.044
22	**0.025±0.016**	**0.086±0.020**	**0.120±0.012**	**0.287±0.025**	0.019±0.003	**-0.559±0.020**	**-0.062±0.026**
23	0.007±0.012	**0.104±0.004**	-0.047±0.010	0.019±0.029	0.001±0.010	**-0.193±0.043**	**-0.039±0.020**
24	**0.041±0.012**	**0.110±0.003**	**0.063±0.033**	**0.146±0.050**	0.032±0.017	**-0.379±0.026**	**-0.029±0.015**
25	**0.060±0.008**	**0.142±0.012**	**0.138±0.021**	**0.285±0.061**	0.034±0.007	**-0.325±0.021**	**-0.020±0.012**

Takeaway 3. Loss upweighting for underrepresented labels yields measurable calibration gains, particularly for positive predictions, but fails to produce consistent or statistically significant improvements in selective reliability.

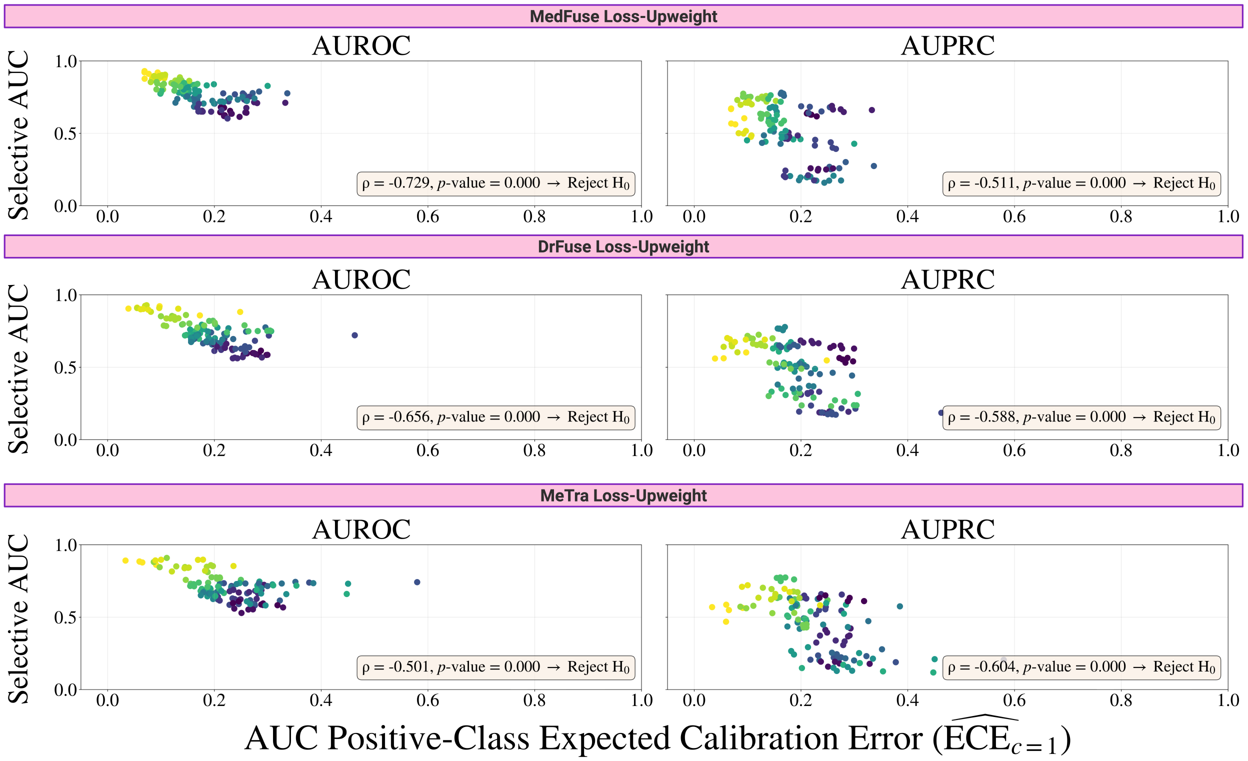

Figure C2: **Can Loss-Upweight Training Mitigate Calibration Failures?** For the loss-upweighted versions of MedFuse, DrFuse, and MeTra, the negative relationship between positive-class ECE AUC and selective AUC remains: conditions with higher stratified ECE still exhibit lower selective AUROC and AUPRC. While points cluster toward lower ECE AUC values, indicating partial calibration improvements, we consistently reject the null hypothesis of no non-negative association between ECE and selective AUC (Spearman's rank $p < 0.05$), showing that this simple loss reweighting does not fully resolve reliability failures in selective prediction for any of the models.

4.4. Regression Analysis of Calibration and Discriminative Performance

To strengthen the link between calibration and selective performance, we fit linear regression models on MedFuse predictions across all 25 clinical conditions using class-stratified ECE (positive and negative) as predictors of both AUROC and AUPRC. Results reveal that stratified ECE, particularly the positive-class component, exhibits a strong negative correlation with AUPRC, reinforcing that miscalibrated confidence in positive predictions reliably forecasts degraded precision and recall under selective evaluation.

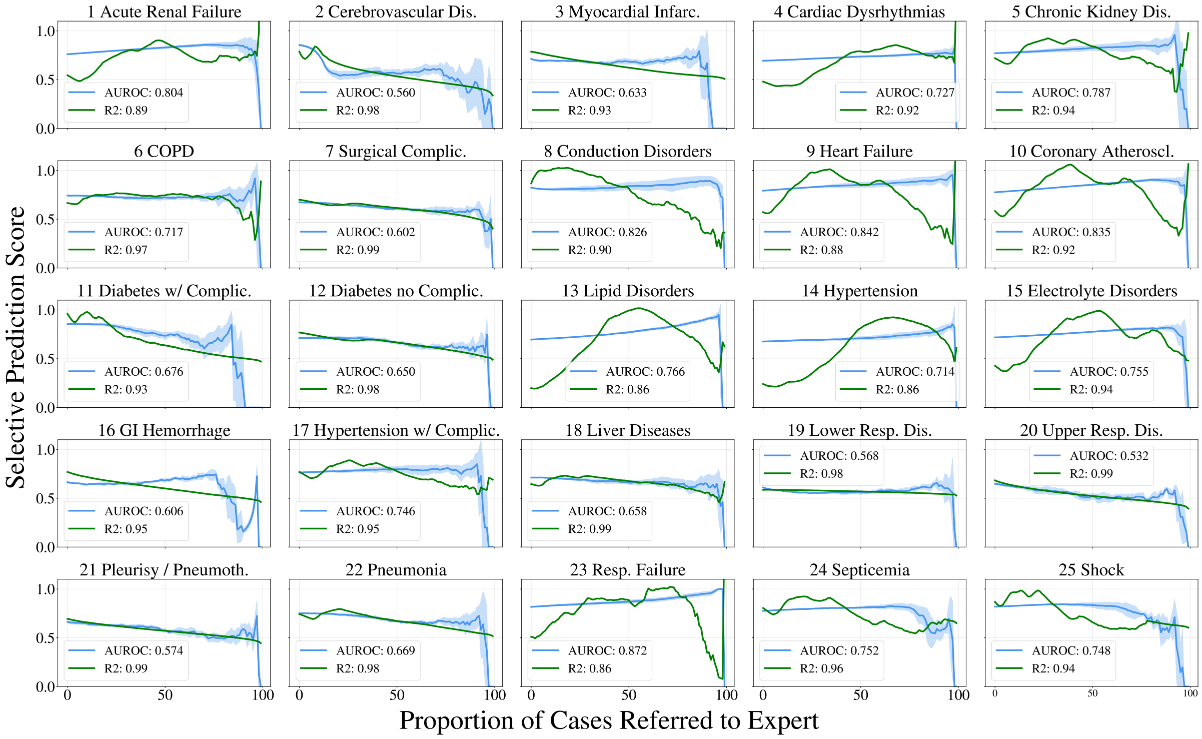

Figure C3: **Linear Regression of AUROC Against Stratified ECE.** Relationship between AUROC and both positive- and negative-class ECE across 25 conditions. Regression slopes demonstrate a mild but consistent negative association, indicating that poorer calibration modestly reduces discriminative performance.

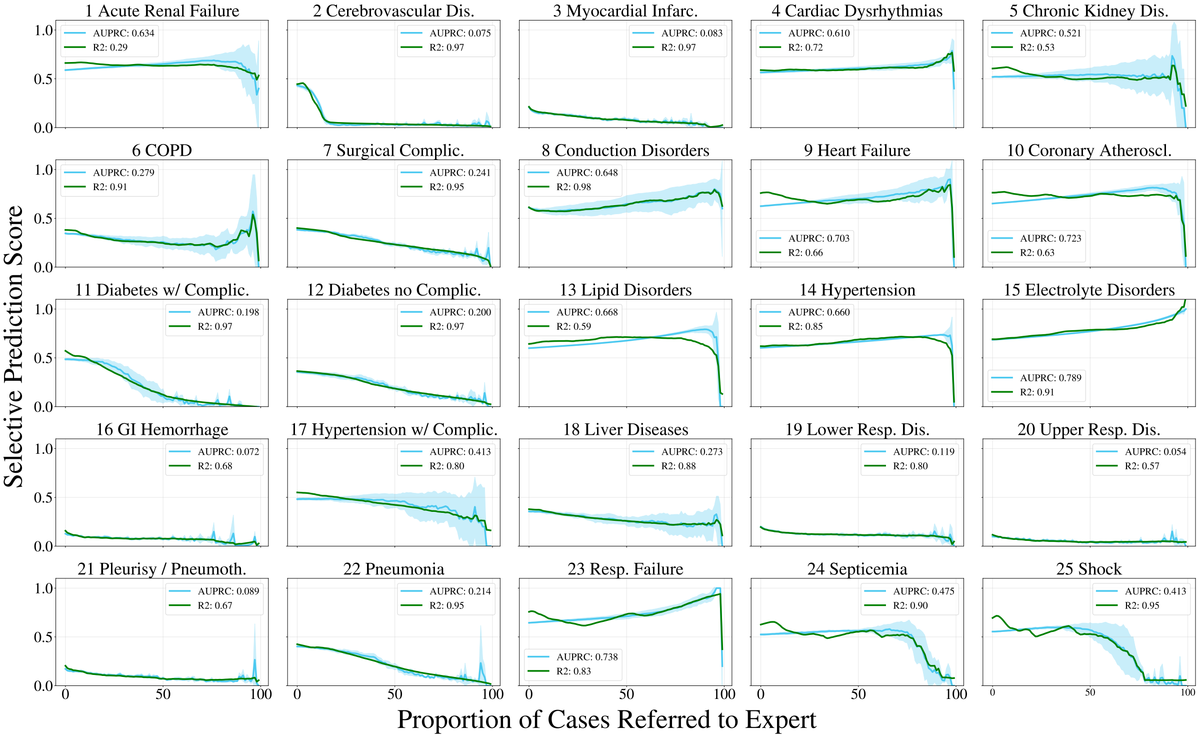

Figure C4: **Linear Regression of AUPRC Against Stratified ECE.** Relationship between AUPRC and class-stratified ECE across 25 conditions. The correlation between ECE and AUPRC exhibits a clear, monotonic decline—confirming that high positive-class miscalibration strongly predicts lower precision-recall performance under selective prediction.

4.5. Stratified Calibration Behavior Across Modalities

To complement the results in Figure 2, we visualize class-stratified ECE and selective AUC for unimodal (CXR, EHR) and multimodal models (MedFuse, DrFuse, and MeTRA) across all 25 clinical conditions. These modality-level grids highlight how calibration dynamics evolve with selective rejection and emphasize that miscalibration, particularly in positive predictions, varies across architectures but follows the same underlying trend.

For the unimodal CXR model, positive-class ECE exhibits substantial variance across seeds, reflecting unstable confidence estimates despite selective AUC patterns that parallel those of MedFuse. The unimodal EHR model shows stratified calibration curves closely aligned with the multimodal results (Figure 2.a), though with consistently higher positive-class ECE for several low-prevalence conditions.

Extending this analysis to DrFuse and MeTRA reveals broadly similar behaviors: both models produce the same selective and calibration trends observed in MedFuse, with only minor condition-specific deviations. These differences, however, do not translate into systematic performance or calibration gains, suggesting that architectural modifications, whether dynamic fusion or transformer-based attention, do not substantially alter the fundamental reliability dynamics across modalities.

Unimodal CXR

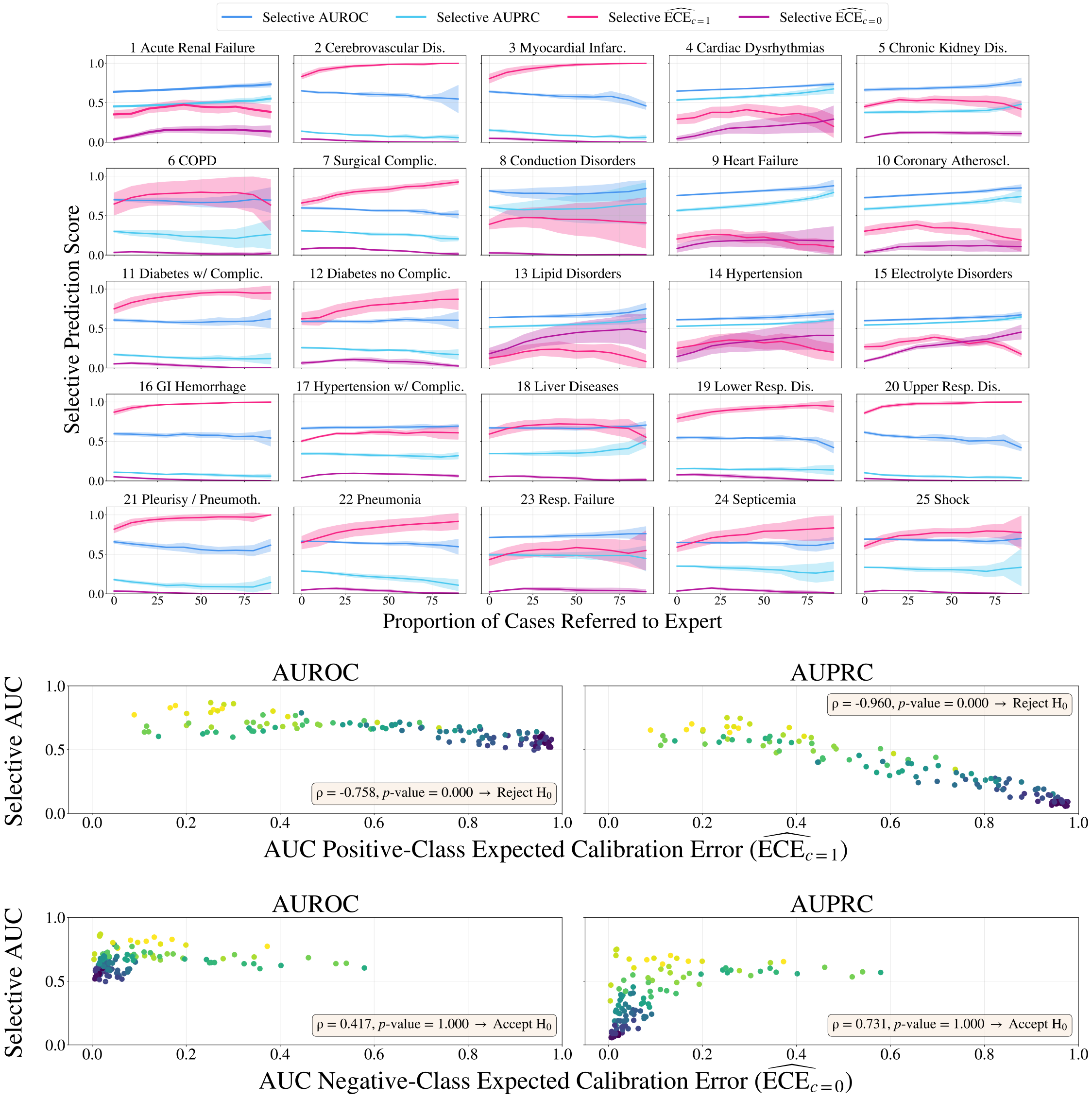

Figure C5: **Stratified calibration and selective prediction for unimodal CXR.** For unimodal CXR, stratified positive-class ECE exhibits substantial variability across seeds and conditions (top), and AUROC/AUPRC remain relatively stable but clearly below the performance of the multimodal MedFuse model. Across conditions, higher positive-class ECE AUC is associated with lower selective AUC (middle), and we reject the null hypothesis of no non-negative association between stratified ECE and Selective AUC (Spearman's rank $p < 0.05$), mirroring the instability observed in the multimodal setting. In contrast, negative-class ECE AUC shows no clear monotonic relationship with selective AUC (bottom), with points spread broadly along the ECE axis, indicating weaker and less interpretable connections between negative-class calibration and selective prediction.

Unimodal EHR

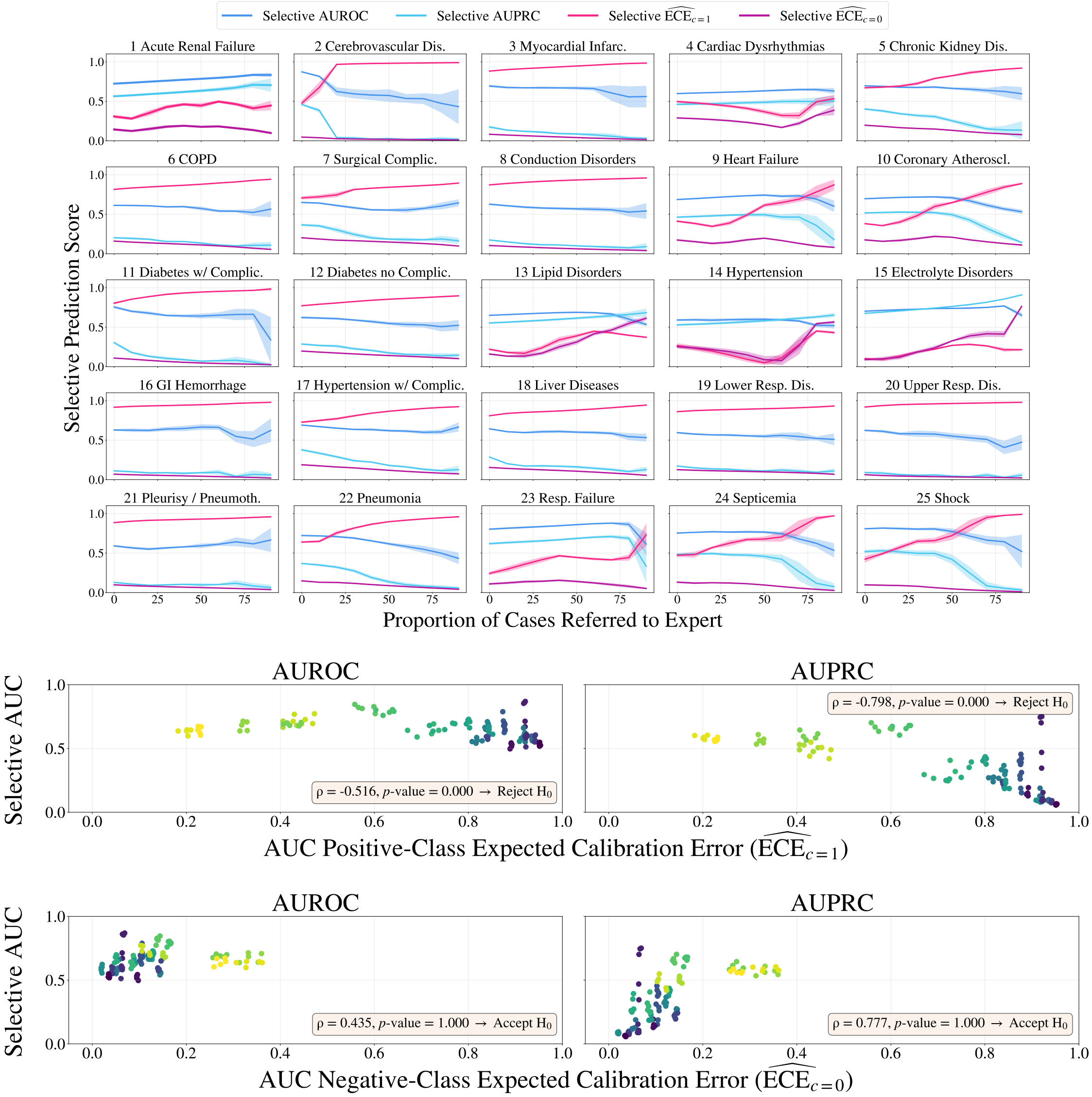

Figure C6: **Stratified calibration and selective prediction for unimodal EHR.** For unimodal EHR, positive-class ECE is higher for several conditions, indicating that EHR-driven predictions are more prone to overconfidence on minority-positive cases. However, both stratified ECE and AUROC/AUPRC exhibit noticeably less variability across seeds and conditions than CXR, suggesting more stable (less noisy) predictions; the resulting trends closely mirror those observed in MedFuse, consistent with EHR driving much of the multimodal performance. In the correlation analysis, positive-class ECE AUC and selective AUC remain negatively associated, though point clusters show more abrupt shifts along the ECE axis and slightly more spread than for CXR. As with CXR, negative-class ECE AUC shows no clear monotonic relationship with Selective AUC, and no consistent trend emerges from the scatter.

MedFuse

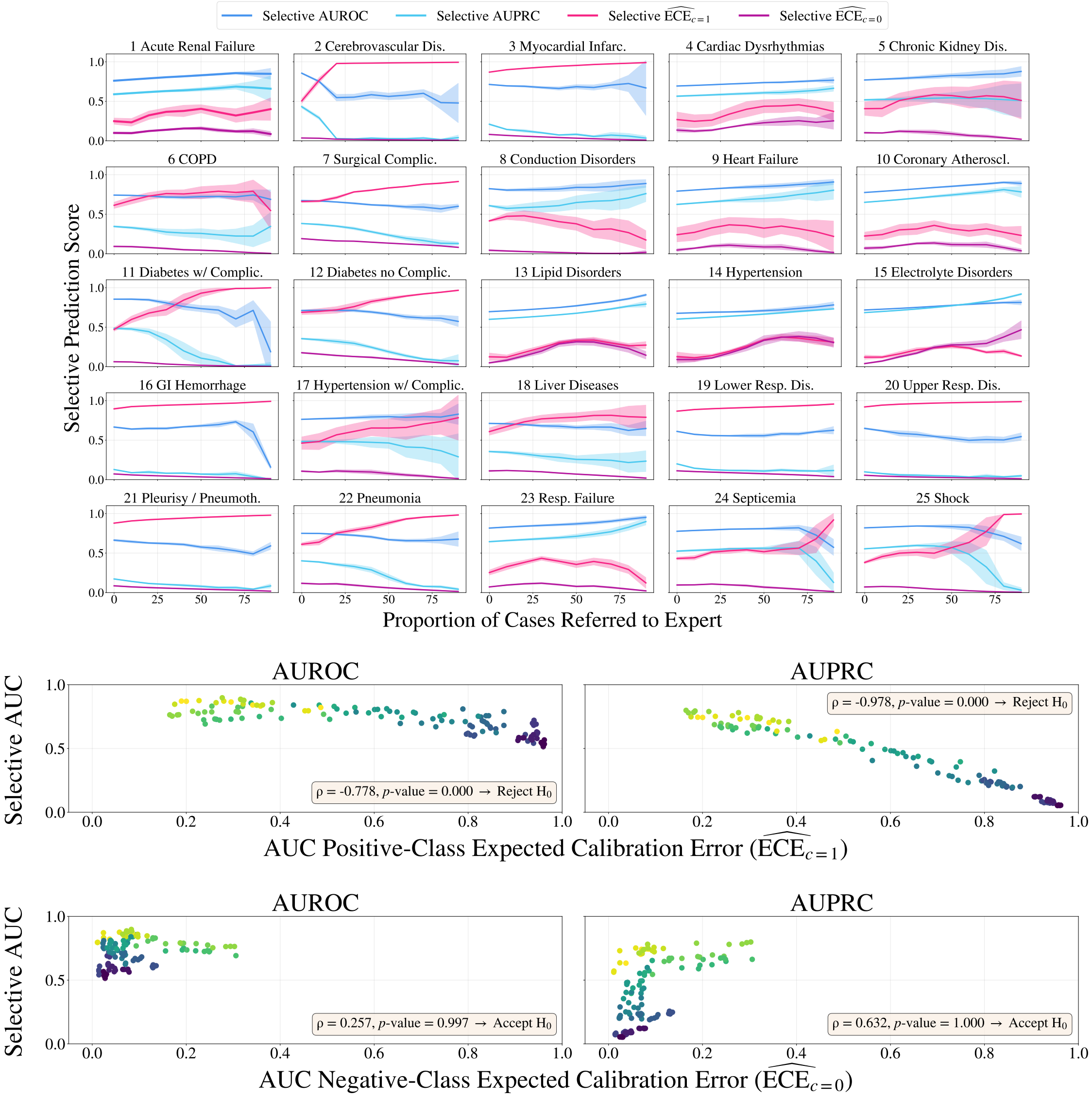

Figure C7: **Stratified calibration and selective prediction for MedFuse.** Across conditions, higher positive-class ECE AUC consistently predicts lower selective AUC, indicating that overconfidence in positive predictions drives instability in selective evaluation. In contrast, most conditions cluster at low negative-class ECE values with positive selective AUC, and no clear monotonic relationship emerges, indicating that well-calibrated negative predictions reliably support selective gains, while miscalibration is primarily driven by overconfident positive predictions.

DrFuse

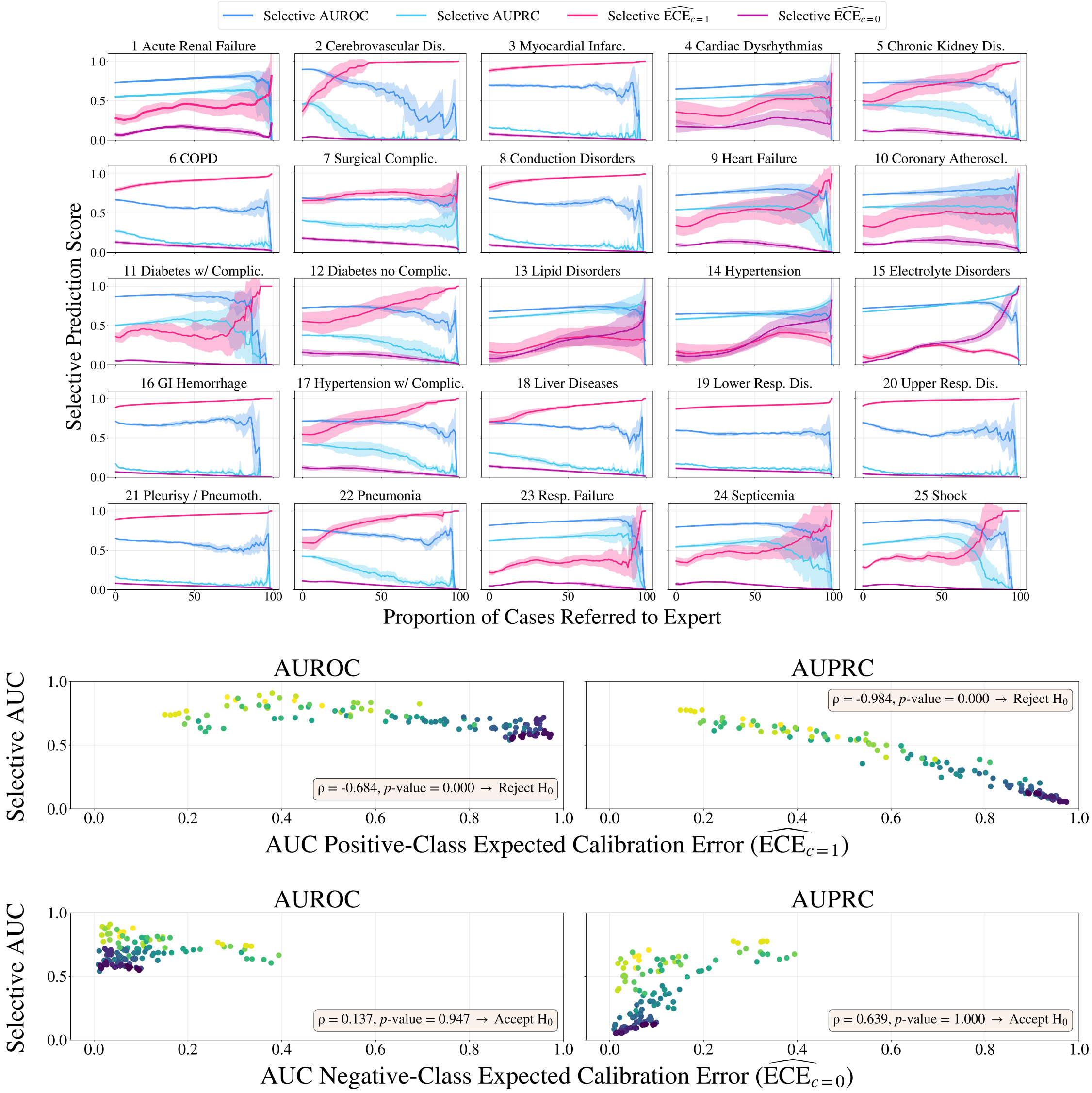

Figure C8: **Stratified calibration and selective prediction for DrFuse.** For DrFuse, stratified ECE and selective AUROC/AUPRC across 25 conditions closely mirror the patterns observed in MedFuse, with only modest gains in aggregate AUROC and AUPRC. As in the other models, higher positive-class ECE AUC is associated with lower Selective AUC, indicating that overconfident positive predictions continue to drive instability in selective evaluation. Negative-class ECE AUC again shows no clear monotonic relationship with Selective AUC, reinforcing that changing the fusion architecture alone does not substantially improve calibration or selective prediction reliability.

MeTRA

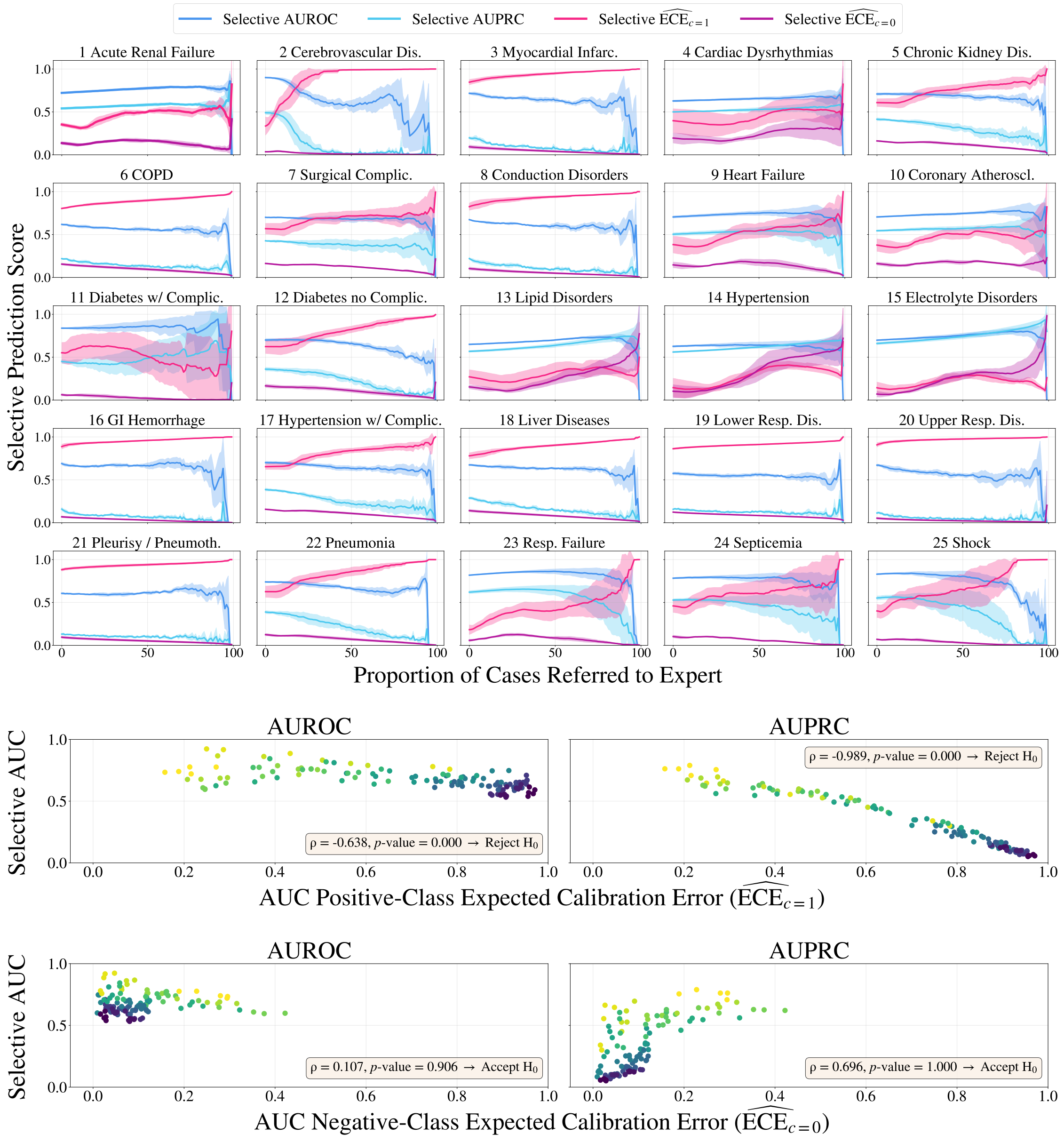

Figure C9: **Stratified calibration and selective prediction for MeTRA.** For MeTRA, stratified ECE and selective AUROC/AUPRC across 25 conditions show overall performance comparable to MedFuse and DrFuse, with similar class-dependent calibration issues. However, positive-class ECE exhibits noticeably higher variance across seeds and slightly noisier selective AUROC/AUPRC trends, which may reflect the adaptation of MeTRA from its original in-hospital mortality task to clinical condition classification. As with the other multimodal architectures, negative-class ECE AUC shows no clear monotonic relationship with selective AUC, indicating that architectural complexity alone does not resolve miscalibration-driven instability in selective prediction.

5. Loss Upweight Training of Multimodal Architectures

To assess whether simple training interventions can mitigate the reliability issues observed in our study, we apply label-dependent loss upweighting to each multimodal backbone (MedFuse, DrFuse, and MeTRA) and analyze the resulting stratified calibration and selective prediction behavior.

MedFuse Upscaled

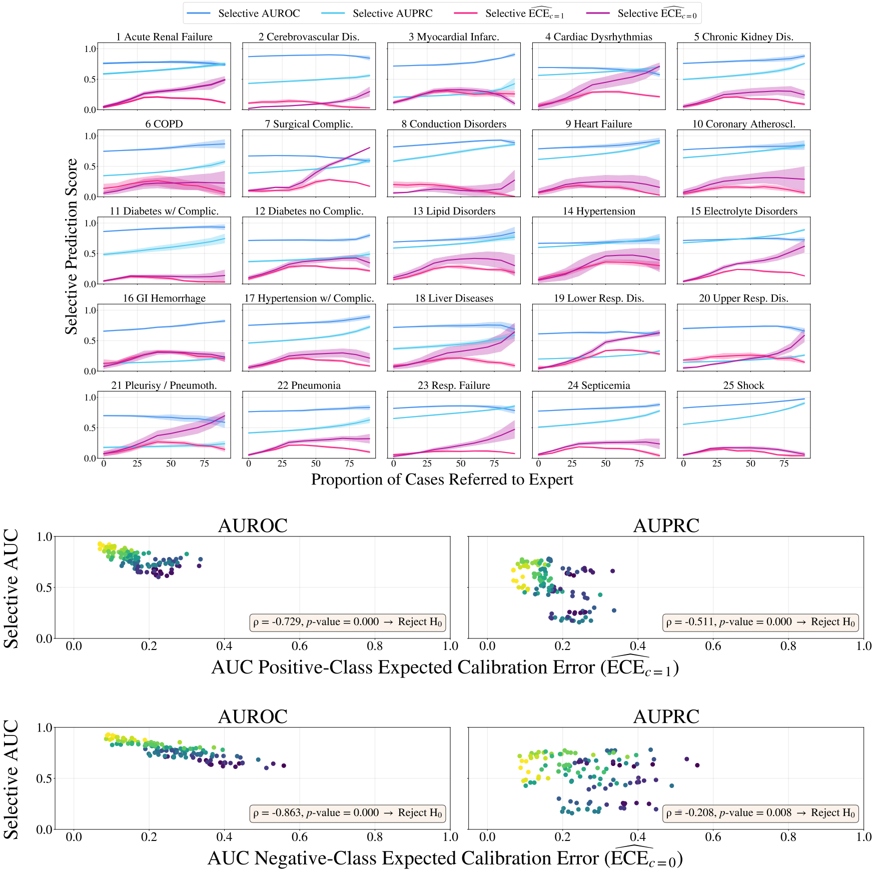

Figure C10: **Stratified calibration and selective prediction for MedFuse (loss-upweighted).** With label-dependent loss upweighting, MedFuse shows reduced variance and lower positive-class ECE across many conditions, although new calibration spikes appear at higher rejection thresholds and selective AUROC/AUPRC remain non-monotonic, indicating that the model is still not well calibrated. Negative-class ECE also worsens for some conditions despite stable discrimination metrics. In the correlation analysis, positive-class ECE AUC and Selective AUC remain negatively associated, with points now clustered more tightly at lower ECE values; for negative-class ECE, a clearer negative relationship with selective AUC emerges and the null of no non-negative association is rejected for both metrics, despite substantial noise in AUPRC. Overall, loss upweighting pulls conditions toward a lower-ECE regime but does not fully resolve the underlying miscalibration driving selective prediction instability.

DrFuse Upscaled

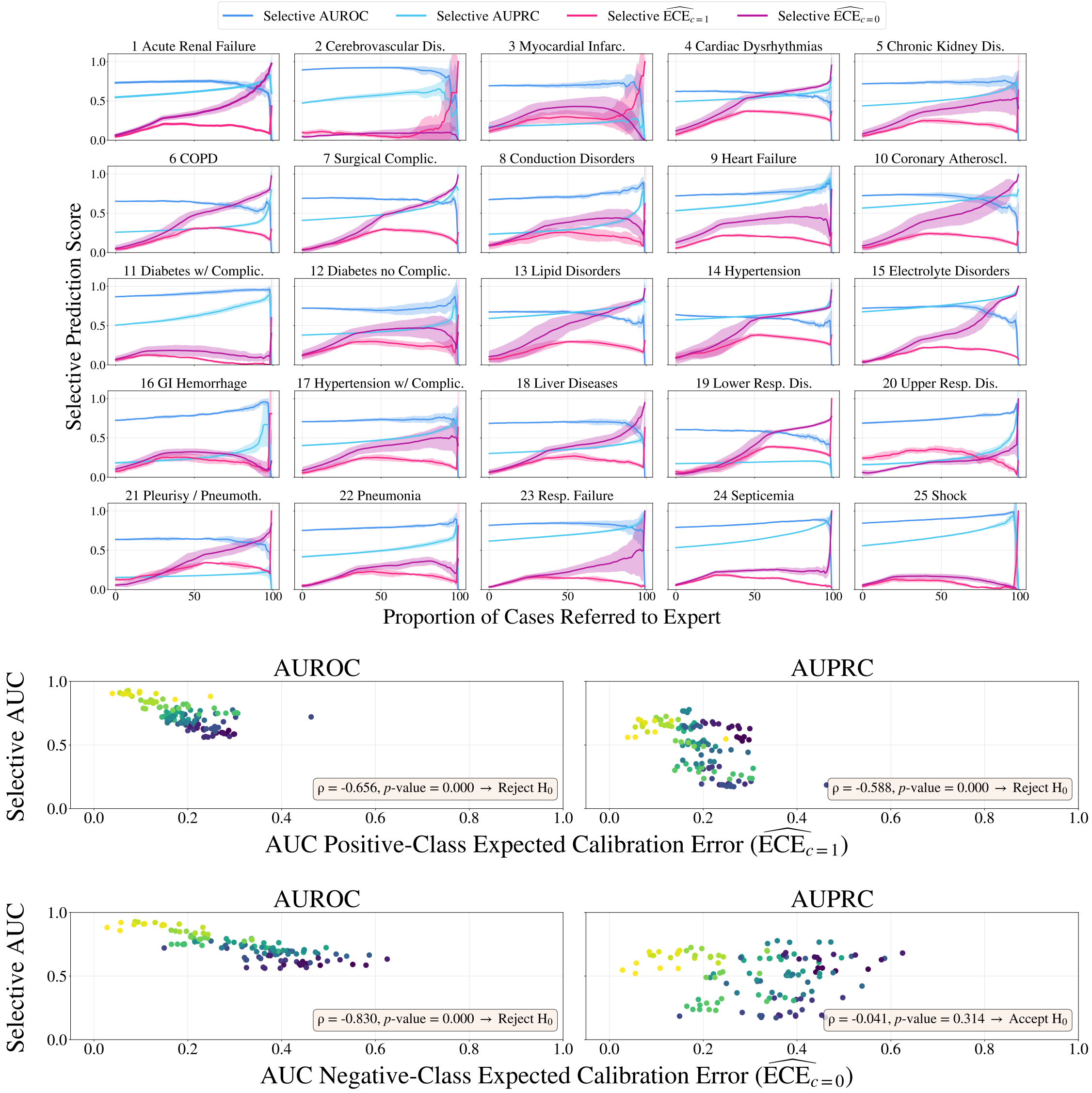

Figure C11: **Stratified calibration and selective prediction for DrFuse (loss-upweighted).** For loss-upweighted DrFuse, stratified ECE and selective AUROC/AUPRC show trends broadly similar to MedFuse, but with smaller apparent gains and substantially higher variance across seeds for negative-class ECE. Across conditions, positive-class ECE AUC remains negatively associated with Selective AUC, indicating that overconfident positive predictions still drive selective instability. For negative-class ECE, however, the correlation analysis is less conclusive: AUROC exhibits a weak negative trend, but AUPRC is highly noisy and the null hypothesis of no non-negative association is not consistently rejected, suggesting that loss upweighting yields limited and architecture-dependent improvements in calibration.

MeTRA Upscaled

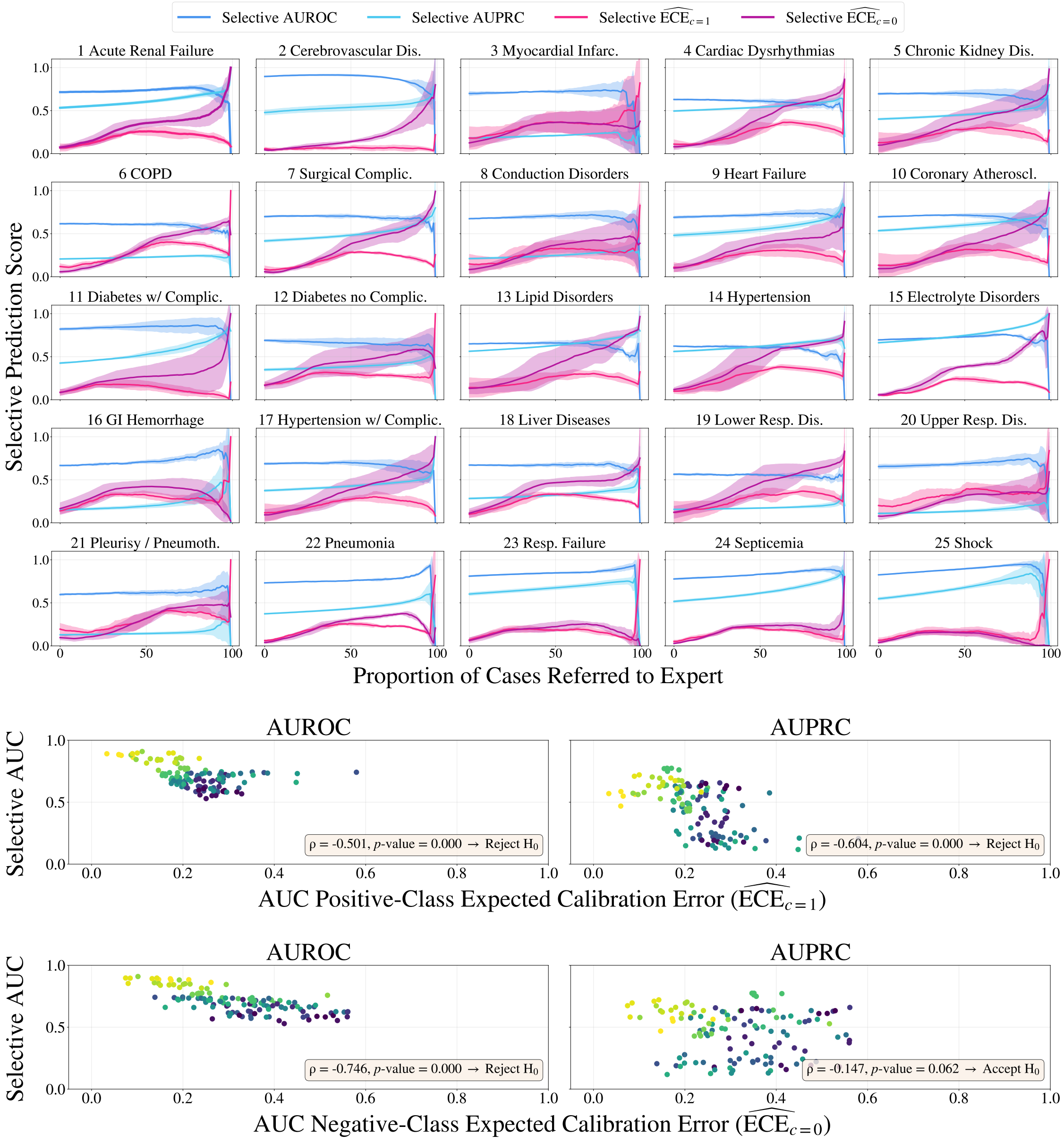

Figure C12: **Stratified calibration and selective prediction for MeTRA (loss-upweighted).** For loss-upweighted MeTRA, stratified ECE and selective AUROC/AUPRC exhibit improvements similar to those observed in MedFuse and DrFuse, with reduced positive-class ECE for several conditions but persistent variability in negative-class ECE. The correlation analysis again shows a negative association between positive-class ECE AUC and Selective AUC, while negative-class ECE displays noisier and less stable trends, particularly for AUPRC, reinforcing that label-dependent loss upweighting yields only partial and architecture-consistent gains in calibration and selective prediction reliability.

6. Per-Condition Comparative Analysis

To provide qualitative insight into how individual modalities and uncertainty-aware training influence selective prediction, we examine three representative conditions: Acute Cerebrovascular Disease (2), Conduction Disorders (8), and Shock (25), across all model variants (EHR, CXR, MedFuse, MedFuse (Loss-Upweight), DrFuse and MeTra). These examples illustrate how modality-specific calibration behaviors interact and how group-aware priors (GAP) mitigate instability and positive-class miscalibration observed in prior sections.

6.1. Label 2: Acute Cerebrovascular Disease

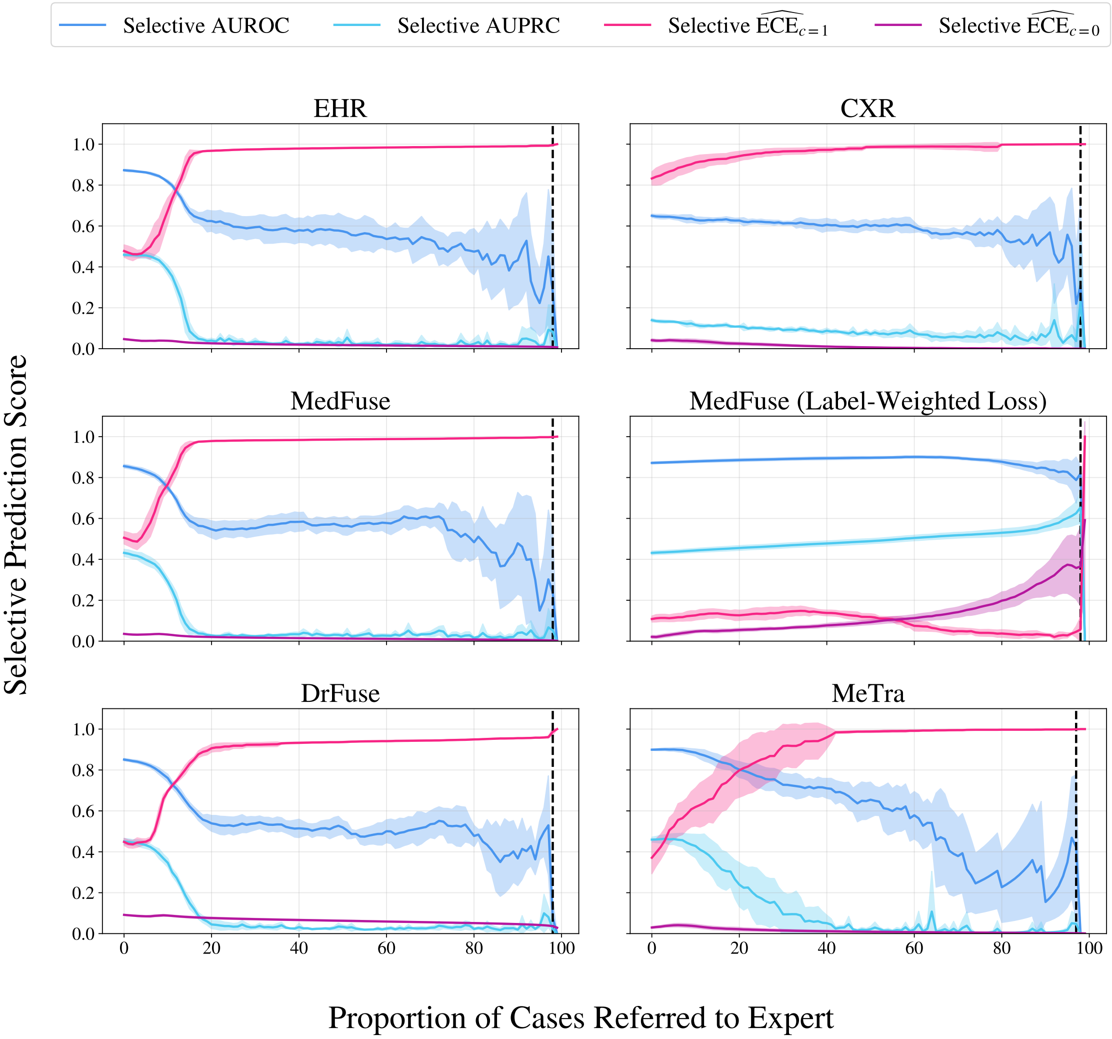

Figure D1: **Comparative analysis for Acute Cerebrovascular Disease.** For this condition, EHR predominantly drives the discriminative performance of MedFuse, while CXR shows poor positive-class calibration and contributes to the oscillatory behavior of selective AUROC. Applying label-weighted loss reduces early-threshold miscalibration in positive predictions but does not fully correct calibration at higher thresholds, leaving the overall selective prediction trend largely unchanged.

6.2. Label 8: Conduction Disorders

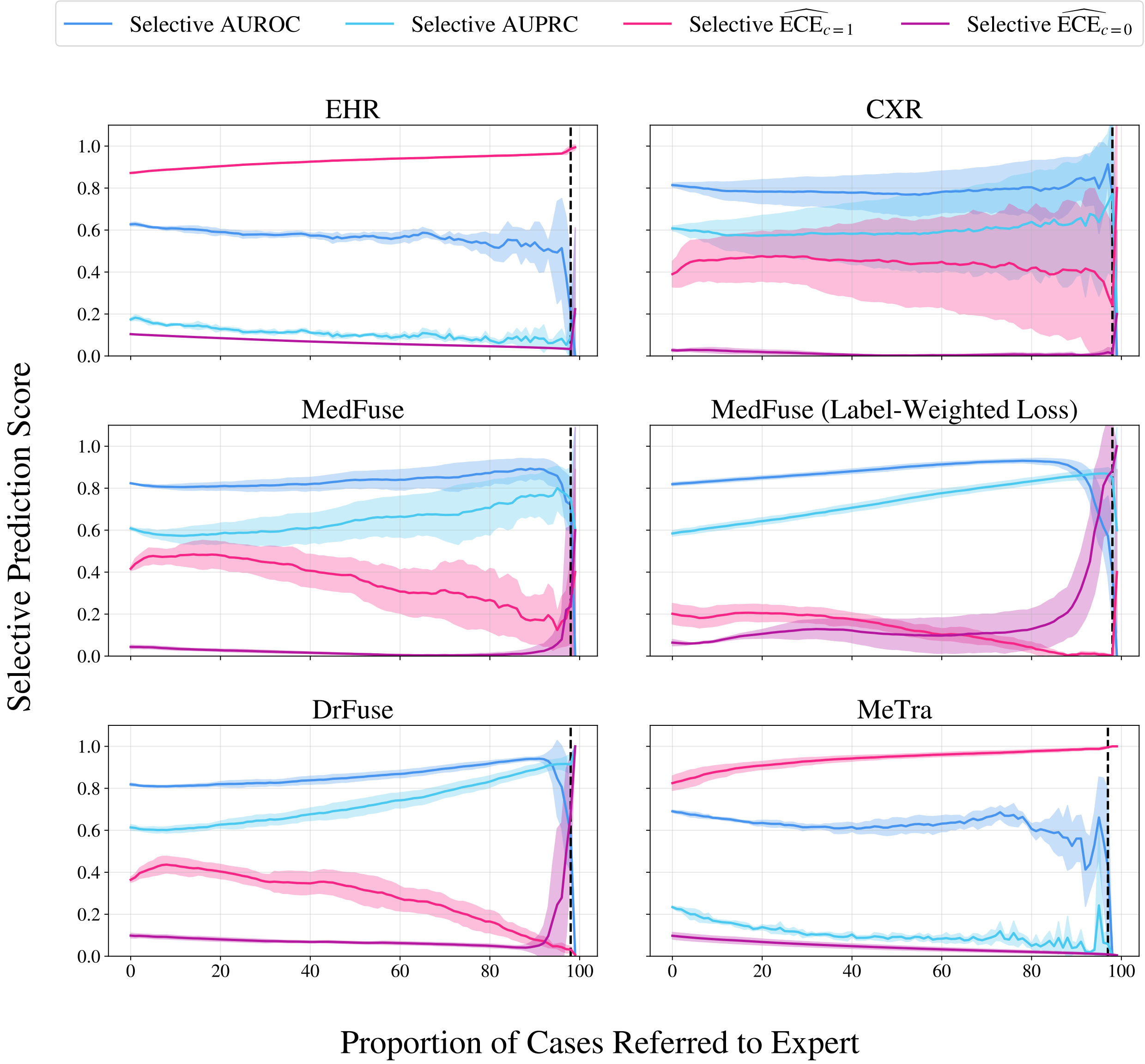

Figure D2: **Comparative analysis for Conduction Disorders.** For this condition, the CXR model shows substantial seed-dependent variability, achieving higher mean performance but poorer calibration consistency than EHR. MedFuse integrates both modalities to improve selective AUROC and stabilize performance, though positive-class ECE remains variable. Label-weighted loss further reduces seed variance and attenuates positive-class ECE at higher confidence thresholds, yielding smoother and more reliable selective behavior, but with similar performance drops at the final thresholds.

6.3. Label 25: Shock

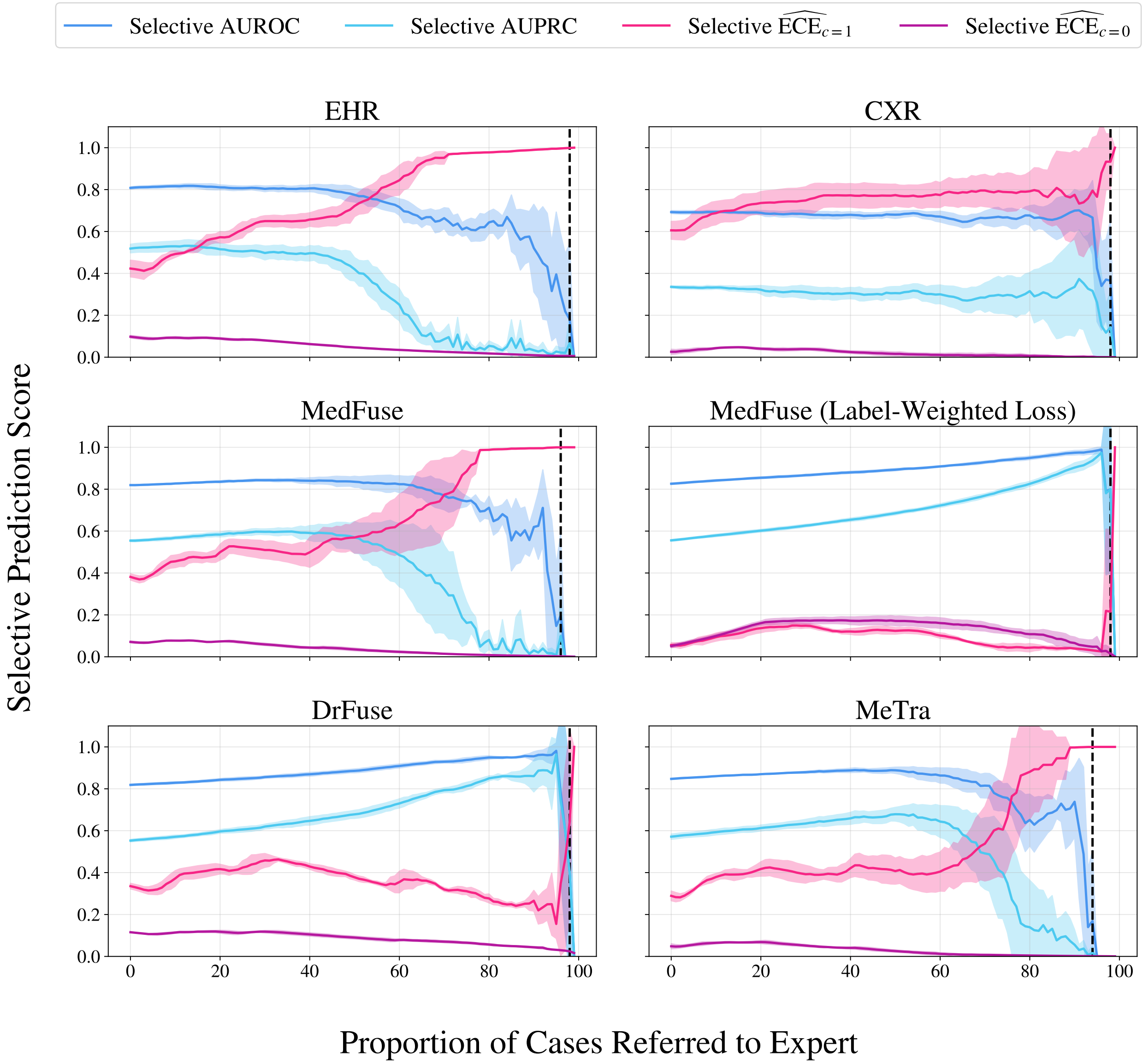

Figure D3: **Comparative analysis for Shock.** For this condition, unimodal EHR and CXR achieve comparable baseline performance, with EHR showing slightly lower positive-class ECE at early thresholds. As rejection increases, however, EHR's AUROC quickly deteriorates while CXR remains uniformly undercalibrated, and MedFuse largely inherits these calibration inconsistencies. Label-weighted loss substantially improves both stability and calibration, suppressing positive-class ECE spikes and producing a clearer monotonic gain in selective AUROC. This condition provides one of the strongest examples of how simple loss upweighting can directly reduce positive-class miscalibration and enhance the reliability of uncertainty estimates.