

An Empirical Analysis of Calibration and Selective Prediction in Multimodal Clinical Condition Classification

L. Julián Lechuga López

Farah E. Shamout

New York University, New York University Abu Dhabi

Tim G. J. Rudner

University of Toronto

LEPOLDO.LECHUGA@NYU.EDU

FARAH.SHAMOUT@NYU.EDU

TIM.RUDNER@UTORONTO.CA

Abstract

As artificial intelligence systems move toward clinical deployment, ensuring reliable prediction behavior is fundamental for safety-critical decision-making tasks. One proposed safeguard is selective prediction, where models can defer uncertain predictions to human experts for review. In this work, we empirically evaluate the reliability of uncertainty-based selective prediction in multilabel clinical condition classification using multimodal ICU data. Across a range of state-of-the-art unimodal and multimodal models, we find that selective prediction can substantially *degrade* performance despite strong standard evaluation metrics. This failure is driven by severe class-dependent miscalibration, whereby models assign high uncertainty to correct predictions and low uncertainty to incorrect ones, particularly for underrepresented clinical conditions. Our results show that commonly used aggregate metrics can obscure these effects, limiting their ability to assess selective prediction behavior in this setting. Taken together, our findings characterize a task-specific failure mode of selective prediction in multimodal clinical condition classification and highlight the need for calibration-aware evaluation to provide strong guarantees of safety and robustness in clinical AI.

Data and Code Availability This study used the publicly available MIMIC-CXR and MIMIC-IV datasets, which can be accessed via PhysioNet subject to completion of the required training and approval of the corresponding data use agreement. Code developed for our study is available [here](#).

Institutional Review Board (IRB) This study does not involve human subjects, so 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 trustworthy and interpretable, allowing 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 from cases of high uncertainty, selective prediction provides a fail-safe mechanism against critical errors that can endanger patients (Figure 1.a). 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 (Figure 1.b).

Unfortunately, evidence from unimodal domains such as computer vision and time-series analysis shows that many models 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

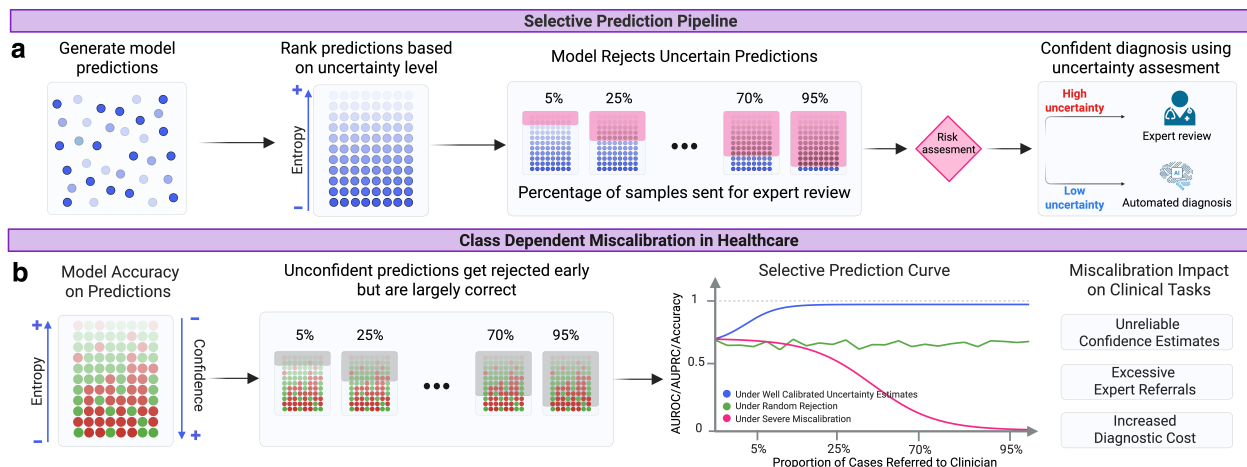


Figure 1: **Selective Prediction Can Serve as a Fail-Safe Mechanism in Safety-Critical Healthcare Settings.** a) A model outputs prediction confidence/uncertainty scores; low-confidence cases are rejected and referred to an expert, yielding coverage-controlled evaluation and diagnosis. b) Miscalibration 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.

records (EHR), imaging, and text are integrated to provide a holistic view of a patient’s state (Simon et al., 2025).

In this paper, we study whether state-of-the-art multimodal models provide reliable uncertainty estimates that enable effective selective prediction in multilabel clinical condition classification. 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 predictions are systematically overconfident, especially for conditions underrepresented in the data, undermining the effectiveness of selective prediction.

Finally, we investigate whether a simple calibration training strategy can mitigate these issues and

find that it only offers modest improvements, falling short of resolving the underlying calibration shortcomings. Our findings underscore calibration-aware selective prediction as a key open challenge for safe deployment of clinical AI systems. To summarize, our key contributions are:

1. We show that selective prediction in multilabel clinical condition classification degrades across a wide range of clinical conditions, and quantify how class-dependent miscalibration correlates with changes in selective AUC.
2. We demonstrate that this degradation persists across state-of-the-art multimodal architectures, indicating that class-dependent miscalibration is not mitigated by architectural choice alone.
3. We evaluate a simple loss up-weighting strategy to correct class-dependent miscalibration, showing limited but measurable improvements in selective prediction performance.
4. We provide robustness analyses across alternative ECE binning strategies, Adaptive ECE, and Brier score, showing that the observed calibration-selective prediction misalignment persists across calibration metrics.

Table 1: **Performance Across Conditions.** Aggregate metrics suggest that multimodal learning (MedFuse) outperforms unimodal baselines. However, these results mask substantial condition-specific variability. Particularly, standard calibration measures (ECE) are deceptively low, concealing severe and consistent miscalibration across individual clinical conditions. This condition-specific variability, motivates the class-level and selective prediction analysis presented in our study. Other multimodal architectures (DrFuse, MeTra) achieve similar aggregate performance, yet, we show that they exhibit comparable calibration failures despite more complex fusion mechanisms. We report improvement gains over EHR given its widespread use as the foundation for most clinical prediction models. (**Best**, **Second Best**).

Model	AUROC	AUPRC	Selective AUROC	Selective AUPRC	$\widehat{ECE} \downarrow$
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	1.42\pm0.335
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
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
MedFuse (Hayat et al., 2022)	0.739 \pm 0.005	0.449\pm0.008	0.705 \pm 0.136	0.396 \pm 0.112	2.08\pm0.725
MedFuse + Temp. Scaling (Guo et al., 2017)	0.739 \pm 0.005	0.449\pm0.008	0.705 \pm 0.135	0.398 \pm 0.110	1.97 \pm 0.60
MedFuse + Loss-Upweighting (King and Zeng, 2001)	0.742\pm0.006	0.447 \pm 0.010	0.764\pm0.098	0.532\pm0.093	4.88 \pm 1.470
MedFuse + Group Aware Priors (Rudner et al., 2024)	0.743\pm0.006	0.451\pm0.008	0.729\pm0.111	0.474\pm0.096	8.31 \pm 1.061
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-Upweighting vs EHR	+0.061 (9.0%)	+0.076 (20.5%)	+0.139 (22.2%)	+0.240 (82.2%)	+3.46 (243.7%)

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 may refer to multiple imaging views of the same modality (e.g., frontal and lateral chest X-rays) (Warner et al., 2024) or 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; Adebisi 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 vitals, free-text notes, and even patient-provider conversations (Mahesh et al., 2024; Zambrano Chaves et al., 2025).

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 robust and reliable probability estimates for safe deployment (Zhao et al., 2024; Deng et al., 2025). Calibration and

selective prediction are closely connected in safety-critical prediction settings. Calibration methods aim to align predicted probabilities with empirical correctness, with common post-hoc approaches including temperature scaling, Platt scaling, and isotonic regression (Guo et al., 2017). Selective prediction instead allows a model to abstain from uncertain cases, but its effectiveness depends on the reliability of the confidence or uncertainty scores used to determine when to abstain (Geifman and El-Yaniv, 2017). Thus, even when a model achieves strong discrimination, miscalibrated confidence estimates can cause selective prediction to reject cases that are relatively safe while retaining cases where additional human review would be needed.

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), and 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 investigat-

ing how multimodal learning impacts selective prediction, calibration, and reliability in clinical settings.

Our work complements prior studies by focusing specifically on the interaction between multimodal fusion, class-dependent calibration, and selective prediction in multilabel clinical condition classification. Rather than proposing a new calibration method, we examine whether existing multimodal architectures provide uncertainty estimates that are reliable enough to support abstention-based safety mechanisms. This perspective addresses a gap in existing evaluations of multimodal clinical models, which often emphasize aggregate discrimination while providing less insight into whether probability estimates remain reliable across labels, classes, and clinically imbalanced conditions.

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:

$$\text{ECE} = \mathbb{E}_{\hat{P}} [|\mathbb{P}(\hat{Y} = Y | \hat{P} = p) - p|]. \quad (1)$$

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

$$\widehat{\text{ECE}} = \sum_{m=1}^M \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|, \quad (2)$$

where $\text{acc}(B_m)$ and $\text{conf}(B_m)$ 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,

$$\text{ECE}_c = \mathbb{E}_{\hat{P}|\hat{Y}=c} [|\mathbb{P}(\hat{Y} = c | \hat{P} = p) - p|], \quad (3)$$

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 $\{B_m^c\}_{m=1}^M$:

$$\widehat{\text{ECE}}_c = \sum_{m=1}^M \frac{|B_m^c|}{n_c} |\text{acc}(B_m^c) - \text{conf}(B_m^c)|, \quad (4)$$

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 $s : \mathcal{X} \rightarrow \mathbb{R}$ 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:

$$(p(\mathbf{y} | \cdot; f), s)(\mathbf{x}) = \begin{cases} p(\mathbf{y} | \mathbf{x}; f), & \text{if } s \leq \tau \\ \perp, & \text{otherwise,} \end{cases} \quad (5)$$

with $p(\mathbf{y} | \cdot; f)$ 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: $s = \mathcal{H}(p(\mathbf{y} | \mathbf{x}; f))$. 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 $(p(\mathbf{y} | \cdot; f), s)(\mathbf{x})$, we compute the predictive performance of the classifier $p(\mathbf{y} | \mathbf{x}; f)$ 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.

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:

$$\mathcal{H}(p(\mathbf{y} | \mathbf{x})) = - \sum_c p(\mathbf{y} = c | \mathbf{x}) \log p(\mathbf{y} = c | \mathbf{x}), \quad (6)$$

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

4. Empirical Setup

4.1. Clinical Task and Dataset

Clinical Condition Classification. We study multilabel clinical condition classification, where the goal is to predict whether a patient presents any of 25 chronic, acute, or mixed clinical conditions during an ICU stay (Hayat et al., 2022). This task is particularly relevant for analyzing selective prediction and calibration because each condition corresponds to a distinct patient subpopulation with varying prevalence, severity, and clinical consequences. As a result, strong average performance across labels may obscure systematic failures for underrepresented or high-risk conditions, leading to unreliable predictions for specific patient groups.

In safety-critical clinical settings, accurately identifying uncertainty at the condition level is essential: miscalibration for even a small subset of conditions can result in missed diagnoses or inappropriate deferral behavior, despite seemingly robust aggregate performance. The multilabel nature of this task therefore provides a natural and clinically meaningful setting to study class-dependent miscalibration and its impact on selective prediction. The full list of clinical conditions and their prevalence is provided in Section A and Table A.1.

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} and x^{cxr} , along with multilabel ground truth y . We follow standard preprocessing procedures and ensure patient-level separation between training and test splits, consistent with prior work (Hayat et al., 2022).

4.2. Model Variants

Throughout this study, we use *unimodal* to refer to models that operate on a single clinical data source, such as structured EHR time-series or chest radiographs alone. We use *multimodal* to refer to models that combine two or more clinical data sources before prediction, allowing the information to be integrated within a single classifier.

Base 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.

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 initialized 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 that integrate EHR and CXR modalities for clinical condition classification: DrFuse (Yao et al., 2024), which learns 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.

Label-Dependent Loss-Upweighting. To probe class-dependent miscalibration and assess whether it can be mitigated, we apply a simple label-dependent loss-upweighting scheme: for each condition c , a weight w_c upweights low-prevalence positive labels during training. This is not intended as a complete solution, but as a lightweight intervention to quantify its effect on class-level calibration and selective prediction across multimodal backbones (MedFuse, DrFuse, MeTra).

4.3. Evaluation Metrics

We report discrimination metrics (AUROC, AUPRC), standard calibration (ECE), and class-stratified calibration metrics ($ECE_{c=0}$, $ECE_{c=1}$). Additional robustness metrics, including ECE across different bins, Adaptive ECE and Brier score, are reported in Section C.4. Selective prediction performance is evaluated using selective AUROC and selective AUPRC. We summarize selective prediction performance by the area under the resulting selective prediction curves.

5. Results

We present empirical results analyzing selective prediction and calibration behavior in multilabel clinical condition classification using multimodal ICU data. Our findings are organized around four guiding questions:

1. Do multimodal models improve or degrade selective prediction performance in this task?
2. How does calibration influence selective prediction behavior?
3. Are class-dependent calibration effects consistent across multimodal architectures?
4. Can loss-upweighting mitigate observed calibration failures?

Our experiments are conducted on a commonly used ICU benchmark pairing structured EHR time-series with chest X-ray images. This multimodal setting has been widely adopted in prior work on clinical risk prediction and patient deterioration modeling, serving as a practical testbed for studying multimodal fusion in critical care contexts (Yang et al., 2021; Khader et al., 2023; Wang et al., 2024; Insalata et al., 2024; Yao et al., 2024; Zheng et al., 2025). Within this setting, our results characterize how calibration and selective prediction behave for multilabel clinical condition classification, and should be interpreted as task- and data-specific rather than as general claims about multimodal clinical models. Additional qualitative results and analysis are provided in Sections C to E, offering complementary views of the findings discussed below.

5.1. Do Multimodal Models Improve or Degrade Selective Prediction Performance in This Task?

Overall Performance. Across the 25 clinical conditions, the multimodal baseline MedFuse generally achieves higher AUROC, AUPRC, and selective AUROC than both unimodal variants, often with statistically significant improvements. These gains confirm that multimodal fusion improves discrimination in this task. However, improvements in discrimination are not accompanied by consistent gains in calibration, revealing a recurrent mismatch between predictive accuracy and probability alignment.

For example, in *Acute and unspecified renal failure (1)*, MedFuse achieves the highest AUROC (0.761) and AUPRC (0.589), and reduces calibration error

relative to CXR (ECE: 12.77 vs. 2.31), yet EHR remains the best calibrated model overall (ECE: 1.62). In *Coronary atherosclerosis (10)*, MedFuse nearly doubles EHR’s calibration error (3.52 vs. 1.80) while still achieving the strongest discrimination across all other metrics. The only condition for which MedFuse significantly outperforms both unimodal variants in calibration is *Other upper respiratory disease (20)* (MedFuse: 0.68 vs. CXR: 3.03, EHR: 0.98); however, this improvement does not translate into statistically superior AUROC, AUPRC, or selective prediction performance. Overall, these results indicate that multimodal fusion reliably improves discrimination but does not consistently improve, and can in some cases worsen, calibration. Detailed per-condition results for unimodal and multimodal models are reported in Table C.2 in Section C.1.

Class-Level Calibration Stratification. To better understand these effects, we decompose expected calibration error into positive- and negative-class components, $ECE_{c=1}$ and $ECE_{c=0}$, as defined in Equation (3). As shown in Table 2, positive-class calibration error is consistently larger than negative-class error across conditions, indicating that overconfidence in positive predictions is the dominant source of miscalibration. This imbalance is obscured when reporting only standard ECE (Table 1), which can therefore lead to misleading conclusions about model robustness.

We further observe that no single model dominates across all calibration components. EHR and MedFuse tend to achieve lower positive-class ECE than CXR, whereas CXR often exhibits the lowest negative-class ECE. This inversion illustrates how aggregate calibration metrics can mask complementary strengths and weaknesses across modalities, and why positive-class miscalibration is particularly detrimental to selective prediction in this task. These findings suggest that any observed improvements in positive-class calibration for the multimodal model are primarily inherited from the EHR modality rather than arising from multimodal fusion itself.

Finding 1

Multimodal fusion improves discrimination in multilabel clinical condition classification, but does not reliably improve calibration; overconfidence in underrepresented class is the dominant source of error.

Table 2: **Class-level Calibration Stratification.** Aggregate ECE (standard metric) appears low but obscures severe class-specific miscalibration. Across conditions and modality settings, positive-class ECE significantly dominates, revealing systematic overconfidence for patients who present the condition and poor calibration across modalities. (Dark-bold: Wilcoxon $p < 0.05$, 5 seeds; Light-bold: highest mean, not significant)

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

5.2. How Does Calibration Influence Selective Prediction Behavior?

Calibration Drives Selective Performance.

We first inspect selective prediction thresholds for all 25 conditions using MedFuse, plotting selective AUROC and AUPRC alongside stratified ECE (Figure 2.a). While selective metrics should monotonically improve as uncertain cases are rejected, we observe systematic departures from this expected behavior across many conditions. At extreme thresholds, selective AUROC and AUPRC collapse when the class distribution becomes highly imbalanced, and in mid-threshold regimes metrics often stagnate or degrade, suggesting 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 con-

trast, high ECE in the underrepresented class, corresponding to the positive class in this task, 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.

Motivated by these trends, we quantify the relationship between calibration and selective behavior across conditions (Figure 2.b). Positive-class ECE AUC and Selective AUC are consistently negatively correlated: higher $ECE_{c=1}$ predicts smaller or even negative changes in selective prediction metrics. The effect is strongest for AUPRC, while AUROC shows greater variability but follows the same direction. We test the null hypothesis of no association between ECE and Selective AUC and reject it for both metrics (Spearman’s rank correlation, $p < 0.05$).

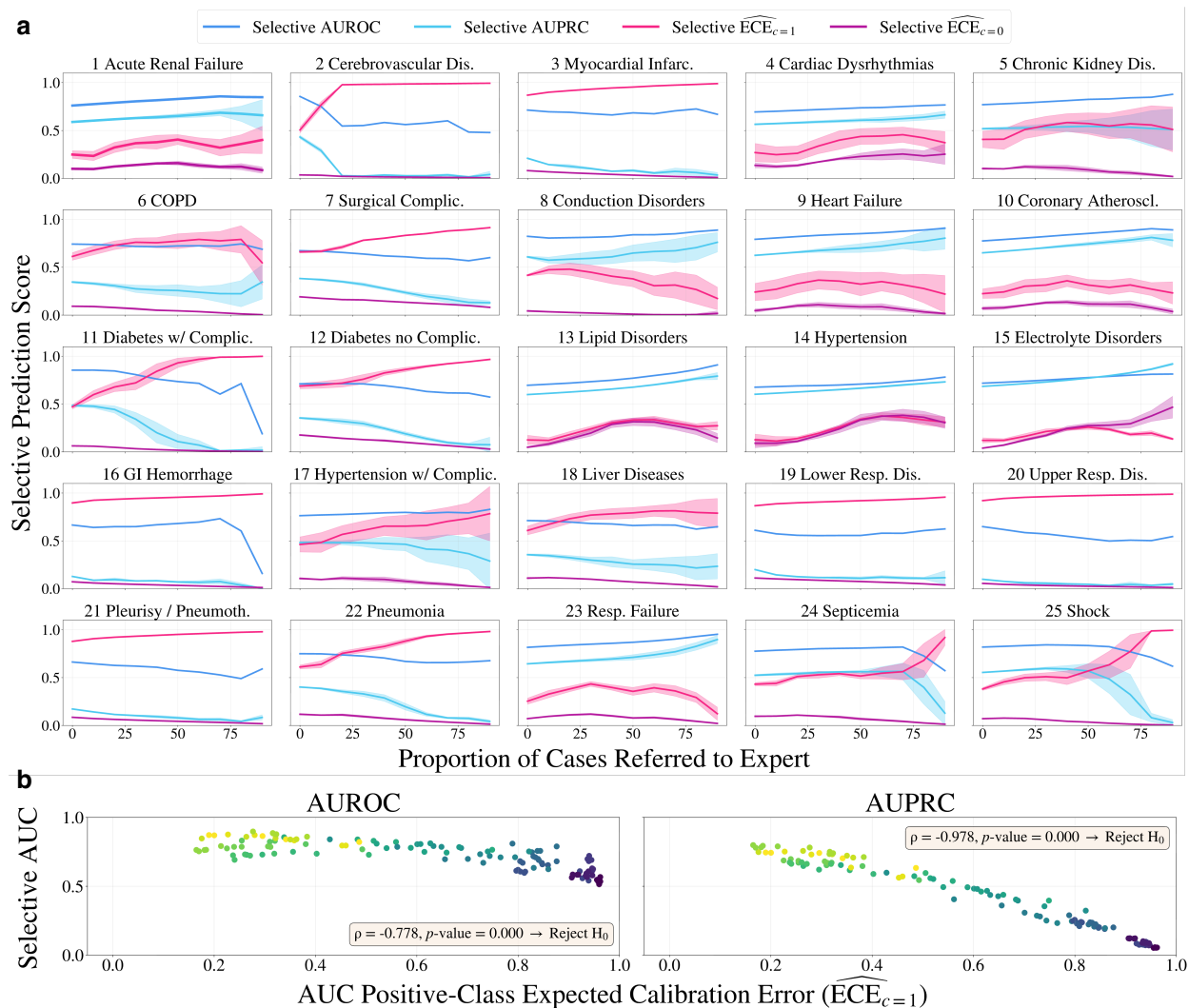


Figure 2: **How Does Calibration Influence Selective Prediction Behavior?** (a) Selective AUROC, selective AUPRC, and class-stratified ECE across 25 clinical conditions for MedFuse as low-confidence predictions are rejected. (b) Relationship between positive-class calibration error ($\widehat{ECE}_{c=1}$ AUC) and selective AUROC/AUPRC across conditions and random seeds, showing worse selective performance with higher $\widehat{ECE}_{c=1}$ (Spearman’s rank, $p < 0.05$. Shaded region: variance across seeds.)

Overall, these findings establish calibration of the underrepresented class as a reliable predictor of selective prediction stability. In this clinical task, miscalibration is dominated by the positive class due to class imbalance; however, high ECE in any minority class serves as an early warning that selective prediction will fail to yield meaningful improvements over standard evaluation.

Finding 2

Calibration of the underrepresented class in multilabel clinical condition classification is a leading indicator of selective reliability: when ECE is high for the minority class, selective AUROC/AUPRC fail to improve and can deteriorate.

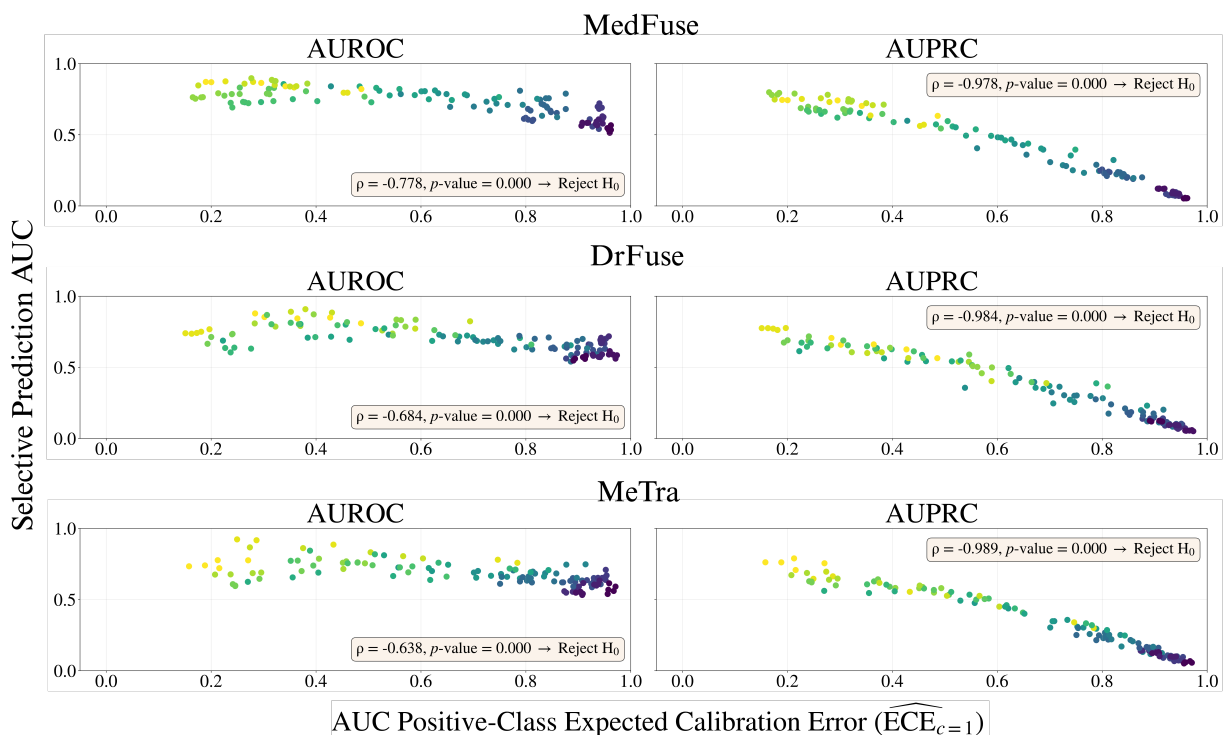


Figure 3: **Are Class-dependent Calibration Effects Consistent Across Multimodal Architectures?** Across MedFuse, DrFuse, and MeTRA, higher calibration error in the underrepresented class (i.e., positive class in this task) is consistently associated with lower selective AUROC and AUPRC across conditions. For all architectures and metrics, the null hypothesis of no association between stratified ECE and selective AUC is rejected, indicating that architectural complexity alone does not mitigate calibration-driven selective prediction failures. (Spearman’s rank H_0 , $p < 0.05$)

5.3. Are Class-Dependent Calibration Effects Consistent Across Multimodal Architectures?

Extending to DrFuse and MeTRA. To assess whether the observed relationship between calibration and selective prediction is specific to MedFuse or persists across architectures, we extend our analysis to DrFuse (Yao et al., 2024) and MeTRA (Khader et al., 2023), two multimodal models with substantially different fusion mechanisms.

To facilitate comparison, we adopt the same analysis relating class-dependent calibration to selective prediction performance, focusing on the underrepresented class, which corresponds to the positive class in this task. As shown in Figure 3, all three multimodal architectures exhibit a consistent downward trend for both AUROC and AUPRC: higher calibration error in the minority class reliably predicts degraded selective performance. This shared behavior indicates that the observed selective prediction fail-

ures are not an artifact of a specific fusion strategy or architectural complexity, but instead reflect a common consequence of class imbalance in this multilabel clinical condition classification setting.

Additional results supporting this comparison across architectures are provided in Section C.2, including aggregate discrimination and selective prediction metrics, class-stratified ECE tables, and per-condition selective prediction curves for DrFuse and MeTRA.

Finding 3

Across multiple multimodal fusion architectures, class-dependent miscalibration in the underrepresented class persists and is consistently associated with degraded selective prediction performance in multilabel clinical classification.

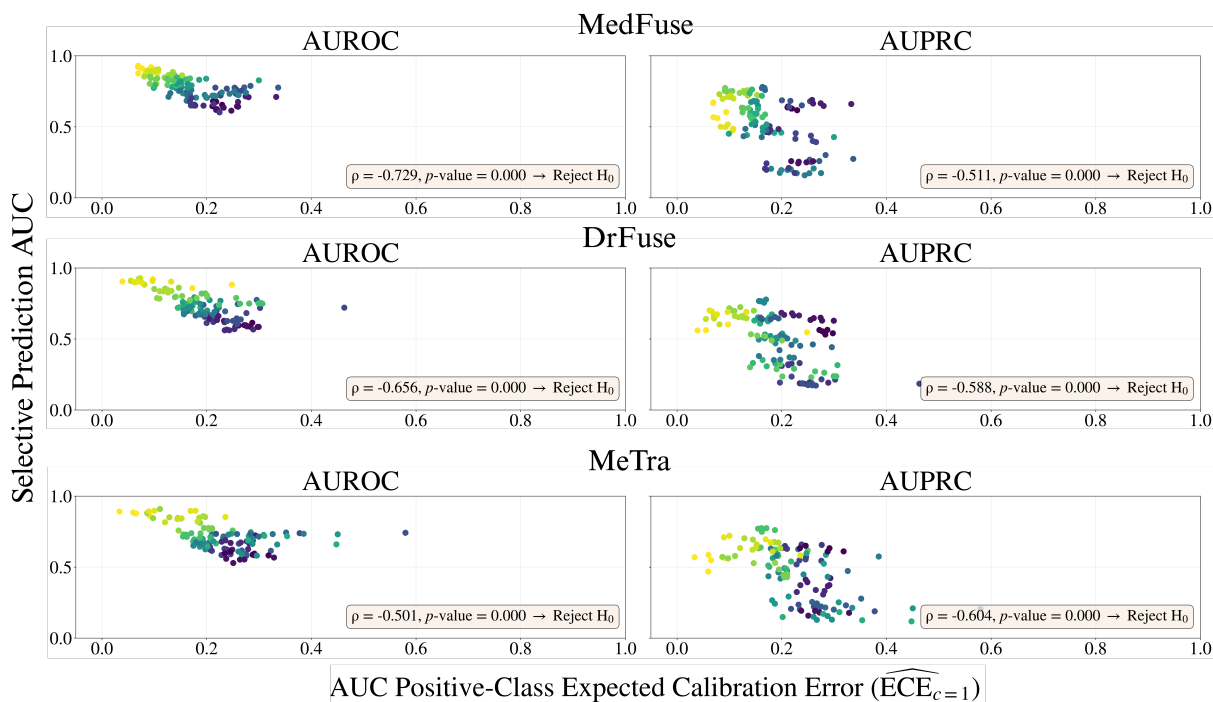


Figure 4: **Can Loss-Upweighting Mitigate Observed Calibration Failures?** Across architectures loss-upweighting reduces calibration error for the underrepresented class, but the negative relationship between stratified ECE and selective AUROC/AUPRC persists across conditions. For all architectures and metrics, the null hypothesis of no association between ECE and selective AUC is rejected (Spearman’s rank, $p < 0.05$).

5.4. Can Loss-Upweighting Mitigate Observed Calibration Failures?

Persistent Reliability Gaps. To assess whether class-dependent miscalibration can be mitigated, we apply a loss-upweighting strategy that emphasizes low-prevalence labels during training. This intervention is evaluated consistently across all multimodal architectures considered in this study, including MedFuse, DrFuse, and MeTra. As shown in Figure 4, loss upweighting systematically reduces calibration error for the underrepresented class across conditions and architectures, indicating decreased overconfidence for rare outcomes. However, selective prediction behavior remains largely unchanged.

Across models, selective AUROC and AUPRC exhibit only small average increases that are highly condition-dependent and rarely statistically significant. Moreover, the negative relationship between class-dependent calibration error and selective performance persists, indicating that improved calibration alone is insufficient to ensure reliable selective prediction in this setting.

Detailed per-condition and architecture-specific analyses are provided in Section C.3. In Section C.4, we further assess whether our findings are sensitive to the calibration metric or binning strategy by reporting results across alternative ECE bin counts, Adaptive ECE, and Brier score.

These robustness analyses show that the observed calibration-selective prediction misalignment persists across evaluation choices: while aggregate ECE shows no consistent global improvement, class-stratified metrics confirm that loss-upweighting primarily affects calibration of the underrepresented class, with weaker and less consistent effects for the majority class.

Finding 4

Loss upweighting for underrepresented labels yields measurable improvements in class-dependent calibration, but does not consistently translate into improved selective prediction reliability in multilabel clinical condition classification.

6. Discussion

This study shows that aggregate performance metrics can overstate reliability by masking condition-specific calibration and selective prediction failures in multi-label clinical condition classification. By moving beyond global averages, we reveal failure modes that are especially relevant in safety-critical settings, where errors affecting specific conditions or patient sub-populations may have disproportionate consequences. These findings underscore the need for calibration-aware evaluation when selective prediction is used as a safeguard in multimodal clinical models.

Why Can Selective Prediction Fail Despite Improved Discrimination? Our results suggest that improvements in aggregate discrimination do not necessarily translate into reliable uncertainty estimates. Although multimodal models can improve AUROC and AUPRC by combining complementary clinical data sources, their confidence estimates may remain poorly aligned with empirical correctness. Since abstention policies often rely on confidence or predictive uncertainty, class-dependent miscalibration can cause models to retain harder or under-represented cases while rejecting easier ones. Thus, selective prediction can fail not because discrimination is poor overall, but because the confidence scores used for abstention are unreliable for specific labels or classes.

Potential Drivers of Class-Dependent Miscalibration. Although our experiments do not establish a causal mechanism by which multimodal fusion induces miscalibration, the results suggest plausible contributors. Calibration failures are concentrated in the minority class and are most pronounced for lower-prevalence conditions, suggesting that label imbalance is a major driver. Moreover, multimodal fusion may improve representation quality and ranking performance while still producing poorly calibrated class-conditional probabilities. Simple mitigation strategies such as loss-upweighting can reduce positive-class calibration error, but may redistribute error toward the negative class, indicating that reweighting alone does not fully resolve class-dependent reliability failures.

Clinical Implications in Deployment. Minority-class overconfidence has important implications for clinical deployment. Low-prevalence findings may correspond to rare but clinically important conditions, where confident false predictions can

be especially harmful. If a model is overconfident on these cases, selective prediction may fail to abstain precisely when additional human review is needed. Consequently, strong aggregate performance may create a misleading impression of safety while reliability failures remain concentrated in clinically meaningful conditions. For clinical AI systems intended to support triage, monitoring, or decision support, evaluation should therefore characterize both class-conditional calibration and selective prediction behavior across clinically relevant labels.

Limitations & Future Work. This study has several limitations. First, the analysis is conducted on a single ICU benchmark, and it remains open whether similar patterns hold across other datasets, modality combinations, or clinical tasks. Second, ECE is sensitive to binning strategy: coarse bins may obscure localized miscalibration, while finer bins may increase estimator variance. We mitigate this concern through robustness analyses using multiple ECE bin counts, Adaptive ECE, and Brier score, which support the same qualitative calibration-selective prediction misalignment. Third, although loss-upweighting improves class-dependent calibration in some settings, it does not consistently improve selective prediction reliability across architectures. Finally, our analysis is retrospective and in-distribution; future work should examine out-of-distribution clinical settings, prospective validation, and interactions with clinical workflows and human-AI decision-making.

7. Conclusion

This work provides an empirical characterization of the interaction between calibration and selective prediction in multimodal multilabel clinical condition classification. We show that gains in discrimination from multimodal fusion do not reliably translate into improved selective prediction due to persistent class-dependent miscalibration. These findings highlight the need for evaluation and training approaches that explicitly account for calibration when selective prediction is used as a safety mechanism. More broadly, our results suggest that reliable deployment of multimodal clinical models requires moving beyond aggregate performance metrics toward task- and class-aware assessments of uncertainty. Future work should develop calibration-aware selective prediction methods that remain reliable across labels, prevalence regimes, and clinical deployment settings.

Acknowledgments

This work was supported by the NYUAD Center for Artificial Intelligence and Robotics, funded by Tamkeen under the NYUAD Research Institute Award CG010. The research was carried out on the High Performance Computing resources at New York University Abu Dhabi.

References

- Abdulmатеen Adebіyі, 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.
- 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.
- Beatrice Insalata, Fabian Schmidt, and Vladimir Vlassov. Multimodal survival prediction using tab-transformer and bioclinicalbert on mimic-iii. In *2024 IEEE International Conference on Big Data (BigData)*, pages 1986–1992. IEEE, 2024.
- 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 Bressemer, 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.
- Gary King and Langche Zeng. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

- Chih-Kuo Lee, Ting-Li Chen, Jeng-En Wu, Min-Tsun Liao, Chieh-hung 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, Chitralkha 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, Bireswar 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.
- Tim GJ Rudner, Ya Shi Zhang, Andrew Gordon Wilson, and Julia Kempe. Mind the gap: Improving robustness to subpopulation shifts with group-aware priors. In *International Conference on Artificial Intelligence and Statistics*, pages 127–135. PMLR, 2024.
- 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.
- Benjamin D Simon, Kutsev Bengisu Ozyoruk, David G Gelikman, Stephanie A Harmon, and Barış Türkbeş. The future of multimodal artificial intelligence models for integrating imaging and clinical metadata: a narrative review. *Diagnostic and Interventional Radiology*, 31(4):303, 2025.
- Yuanlong Wang, Changchang Yin, and Ping Zhang. Multimodal risk prediction with physiological signals, medical images and clinical notes. *Heliyon*, 10(5), 2024.
- 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.
- Haiyang Yang, Li Kuang, and FengQiang Xia. Multimodal temporal-clinical note network for mortality prediction. *Journal of Biomedical Semantics*, 12(1):3, 2021.
- 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.
- Yi Zheng, Fei Zhao, Xiaohua Liu, Min Liu, Ming Yang, Chengkai Zhang, and Xiangfeng Luo. A multimodal deep learning framework for predicting cardiovascular deterioration based on mimic-iv dataset. In *Proceedings of the 2025 6th International Symposium on Artificial Intelligence for Medical Sciences*, pages 837–842, 2025.

Appendix

Table of Contents

A Clinical Condition Dataset Distribution	809
B Training Details	810
C Extended Empirical Analysis	811
C.1 Do Multimodal Models Improve or Degrade Selective Prediction Performance in This Task?	811
C.2 Are Class-Dependent Calibration Effects Consistent Across Multimodal Architectures?	811
C.3 Can Loss-Upweighting Mitigate Observed Calibration Failures?	814
C.4 Extended Calibration Metrics and Binning Strategies	816
C.5 Stratified Calibration Behavior Across Modalities	823
D Per-Condition Comparative Analysis	829
D.1 Label 2: Acute Cerebrovascular Disease	829
D.2 Label 8: Conduction Disorders	830
D.3 Label 25: Respiratory Shock	831
E Regression Analysis of Calibration and Discriminative Performance	832

Reproducibility

We have made a significant effort to ensure the reproducibility of our results. An anonymized implementation of our method is provided at <https://github.com/jlaitue/medcertain>, which includes training, evaluation, and analysis scripts. The experimental setup, including hyperparameters, model configurations, and sampling parameters are described in Section B. 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 upon request, with instructions for reproducing results in the README.md file in the released code.

Appendix A. Clinical Condition Dataset Distribution

Table A.1: **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.

Clinical Condition	Organ	Modality		Train Set Prevalence	
		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)

Appendix B. Training Details

Hyperparameter Optimization and Model Selection

We conducted 50 experiments corresponding to randomly sampled hyperparameter sets, each 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. This procedure was applied to unimodal CXR, EHR, and MedFuse models, following the original MedFuse protocol for model selection based on AUROC.

For DrFuse, we used the authors' publicly released code and evaluation pipeline without modification, training and evaluating the model according to the specifications provided in their scripts. For MeTra, we adapted the authors' implementation originally designed for in-hospital mortality prediction by modifying only the final output layer to support clinical condition classification; all other training and optimization settings were kept consistent with the original implementation.

We report test performance as means and standard errors over five random seeds. All models were trained using the Adam optimizer. Experiments were executed on NVIDIA A100 and V100 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.

Appendix C. Extended Empirical Analysis

This section provides extended quantitative and qualitative evidence supporting the cross-architecture trends summarized in the main text. We report detailed evaluation metrics, class-stratified calibration statistics, and full selective prediction curves for all architectures.

C.1. Do Multimodal Models Improve or Degrade Selective Prediction Performance in This Task?

Table C.2: Do Multimodal Models Improve or Degrade Selective Prediction Performance in this Task? Condition-level performance of standard and selective prediction metrics. MedFuse generally improves AUROC, AUPRC and selective AUROC, but calibration gains are very 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	ECE ↓	AUROC	AUPRC	Selective AUROC	Selective AUPRC	ECE ↓	AUROC	AUPRC	Selective AUROC	Selective AUPRC	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±0.07	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±1.05
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±1.05	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±0.43	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±0.37	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

C.2. Are Class-Dependent Calibration Effects Consistent Across Multimodal Architectures?

Table C.3 reports aggregate discrimination and selective prediction metrics across all clinical conditions. While MedFuse often achieves the strongest discrimination performance, no single multimodal architecture consistently attains lower calibration error across conditions. Class-stratified ECE results in Table C.4 further show that increased architectural complexity does not resolve class-dependent miscalibration, particularly for underrepresented positive outcomes.

Additional stratified calibration plots and selective prediction analyses for unimodal and multimodal models (EHR, CXR, MedFuse, DrFuse, and MeTRA) are provided in Section C.5, offering a complete view of per-condition behavior across model families. Specifically, selective prediction curves for DrFuse and MeTRA are shown in Figures C.4 and C.5. Across conditions, these curves closely mirror the MedFuse behavior observed in the main analysis, with moderate gains at intermediate rejection thresholds followed by degradation at extreme coverage levels. This consistency indicates that the observed selective prediction failures are driven by shared calibration dynamics rather than architecture-specific fusion mechanisms.

Table C.3: **Condition-level Performance of Evaluation Metrics Across Multimodal Architectures.** MedFuse, DrFuse, and MeTra achieve broadly comparable discrimination performance across conditions, with MedFuse consistently attaining the highest AUROC/AUPRC values, while all models exhibit elevated calibration error. Shading and bolding are applied *within each model* to compare standard AUC metrics against their selective AUC counterparts for the same condition. Under correct uncertainty estimation, selective AUROC/AUPRC should be statistically significantly higher than standard metrics, reflecting effective rejection of uncertain predictions. Instead, we observe no consistent improvement across architectures: standard AUC metrics frequently dominate their selective counterparts, indicating that selective prediction fails to provide systematic gains. These results further suggest that increased architectural complexity alone does not improve reliability or resolve calibration-driven failures in this 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	$\widehat{ECE} \downarrow$	AUROC	AUPRC	Selective AUROC	Selective AUPRC	$\widehat{ECE} \downarrow$	AUROC	AUPRC	Selective AUROC	Selective AUPRC	$\widehat{ECE} \downarrow$
1	0.761 \pm 0.005	0.589 \pm 0.008	0.804 \pm 0.111	0.634 \pm 0.109	2.31 \pm 0.97	0.732 \pm 0.004	0.551 \pm 0.008	0.767 \pm 0.104	0.585 \pm 0.113	4.06 \pm 1.21	0.722 \pm 0.009	0.541 \pm 0.011	0.759 \pm 0.083	0.573 \pm 0.071	2.73 \pm 0.39
2	0.856 \pm 0.010	0.432 \pm 0.019	0.560 \pm 0.176	0.075 \pm 0.116	0.88 \pm 0.24	0.899 \pm 0.004	0.460 \pm 0.013	0.584 \pm 0.372	0.118 \pm 0.166	1.06 \pm 0.44	0.900 \pm 0.004	0.490 \pm 0.022	0.602 \pm 0.253	0.112 \pm 0.172	1.13 \pm 0.32
3	0.713 \pm 0.005	0.209 \pm 0.010	0.633 \pm 0.207	0.083 \pm 0.049	0.76 \pm 0.23	0.697 \pm 0.013	0.166 \pm 0.010	0.637 \pm 0.169	0.090 \pm 0.069	2.21 \pm 0.58	0.717 \pm 0.010	0.197 \pm 0.014	0.610 \pm 0.158	0.085 \pm 0.058	1.98 \pm 1.12
4	0.694 \pm 0.006	0.565 \pm 0.010	0.727 \pm 0.079	0.610 \pm 0.071	2.61 \pm 1.65	0.650 \pm 0.012	0.520 \pm 0.013	0.692 \pm 0.084	0.547 \pm 0.089	3.42 \pm 1.66	0.628 \pm 0.011	0.500 \pm 0.008	0.648 \pm 0.076	0.529 \pm 0.079	3.59 \pm 1.65
5	0.770 \pm 0.003	0.519 \pm 0.010	0.787 \pm 0.177	0.521 \pm 0.169	2.34 \pm 2.04	0.726 \pm 0.006	0.448 \pm 0.009	0.693 \pm 0.127	0.334 \pm 0.152	3.59 \pm 2.23	0.709 \pm 0.017	0.414 \pm 0.019	0.668 \pm 0.101	0.299 \pm 0.105	2.68 \pm 1.31
6	0.742 \pm 0.004	0.345 \pm 0.011	0.717 \pm 0.120	0.279 \pm 0.127	2.17 \pm 0.45	0.670 \pm 0.013	0.274 \pm 0.015	0.573 \pm 0.099	0.137 \pm 0.060	1.94 \pm 0.68	0.618 \pm 0.010	0.219 \pm 0.007	0.557 \pm 0.086	0.140 \pm 0.042	2.52 \pm 0.61
7	0.673 \pm 0.001	0.383 \pm 0.003	0.602 \pm 0.094	0.241 \pm 0.100	1.46 \pm 0.31	0.693 \pm 0.008	0.409 \pm 0.008	0.662 \pm 0.097	0.349 \pm 0.097	1.90 \pm 0.93	0.701 \pm 0.003	0.427 \pm 0.005	0.673 \pm 0.091	0.380 \pm 0.104	1.76 \pm 0.51
8	0.824 \pm 0.004	0.609 \pm 0.009	0.826 \pm 0.099	0.648 \pm 0.110	2.32 \pm 0.45	0.691 \pm 0.007	0.234 \pm 0.008	0.611 \pm 0.121	0.104 \pm 0.066	2.30 \pm 0.70	0.674 \pm 0.007	0.220 \pm 0.011	0.592 \pm 0.111	0.107 \pm 0.051	2.61 \pm 0.72
9	0.792 \pm 0.004	0.625 \pm 0.008	0.842 \pm 0.110	0.703 \pm 0.123	4.50 \pm 1.71	0.732 \pm 0.006	0.544 \pm 0.009	0.751 \pm 0.137	0.527 \pm 0.169	3.95 \pm 2.06	0.705 \pm 0.013	0.503 \pm 0.019	0.723 \pm 0.103	0.515 \pm 0.120	2.02 \pm 0.62
10	0.776 \pm 0.002	0.651 \pm 0.004	0.835 \pm 0.109	0.723 \pm 0.118	3.52 \pm 1.29	0.735 \pm 0.005	0.575 \pm 0.005	0.764 \pm 0.123	0.574 \pm 0.154	4.17 \pm 3.96	0.705 \pm 0.008	0.545 \pm 0.009	0.734 \pm 0.093	0.560 \pm 0.102	3.28 \pm 0.71
11	0.855 \pm 0.006	0.485 \pm 0.012	0.676 \pm 0.269	0.198 \pm 0.193	1.15 \pm 0.22	0.867 \pm 0.005	0.502 \pm 0.015	0.775 \pm 0.267	0.464 \pm 0.226	1.41 \pm 0.45	0.838 \pm 0.008	0.450 \pm 0.027	0.822 \pm 0.192	0.487 \pm 0.234	2.22 \pm 1.11
12	0.711 \pm 0.006	0.355 \pm 0.010	0.650 \pm 0.121	0.200 \pm 0.110	1.74 \pm 0.56	0.726 \pm 0.007	0.377 \pm 0.007	0.673 \pm 0.131	0.262 \pm 0.132	3.52 \pm 1.24	0.700 \pm 0.020	0.361 \pm 0.024	0.620 \pm 0.116	0.223 \pm 0.117	2.22 \pm 0.45
13	0.696 \pm 0.002	0.599 \pm 0.004	0.766 \pm 0.130	0.668 \pm 0.125	3.61 \pm 0.64	0.678 \pm 0.005	0.596 \pm 0.010	0.706 \pm 0.080	0.670 \pm 0.094	3.72 \pm 1.65	0.650 \pm 0.008	0.568 \pm 0.011	0.682 \pm 0.083	0.647 \pm 0.085	2.57 \pm 1.67
14	0.676 \pm 0.005	0.603 \pm 0.006	0.714 \pm 0.087	0.660 \pm 0.085	1.86 \pm 0.62	0.647 \pm 0.008	0.580 \pm 0.013	0.639 \pm 0.086	0.650 \pm 0.074	2.02 \pm 1.20	0.627 \pm 0.011	0.560 \pm 0.009	0.623 \pm 0.087	0.620 \pm 0.073	2.58 \pm 1.81
15	0.718 \pm 0.003	0.685 \pm 0.004	0.755 \pm 0.114	0.789 \pm 0.082	3.39 \pm 0.92	0.720 \pm 0.005	0.676 \pm 0.005	0.738 \pm 0.099	0.773 \pm 0.074	3.78 \pm 0.92	0.698 \pm 0.007	0.660 \pm 0.017	0.742 \pm 0.096	0.754 \pm 0.085	3.54 \pm 0.48
16	0.665 \pm 0.009	0.128 \pm 0.003	0.606 \pm 0.174	0.072 \pm 0.043	1.05 \pm 0.36	0.711 \pm 0.006	0.171 \pm 0.009	0.625 \pm 0.221	0.065 \pm 0.044	1.01 \pm 0.19	0.690 \pm 0.017	0.159 \pm 0.023	0.625 \pm 0.168	0.069 \pm 0.055	1.34 \pm 0.49
17	0.763 \pm 0.004	0.482 \pm 0.014	0.746 \pm 0.188	0.413 \pm 0.175	2.70 \pm 1.70	0.714 \pm 0.006	0.413 \pm 0.008	0.665 \pm 0.123	0.281 \pm 0.138	3.71 \pm 2.60	0.701 \pm 0.017	0.386 \pm 0.018	0.635 \pm 0.103	0.247 \pm 0.098	2.52 \pm 0.89
18	0.712 \pm 0.003	0.356 \pm 0.009	0.658 \pm 0.101	0.273 \pm 0.097	1.75 \pm 0.19	0.700 \pm 0.005	0.315 \pm 0.008	0.625 \pm 0.115	0.167 \pm 0.094	2.10 \pm 0.48	0.674 \pm 0.010	0.288 \pm 0.016	0.612 \pm 0.113	0.157 \pm 0.073	2.37 \pm 1.06
19	0.611 \pm 0.005	0.199 \pm 0.008	0.568 \pm 0.095	0.119 \pm 0.036	1.50 \pm 0.29	0.598 \pm 0.013	0.171 \pm 0.006	0.556 \pm 0.082	0.120 \pm 0.032	1.57 \pm 0.58	0.575 \pm 0.016	0.157 \pm 0.006	0.538 \pm 0.080	0.124 \pm 0.036	1.67 \pm 0.44
20	0.649 \pm 0.011	0.098 \pm 0.003	0.532 \pm 0.109	0.054 \pm 0.027	0.68 \pm 0.10	0.694 \pm 0.015	0.139 \pm 0.016	0.565 \pm 0.137	0.053 \pm 0.055	1.23 \pm 0.58	0.671 \pm 0.016	0.113 \pm 0.013	0.554 \pm 0.130	0.057 \pm 0.056	1.29 \pm 0.44
21	0.663 \pm 0.014	0.172 \pm 0.006	0.574 \pm 0.103	0.089 \pm 0.056	1.43 \pm 0.48	0.649 \pm 0.003	0.166 \pm 0.007	0.569 \pm 0.108	0.093 \pm 0.046	1.37 \pm 0.77	0.606 \pm 0.010	0.131 \pm 0.007	0.596 \pm 0.114	0.097 \pm 0.036	2.26 \pm 0.22
22	0.750 \pm 0.004	0.402 \pm 0.005	0.669 \pm 0.154	0.214 \pm 0.140	2.04 \pm 0.83	0.762 \pm 0.002	0.423 \pm 0.008	0.668 \pm 0.152	0.209 \pm 0.138	2.20 \pm 1.12	0.740 \pm 0.008	0.388 \pm 0.019	0.649 \pm 0.162	0.210 \pm 0.124	2.00 \pm 1.18
23	0.817 \pm 0.005	0.645 \pm 0.007	0.872 \pm 0.101	0.738 \pm 0.115	2.09 \pm 0.85	0.819 \pm 0.003	0.621 \pm 0.008	0.829 \pm 0.151	0.630 \pm 0.169	2.94 \pm 1.00	0.818 \pm 0.005	0.620 \pm 0.019	0.800 \pm 0.170	0.543 \pm 0.209	2.36 \pm 0.82
24	0.776 \pm 0.004	0.525 \pm 0.010	0.752 \pm 0.151	0.475 \pm 0.178	2.38 \pm 0.68	0.797 \pm 0.005	0.545 \pm 0.012	0.776 \pm 0.155	0.513 \pm 0.197	2.76 \pm 0.80	0.783 \pm 0.008	0.528 \pm 0.015	0.747 \pm 0.155	0.436 \pm 0.190	2.57 \pm 0.80
25	0.820 \pm 0.003	0.554 \pm 0.005	0.748 \pm 0.203	0.413 \pm 0.242	1.65 \pm 0.35	0.847 \pm 0.004	0.571 \pm 0.014	0.772 \pm 0.233	0.484 \pm 0.253	1.71 \pm 0.67	0.830 \pm 0.008	0.552 \pm 0.026	0.719 \pm 0.205	0.367 \pm 0.236	2.18 \pm 1.00
Average	0.739 \pm 0.005	0.449 \pm 0.008	0.705 \pm 0.136	0.396 \pm 0.112	2.08 \pm 0.73	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	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

Table C.4: **Condition-level Performance of ECE Stratification Across Multimodal Architectures.** Positive-class ECE dominates across conditions, highlighting systematic overconfidence in positive predictions across all multimodal architectures, independent of fusion complexity. Although MedFuse attains slightly lower mean ECE values, these differences are not consistently statistically significant across conditions. This indicates that no state-of-the-art multimodal architecture reliably resolves class-dependent calibration errors for 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		
	$\widehat{\text{ECE}} \downarrow$	$\widehat{\text{ECE}}_{c=1} \downarrow$	$\widehat{\text{ECE}}_{c=0} \downarrow$	$\widehat{\text{ECE}} \downarrow$	$\widehat{\text{ECE}}_{c=1} \downarrow$	$\widehat{\text{ECE}}_{c=0} \downarrow$	$\widehat{\text{ECE}} \downarrow$	$\widehat{\text{ECE}}_{c=1} \downarrow$	$\widehat{\text{ECE}}_{c=0} \downarrow$
1	2.31 ± 0.97	24.85 ± 3.96	9.97 ± 1.14	4.06 ± 1.21	27.74 ± 5.78	7.01 ± 2.40	2.73 ± 0.39	35.27 ± 2.04	13.65 ± 1.70
2	0.88 ± 0.24	50.56 ± 3.07	3.56 ± 0.19	1.06 ± 0.44	37.08 ± 7.82	2.98 ± 0.46	1.13 ± 0.32	33.53 ± 10.49	3.52 ± 0.15
3	0.76 ± 0.23	86.93 ± 0.14	8.05 ± 0.25	2.21 ± 0.58	88.15 ± 2.27	7.83 ± 1.71	1.98 ± 1.12	84.65 ± 2.99	9.32 ± 1.95
4	2.61 ± 1.65	26.90 ± 9.52	13.57 ± 2.63	3.42 ± 1.66	35.79 ± 13.68	17.54 ± 7.67	3.59 ± 1.65	39.75 ± 13.97	19.53 ± 5.75
5	2.34 ± 2.04	40.71 ± 8.56	10.21 ± 1.06	3.59 ± 2.23	49.62 ± 9.38	12.31 ± 1.96	2.68 ± 1.31	60.85 ± 4.41	16.17 ± 0.66
6	2.17 ± 0.45	61.34 ± 3.97	9.27 ± 0.68	1.94 ± 0.66	79.53 ± 1.91	13.44 ± 1.82	2.52 ± 0.61	80.43 ± 0.66	15.52 ± 1.17
7	1.46 ± 0.31	66.13 ± 1.23	19.12 ± 0.40	1.90 ± 0.93	65.82 ± 3.70	18.38 ± 1.68	1.76 ± 0.51	56.89 ± 6.12	16.28 ± 0.59
8	2.32 ± 0.45	41.64 ± 0.88	4.38 ± 0.89	2.30 ± 0.70	82.47 ± 3.45	9.75 ± 1.62	2.61 ± 0.72	82.51 ± 4.15	10.35 ± 1.64
9	4.50 ± 1.71	24.13 ± 9.07	4.69 ± 1.96	3.95 ± 2.06	34.55 ± 11.33	10.07 ± 2.26	2.02 ± 0.62	38.30 ± 9.77	14.65 ± 3.27
10	3.52 ± 1.29	22.44 ± 4.96	7.08 ± 1.88	4.17 ± 3.96	34.09 ± 13.14	11.16 ± 2.60	3.28 ± 0.71	37.68 ± 7.23	16.13 ± 1.71
11	1.15 ± 0.22	47.13 ± 2.20	6.30 ± 0.34	1.41 ± 0.45	35.95 ± 5.61	5.01 ± 0.89	2.22 ± 1.11	55.17 ± 11.94	6.25 ± 1.24
12	1.74 ± 0.56	68.84 ± 2.85	17.64 ± 0.79	3.52 ± 1.24	55.32 ± 11.18	16.14 ± 3.87	2.22 ± 0.45	62.38 ± 8.01	16.60 ± 1.96
13	3.61 ± 0.64	12.51 ± 4.83	4.88 ± 0.85	3.72 ± 1.65	17.27 ± 12.41	8.27 ± 5.89	2.57 ± 1.67	27.10 ± 8.21	15.39 ± 2.24
14	1.86 ± 0.62	12.64 ± 5.92	9.09 ± 4.73	2.02 ± 1.20	16.93 ± 11.53	12.38 ± 6.58	2.58 ± 1.81	14.50 ± 10.83	10.32 ± 7.64
15	3.39 ± 0.92	12.05 ± 2.66	3.95 ± 0.52	3.78 ± 0.92	10.59 ± 3.32	3.01 ± 1.87	3.54 ± 0.48	14.38 ± 6.59	7.02 ± 3.27
16	1.05 ± 0.36	89.64 ± 0.80	7.21 ± 0.42	1.01 ± 0.19	88.91 ± 0.80	6.84 ± 0.74	1.34 ± 0.49	88.77 ± 2.63	6.89 ± 0.98
17	2.70 ± 1.70	46.22 ± 7.91	10.70 ± 0.96	3.71 ± 2.60	54.74 ± 9.45	12.52 ± 2.57	2.52 ± 0.89	65.36 ± 3.14	15.65 ± 0.56
18	1.75 ± 0.19	60.84 ± 4.55	11.05 ± 0.92	2.10 ± 0.48	70.47 ± 3.99	14.39 ± 1.54	2.37 ± 1.06	78.01 ± 3.50	14.10 ± 1.39
19	1.50 ± 0.29	86.71 ± 0.43	11.15 ± 0.47	1.57 ± 0.58	86.99 ± 1.08	11.54 ± 1.14	1.67 ± 0.44	86.28 ± 0.82	12.37 ± 0.55
20	0.68 ± 0.10	92.06 ± 0.39	5.53 ± 0.33	1.23 ± 0.58	91.04 ± 0.93	4.63 ± 0.81	1.29 ± 0.44	90.78 ± 2.42	5.28 ± 1.12
21	1.43 ± 0.48	87.93 ± 0.48	8.56 ± 0.60	1.37 ± 0.77	89.23 ± 0.93	8.20 ± 0.81	2.26 ± 0.22	88.25 ± 1.49	9.52 ± 1.33
22	2.04 ± 0.83	61.13 ± 2.02	11.72 ± 0.84	2.20 ± 1.12	59.74 ± 7.21	11.26 ± 1.04	2.00 ± 1.18	62.64 ± 6.49	12.51 ± 1.24
23	2.09 ± 0.85	25.38 ± 3.17	7.17 ± 0.15	2.94 ± 1.00	22.12 ± 3.08	4.87 ± 0.58	2.36 ± 0.82	18.30 ± 5.56	5.82 ± 1.39
24	2.38 ± 0.68	43.16 ± 1.67	9.60 ± 0.84	2.76 ± 0.80	36.35 ± 4.61	7.43 ± 1.94	2.57 ± 0.80	45.99 ± 7.53	10.29 ± 1.44
25	1.65 ± 0.35	38.14 ± 1.48	7.12 ± 0.45	1.71 ± 0.67	28.90 ± 2.32	4.83 ± 1.20	2.18 ± 1.00	40.13 ± 7.60	7.04 ± 0.57
Average	2.08 ± 0.73	49.20 ± 3.47	8.86 ± 0.97	2.54 ± 1.15	51.98 ± 6.04	9.67 ± 2.23	2.32 ± 0.83	55.52 ± 5.94	11.61 ± 1.82

C.3. Can Loss-Upweighting Mitigate Observed Calibration Failures?

This section provides extended quantitative evidence supporting the loss-upweighting results summarized in the main text. We report per-condition class-stratified calibration statistics, and cross-architecture analyses to assess whether simple loss reweighting can mitigate calibration-driven selective prediction failures.

Cross-Architecture Consistency. Table C.5 reports the quantitative evaluation metrics for each multimodal architecture under loss upweighting, confirming that while selective AUROC/AUPRC exhibit modest gains, these improvements are inconsistent across conditions and do not fundamentally alter the observed reliability trends.

Calibration Effects. Using the standard aggregate ECE, we observe no meaningful global improvement under loss upweighting. In contrast, class-stratified analysis reveals statistically significant reductions in positive-class ECE for 23 of 25 conditions, while negative-class calibration improves in 11 of 25. These effects are quantified in Table C.6, which reports class-stratified calibration metrics for each architecture under loss upweighting. Together, these results confirm that loss reweighting reduces overconfidence for rare positive labels, but that these calibration gains do not reliably translate into improved selective prediction.

Table C.5: **Condition-level Performance of Evaluation Metrics Across Multimodal Architectures under Loss-Upweighting.** Although the performance of all three models is comparable, MedFuse outperforms both DrFuse and MeTra across different conditions and evaluation metrics, with the downside that all models present high ECE scores and the difference between models is not statistically significant. 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 (Loss-Upweighting)					DrFuse (Loss-Upweighting)					MeTra (Loss-Upweighting)				
	AUROC	AUPRC	Selective AUROC	Selective AUPRC	$\widehat{ECE} \downarrow$	AUROC	AUPRC	Selective AUROC	Selective AUPRC	$\widehat{ECE} \downarrow$	AUROC	AUPRC	Selective AUROC	Selective AUPRC	$\widehat{ECE} \downarrow$
1	0.761 \pm 0.003	0.588 \pm 0.008	0.756 \pm 0.088	0.672 \pm 0.065	3.09 \pm 0.86	0.731 \pm 0.004	0.547 \pm 0.007	0.717 \pm 0.090	0.628 \pm 0.081	4.58 \pm 0.98	0.715 \pm 0.008	0.533 \pm 0.012	0.721 \pm 0.088	0.618 \pm 0.077	3.35 \pm 1.48
2	0.871 \pm 0.006	0.431 \pm 0.010	0.872 \pm 0.093	0.491 \pm 0.069	2.38 \pm 0.66	0.895 \pm 0.005	0.473 \pm 0.010	0.878 \pm 0.153	0.547 \pm 0.135	4.44 \pm 1.50	0.897 \pm 0.007	0.478 \pm 0.034	0.871 \pm 0.111	0.546 \pm 0.078	3.79 \pm 0.79
3	0.716 \pm 0.003	0.205 \pm 0.005	0.744 \pm 0.167	0.260 \pm 0.097	9.58 \pm 2.12	0.693 \pm 0.007	0.175 \pm 0.007	0.687 \pm 0.118	0.205 \pm 0.073	13.34 \pm 6.55	0.699 \pm 0.027	0.182 \pm 0.012	0.694 \pm 0.143	0.205 \pm 0.070	10.09 \pm 3.99
4	0.692 \pm 0.005	0.563 \pm 0.007	0.650 \pm 0.084	0.631 \pm 0.060	2.89 \pm 1.26	0.621 \pm 0.005	0.492 \pm 0.005	0.589 \pm 0.073	0.550 \pm 0.066	4.58 \pm 2.03	0.630 \pm 0.009	0.496 \pm 0.010	0.595 \pm 0.073	0.543 \pm 0.068	1.93 \pm 0.65
5	0.760 \pm 0.006	0.496 \pm 0.012	0.805 \pm 0.094	0.599 \pm 0.113	4.60 \pm 1.52	0.718 \pm 0.002	0.437 \pm 0.006	0.726 \pm 0.085	0.522 \pm 0.088	5.90 \pm 2.72	0.696 \pm 0.015	0.402 \pm 0.013	0.678 \pm 0.095	0.467 \pm 0.080	4.98 \pm 3.28
6	0.747 \pm 0.007	0.348 \pm 0.009	0.803 \pm 0.099	0.444 \pm 0.102	4.78 \pm 2.64	0.653 \pm 0.005	0.259 \pm 0.004	0.619 \pm 0.076	0.320 \pm 0.089	4.36 \pm 2.76	0.615 \pm 0.009	0.206 \pm 0.007	0.590 \pm 0.072	0.223 \pm 0.032	3.57 \pm 3.35
7	0.670 \pm 0.003	0.390 \pm 0.009	0.645 \pm 0.074	0.476 \pm 0.095	5.90 \pm 1.22	0.690 \pm 0.003	0.409 \pm 0.004	0.669 \pm 0.074	0.512 \pm 0.106	2.46 \pm 1.70	0.699 \pm 0.012	0.416 \pm 0.018	0.683 \pm 0.078	0.505 \pm 0.093	2.96 \pm 0.57
8	0.819 \pm 0.009	0.584 \pm 0.014	0.860 \pm 0.122	0.734 \pm 0.303	4.61 \pm 0.84	0.673 \pm 0.007	0.233 \pm 0.009	0.720 \pm 0.090	0.325 \pm 0.115	8.29 \pm 2.82	0.674 \pm 0.005	0.210 \pm 0.012	0.684 \pm 0.093	0.250 \pm 0.064	6.13 \pm 3.48
9	0.789 \pm 0.007	0.615 \pm 0.007	0.840 \pm 0.101	0.727 \pm 0.121	5.97 \pm 2.62	0.722 \pm 0.004	0.533 \pm 0.011	0.754 \pm 0.090	0.648 \pm 0.108	8.07 \pm 2.26	0.690 \pm 0.017	0.481 \pm 0.025	0.708 \pm 0.086	0.571 \pm 0.094	4.83 \pm 3.15
10	0.773 \pm 0.005	0.641 \pm 0.009	0.809 \pm 0.093	0.732 \pm 0.100	5.86 \pm 2.01	0.725 \pm 0.007	0.566 \pm 0.007	0.689 \pm 0.102	0.644 \pm 0.066	4.17 \pm 3.15	0.696 \pm 0.012	0.535 \pm 0.015	0.681 \pm 0.093	0.608 \pm 0.073	2.16 \pm 1.18
11	0.861 \pm 0.006	0.484 \pm 0.026	0.901 \pm 0.105	0.607 \pm 0.126	4.62 \pm 0.72	0.867 \pm 0.003	0.506 \pm 0.004	0.904 \pm 0.097	0.639 \pm 0.120	6.14 \pm 2.96	0.821 \pm 0.017	0.426 \pm 0.012	0.833 \pm 0.109	0.553 \pm 0.117	6.68 \pm 2.72
12	0.713 \pm 0.008	0.367 \pm 0.013	0.720 \pm 0.107	0.413 \pm 0.091	6.83 \pm 1.26	0.722 \pm 0.004	0.379 \pm 0.005	0.712 \pm 0.101	0.444 \pm 0.098	7.58 \pm 4.02	0.690 \pm 0.012	0.349 \pm 0.017	0.651 \pm 0.095	0.390 \pm 0.064	9.57 \pm 3.54
13	0.689 \pm 0.004	0.590 \pm 0.008	0.733 \pm 0.123	0.655 \pm 0.113	5.94 \pm 0.46	0.674 \pm 0.004	0.592 \pm 0.006	0.639 \pm 0.081	0.672 \pm 0.073	3.74 \pm 3.10	0.650 \pm 0.004	0.563 \pm 0.006	0.631 \pm 0.083	0.644 \pm 0.076	3.22 \pm 1.88
14	0.666 \pm 0.009	0.600 \pm 0.007	0.684 \pm 0.085	0.659 \pm 0.094	2.25 \pm 0.77	0.641 \pm 0.005	0.573 \pm 0.005	0.585 \pm 0.072	0.637 \pm 0.066	2.47 \pm 1.43	0.624 \pm 0.004	0.561 \pm 0.009	0.588 \pm 0.077	0.625 \pm 0.069	2.47 \pm 0.35
15	0.714 \pm 0.002	0.676 \pm 0.004	0.727 \pm 0.086	0.772 \pm 0.078	3.22 \pm 0.62	0.721 \pm 0.005	0.676 \pm 0.007	0.698 \pm 0.100	0.767 \pm 0.068	2.17 \pm 1.03	0.696 \pm 0.008	0.663 \pm 0.009	0.712 \pm 0.080	0.766 \pm 0.083	3.05 \pm 0.91
16	0.656 \pm 0.010	0.128 \pm 0.005	0.731 \pm 0.095	0.169 \pm 0.050	7.21 \pm 3.68	0.725 \pm 0.009	0.183 \pm 0.013	0.790 \pm 0.125	0.271 \pm 0.135	9.68 \pm 3.80	0.667 \pm 0.008	0.151 \pm 0.013	0.713 \pm 0.115	0.226 \pm 0.130	14.73 \pm 6.17
17	0.752 \pm 0.007	0.461 \pm 0.010	0.803 \pm 0.105	0.563 \pm 0.117	5.09 \pm 1.61	0.707 \pm 0.003	0.404 \pm 0.010	0.720 \pm 0.085	0.491 \pm 0.093	6.41 \pm 2.93	0.687 \pm 0.016	0.374 \pm 0.015	0.667 \pm 0.093	0.441 \pm 0.077	5.00 \pm 3.28
18	0.717 \pm 0.007	0.366 \pm 0.020	0.722 \pm 0.090	0.467 \pm 0.103	5.71 \pm 2.67	0.686 \pm 0.006	0.302 \pm 0.004	0.681 \pm 0.083	0.363 \pm 0.070	5.24 \pm 2.51	0.671 \pm 0.009	0.283 \pm 0.011	0.660 \pm 0.075	0.333 \pm 0.070	7.99 \pm 3.32
19	0.612 \pm 0.005	0.198 \pm 0.004	0.623 \pm 0.067	0.251 \pm 0.065	3.72 \pm 1.10	0.606 \pm 0.005	0.173 \pm 0.003	0.560 \pm 0.078	0.187 \pm 0.023	3.36 \pm 2.03	0.564 \pm 0.013	0.150 \pm 0.006	0.547 \pm 0.062	0.179 \pm 0.051	9.74 \pm 4.91
20	0.699 \pm 0.010	0.144 \pm 0.007	0.707 \pm 0.079	0.200 \pm 0.086	4.44 \pm 1.45	0.689 \pm 0.011	0.159 \pm 0.009	0.741 \pm 0.088	0.260 \pm 0.136	5.23 \pm 2.60	0.654 \pm 0.030	0.109 \pm 0.010	0.676 \pm 0.107	0.148 \pm 0.074	6.48 \pm 3.84
21	0.699 \pm 0.009	0.178 \pm 0.013	0.653 \pm 0.088	0.205 \pm 0.083	6.53 \pm 3.43	0.638 \pm 0.010	0.154 \pm 0.002	0.622 \pm 0.078	0.178 \pm 0.031	4.69 \pm 1.83	0.596 \pm 0.013	0.129 \pm 0.009	0.615 \pm 0.089	0.154 \pm 0.077	7.05 \pm 1.97
22	0.763 \pm 0.004	0.415 \pm 0.010	0.788 \pm 0.093	0.502 \pm 0.096	3.89 \pm 0.89	0.752 \pm 0.004	0.417 \pm 0.012	0.786 \pm 0.086	0.507 \pm 0.094	3.10 \pm 1.37	0.732 \pm 0.004	0.373 \pm 0.008	0.757 \pm 0.117	0.440 \pm 0.095	2.57 \pm 0.98
23	0.818 \pm 0.006	0.652 \pm 0.006	0.826 \pm 0.099	0.757 \pm 0.075	2.19 \pm 0.77	0.818 \pm 0.003	0.616 \pm 0.006	0.820 \pm 0.103	0.708 \pm 0.072	2.36 \pm 1.04	0.810 \pm 0.009	0.602 \pm 0.020	0.834 \pm 0.124	0.663 \pm 0.110	3.16 \pm 1.60
24	0.774 \pm 0.005	0.510 \pm 0.010	0.815 \pm 0.090	0.621 \pm 0.103	5.60 \pm 0.95	0.792 \pm 0.003	0.534 \pm 0.007	0.829 \pm 0.091	0.656 \pm 0.099	4.51 \pm 0.86	0.777 \pm 0.007	0.517 \pm 0.014	0.823 \pm 0.092	0.632 \pm 0.098	2.02 \pm 0.35
25	0.826 \pm 0.004	0.556 \pm 0.004	0.886 \pm 0.115	0.698 \pm 0.142	5.09 \pm 0.55	0.845 \pm 0.002	0.557 \pm 0.008	0.888 \pm 0.120	0.679 \pm 0.133	5.20 \pm 1.25	0.826 \pm 0.004	0.548 \pm 0.019	0.867 \pm 0.145	0.661 \pm 0.160	2.99 \pm 1.34
Average	0.742 \pm 0.006	0.447 \pm 0.010	0.764 \pm 0.098	0.532 \pm 0.093	4.88 \pm 1.47	0.720 \pm 0.005	0.414 \pm 0.007	0.721 \pm 0.094	0.494 \pm 0.089	5.28 \pm 2.37	0.699 \pm 0.011	0.390 \pm 0.014	0.699 \pm 0.096	0.456 \pm 0.083	5.22 \pm 2.28

Table C.6: **Condition-level Performance of ECE Stratification Across Multimodal Architectures under Loss-Upweighting.** Class-stratified calibration metrics under loss upweighting show reduced positive-class ECE for many conditions, but it still remains dominant across architectures. No single multimodal model consistently achieves the lowest ECE across conditions, indicating that loss upweighting mitigates but does not resolve class-dependent calibration errors. (Dark-bold: $p < 0.05$, Wilcoxon signed-rank test, 5 seeds; Light-bold: highest mean, not significant)

Clinical Condition	MedFuse (Loss-Upweighting)			DrFuse (Loss-Upweighting)			MeTra (Loss-Upweighting)		
	$\widehat{\text{ECE}} \downarrow$	$\widehat{\text{ECE}}_{c=1} \downarrow$	$\widehat{\text{ECE}}_{c=0} \downarrow$	$\widehat{\text{ECE}} \downarrow$	$\widehat{\text{ECE}}_{c=1} \downarrow$	$\widehat{\text{ECE}}_{c=0} \downarrow$	$\widehat{\text{ECE}} \downarrow$	$\widehat{\text{ECE}}_{c=1} \downarrow$	$\widehat{\text{ECE}}_{c=0} \downarrow$
1	3.09 ± 0.86	3.81 ± 1.13	5.01 ± 1.96	4.58 ± 0.98	4.57 ± 0.84	6.32 ± 1.96	3.35 ± 1.48	7.19 ± 4.48	7.36 ± 2.53
2	2.38 ± 0.66	10.82 ± 1.46	2.10 ± 0.76	4.44 ± 1.50	9.75 ± 2.09	4.69 ± 1.66	3.79 ± 0.79	5.78 ± 1.66	4.41 ± 0.85
3	9.58 ± 2.12	12.14 ± 1.93	11.33 ± 2.29	13.34 ± 6.55	11.71 ± 2.50	15.62 ± 7.08	10.09 ± 3.99	18.20 ± 13.24	12.53 ± 5.48
4	2.89 ± 1.26	7.52 ± 2.45	5.32 ± 5.21	4.58 ± 2.03	7.31 ± 2.60	11.83 ± 4.69	1.93 ± 0.65	11.11 ± 5.16	8.07 ± 3.52
5	4.60 ± 1.52	5.05 ± 2.19	5.40 ± 2.97	5.90 ± 2.72	5.60 ± 2.70	8.21 ± 4.41	4.98 ± 3.28	12.78 ± 12.05	9.92 ± 7.48
6	4.78 ± 2.64	13.89 ± 7.07	5.96 ± 3.21	4.36 ± 2.76	4.41 ± 1.88	5.47 ± 3.56	3.57 ± 1.35	11.84 ± 3.42	5.74 ± 2.10
7	5.90 ± 1.22	9.79 ± 1.85	10.40 ± 1.49	2.46 ± 1.79	4.01 ± 1.00	3.26 ± 2.91	2.96 ± 0.57	8.72 ± 3.99	5.77 ± 2.06
8	4.61 ± 0.84	20.18 ± 4.92	6.43 ± 1.51	8.29 ± 2.82	9.40 ± 5.07	9.16 ± 2.14	6.13 ± 3.48	14.85 ± 11.31	8.36 ± 4.59
9	5.97 ± 2.62	6.67 ± 1.70	7.93 ± 5.47	8.07 ± 2.26	5.75 ± 1.75	12.77 ± 4.47	4.83 ± 3.15	11.04 ± 5.53	10.03 ± 4.27
10	5.86 ± 2.01	5.35 ± 3.38	7.46 ± 4.30	4.17 ± 3.15	6.17 ± 2.38	8.70 ± 5.61	2.16 ± 1.13	13.40 ± 14.01	9.35 ± 5.23
11	4.62 ± 0.72	4.86 ± 1.49	4.98 ± 0.82	6.14 ± 2.96	6.17 ± 1.61	7.17 ± 3.73	6.68 ± 2.72	8.75 ± 3.53	8.44 ± 3.56
12	6.83 ± 1.26	8.12 ± 2.16	10.47 ± 2.69	7.58 ± 4.02	11.77 ± 3.28	12.57 ± 5.78	9.57 ± 3.54	17.23 ± 2.96	16.67 ± 4.44
13	5.94 ± 0.46	7.16 ± 1.24	10.32 ± 3.75	3.74 ± 3.10	7.17 ± 3.51	10.62 ± 7.21	3.22 ± 1.88	14.21 ± 11.52	13.23 ± 6.32
14	2.25 ± 0.77	7.84 ± 3.34	6.50 ± 4.92	2.47 ± 1.43	9.74 ± 4.32	8.68 ± 7.05	2.47 ± 0.35	9.95 ± 2.70	11.95 ± 2.34
15	3.22 ± 0.62	4.22 ± 1.38	4.33 ± 1.55	2.17 ± 1.03	4.51 ± 2.29	3.34 ± 1.86	3.05 ± 0.91	5.77 ± 1.29	5.72 ± 1.70
16	7.21 ± 3.68	12.76 ± 6.14	7.65 ± 3.81	9.68 ± 3.80	7.86 ± 2.60	10.48 ± 4.14	14.73 ± 6.17	13.94 ± 3.39	16.54 ± 6.96
17	5.09 ± 1.61	6.16 ± 2.69	5.87 ± 3.00	6.41 ± 2.93	5.70 ± 2.92	8.56 ± 4.13	5.00 ± 3.25	11.68 ± 10.50	9.28 ± 6.46
18	5.71 ± 2.67	8.68 ± 3.09	6.80 ± 3.54	5.24 ± 2.51	6.50 ± 3.35	6.22 ± 3.20	7.99 ± 3.32	9.92 ± 4.03	11.24 ± 5.20
19	3.72 ± 1.10	6.00 ± 2.71	3.86 ± 1.39	3.36 ± 2.03	7.05 ± 2.53	4.39 ± 2.81	9.74 ± 4.91	12.10 ± 13.11	12.23 ± 7.18
20	4.44 ± 1.45	18.17 ± 1.09	5.18 ± 1.68	5.23 ± 2.60	24.25 ± 2.43	6.09 ± 3.30	6.48 ± 3.84	20.08 ± 8.64	7.84 ± 3.91
21	6.53 ± 3.43	7.57 ± 3.75	7.36 ± 4.02	4.69 ± 1.85	12.96 ± 2.23	5.69 ± 2.19	7.05 ± 1.97	19.47 ± 6.77	9.59 ± 2.77
22	3.89 ± 0.89	5.27 ± 1.06	5.56 ± 1.54	3.10 ± 1.37	5.57 ± 1.14	4.29 ± 1.37	2.57 ± 0.98	6.21 ± 0.90	3.95 ± 1.18
23	2.19 ± 0.77	6.05 ± 0.85	3.27 ± 1.91	2.36 ± 1.04	3.67 ± 0.81	3.32 ± 1.17	3.16 ± 1.69	7.09 ± 2.35	6.23 ± 3.19
24	5.60 ± 0.95	5.24 ± 0.89	6.70 ± 1.13	4.51 ± 0.86	5.25 ± 0.73	6.30 ± 1.23	2.02 ± 0.35	5.23 ± 1.63	2.87 ± 0.69
25	5.09 ± 0.55	5.63 ± 1.12	5.11 ± 0.80	5.20 ± 1.25	4.11 ± 1.67	5.83 ± 1.52	2.99 ± 1.34	6.58 ± 2.98	3.94 ± 1.55
Average	4.88 ± 1.47	8.36 ± 2.44	6.45 ± 2.63	5.28 ± 2.37	7.64 ± 2.33	7.58 ± 3.57	5.22 ± 2.28	11.32 ± 6.05	8.85 ± 3.82

C.4. Extended Calibration Metrics and Binning Strategies

This section provides additional robustness analyses for the calibration results reported in the main text. Since Expected Calibration Error (ECE) depends on the choice of confidence bins, we first evaluate whether the observed class-dependent miscalibration is sensitive to the number of bins used for estimation. We then complement ECE with Adaptive Expected Calibration Error (AECE) and Brier score, which provide alternative views of calibration quality. Across these analyses, the central trend remains consistent: calibration errors are substantially larger for the positive class than for the negative class, and this class-dependent miscalibration persists across unimodal and multimodal architectures.

C.4.1. VARIABLE BINNING FOR ECE

Binning-based ECE estimates involve an inherent bias–variance tradeoff. Using fewer bins can reduce estimator variance but may obscure localized miscalibration, while using more bins reveals sharper calibration failures at the cost of increased variance, particularly for low-prevalence labels. To assess whether our conclusions depend on a specific binning choice, we recompute aggregate and class-stratified ECE using $n \in \{5, 15, 20, 50\}$ bins (Table C.7, Table C.8).

Across bin counts, the same qualitative pattern is observed. Although aggregate ECE changes moderately with the number of bins, positive-class calibration error remains substantially higher than negative-class calibration error across model families. This indicates that the observed class-dependent miscalibration is not an artifact of a single ECE binning configuration.

Table C.7: **Robustness of ECE Estimates Across Binning Strategies for Unimodal Models.** Aggregate and class-stratified ECE are reported for unimodal baselines and MedFuse across different numbers of bins. Positive-class calibration error remains substantially higher than negative-class calibration error across binning choices, indicating that the observed class-dependent miscalibration is not driven by a single ECE bin configuration.

Bins	EHR			CXR			MedFuse		
	$\widehat{\text{ECE}} \downarrow$	$\widehat{\text{ECE}}_{c=1} \downarrow$	$\widehat{\text{ECE}}_{c=0} \downarrow$	$\widehat{\text{ECE}} \downarrow$	$\widehat{\text{ECE}}_{c=1} \downarrow$	$\widehat{\text{ECE}}_{c=0} \downarrow$	$\widehat{\text{ECE}} \downarrow$	$\widehat{\text{ECE}}_{c=1} \downarrow$	$\widehat{\text{ECE}}_{c=0} \downarrow$
5	1.07±0.30	60.35±1.08	14.26±0.48	9.26±0.88	54.64±5.64	5.68±1.75	1.85±0.78	49.19±3.46	8.85±0.98
10	1.42±0.34	60.35±1.08	14.27±0.48	9.41±0.83	54.67±5.63	5.74±1.71	2.08±0.73	49.20±3.47	8.86±0.97
15	1.64±0.32	60.70±1.14	14.39±0.44	9.55±0.81	54.97±5.65	5.87±1.67	2.30±0.71	49.48±3.45	8.94±0.95
20	1.84±0.40	60.81±1.11	14.42±0.45	9.59±0.83	54.95±5.66	5.93±1.68	2.41±0.70	49.39±3.46	8.97±0.95
50	2.81±0.42	61.27±1.14	14.61±0.45	10.01±0.77	55.59±5.63	6.39±1.54	3.26±0.62	49.83±3.45	9.17±0.89

Table C.8: **Robustness of ECE Estimates Across Binning Strategies for Multimodal Models.** Aggregate and class-stratified ECE are reported for multimodal architectures across different numbers of bins. Across MedFuse, DrFuse, and MeTra, positive-class ECE remains consistently larger than negative-class ECE, showing that class-dependent miscalibration persists across binning strategies and model families.

Bins	MedFuse			DrFuse			MeTra		
	$\widehat{\text{ECE}} \downarrow$	$\widehat{\text{ECE}}_{c=1} \downarrow$	$\widehat{\text{ECE}}_{c=0} \downarrow$	$\widehat{\text{ECE}} \downarrow$	$\widehat{\text{ECE}}_{c=1} \downarrow$	$\widehat{\text{ECE}}_{c=0} \downarrow$	$\widehat{\text{ECE}} \downarrow$	$\widehat{\text{ECE}}_{c=1} \downarrow$	$\widehat{\text{ECE}}_{c=0} \downarrow$
5	1.85±0.78	49.19±3.46	8.85±0.98	2.42±1.21	51.95±6.04	9.64±2.26	2.08±0.90	55.47±5.96	11.58±1.84
10	2.08±0.73	49.20±3.47	8.86±0.97	2.54±1.15	51.98±6.04	9.67±2.23	2.32±0.83	55.52±5.94	11.61±1.82
15	2.30±0.71	49.48±3.45	8.94±0.95	2.76±1.11	52.18±5.98	9.79±2.22	2.55±0.78	55.74±5.85	11.71±1.77
20	2.41±0.70	49.39±3.46	8.97±0.95	2.87±1.07	52.13±6.00	9.79±2.19	2.67±0.74	55.72±5.84	11.69±1.76
50	3.26±0.62	49.83±3.45	9.17±0.89	3.62±0.96	52.69±5.86	10.08±2.09	3.47±0.63	56.10±5.61	11.87±1.71

C.4.2. ADAPTIVE EXPECTED CALIBRATION ERROR

We further report Adaptive Expected Calibration Error (AECE), which uses adaptive bins with approximately equal sample counts rather than fixed-width confidence intervals. This provides a complementary estimate of calibration error that is less directly tied to the distribution of confidence scores across fixed bin boundaries. As with ECE, we report aggregate AECE as well as class-stratified AECE for the positive and negative classes for the unimodal, multimodal and loss-upweighted multimodal models.

Overall, the AECE results are consistent with the fixed-bin ECE analysis. For standard unimodal and multimodal models (Table C.9, Table C.10), positive-class AECE remains substantially higher than negative-class AECE, confirming that calibration failures are concentrated in the minority class. Under loss upweighting (Table C.11), positive-class AECE decreases, but this improvement equally accompanied by a redistribution of calibration error toward the negative class. Thus, while loss upweighting partially mitigates positive-class overconfidence, it does not eliminate class-dependent calibration imbalance.

Table C.9: **Adaptive ECE for Unimodal Models.** Aggregate and class-stratified Adaptive Expected Calibration Error (AECE) are reported for unimodal baselines. Positive-class AECE remains substantially higher than negative-class AECE, confirming that minority-class miscalibration persists when calibration error is estimated using adaptive equal-frequency bins. (Dark-bold: $p < 0.05$, Wilcoxon signed-rank test, 5 seeds; Light-bold: highest mean, not significant)

Clinical Condition	EHR			CXR			MedFuse		
	AECE \downarrow	AECE $_{c=1}\downarrow$	AECE $_{c=0}\downarrow$	AECE \downarrow	AECE $_{c=1}\downarrow$	AECE $_{c=0}\downarrow$	AECE \downarrow	AECE $_{c=1}\downarrow$	AECE $_{c=0}\downarrow$
1	2.32 ± 0.48	57.80 ± 0.36	27.38 ± 0.09	14.39 ± 1.15	56.87 ± 3.22	29.87 ± 3.00	3.22 ± 0.91	54.09 ± 2.26	25.83 ± 2.03
2	1.02 ± 0.21	66.59 ± 0.37	6.23 ± 0.14	5.87 ± 0.47	86.92 ± 2.62	6.84 ± 1.62	1.29 ± 0.26	67.96 ± 1.40	5.40 ± 0.18
3	1.37 ± 0.12	88.27 ± 0.33	8.22 ± 0.06	6.75 ± 1.33	85.30 ± 3.50	8.38 ± 2.50	1.46 ± 0.30	86.93 ± 0.14	8.05 ± 0.25
4	2.37 ± 0.50	60.63 ± 0.13	36.44 ± 0.12	14.13 ± 1.07	52.81 ± 4.49	32.66 ± 4.52	3.93 ± 1.67	55.81 ± 3.90	32.44 ± 3.51
5	2.65 ± 0.37	70.95 ± 0.28	21.08 ± 0.07	17.08 ± 0.70	59.96 ± 1.82	22.65 ± 1.99	3.64 ± 1.62	61.94 ± 4.35	18.95 ± 3.07
6	2.58 ± 0.14	81.50 ± 0.22	16.00 ± 0.07	8.93 ± 1.03	74.71 ± 9.10	11.19 ± 4.40	3.30 ± 0.47	72.08 ± 2.20	13.20 ± 1.03
7	1.99 ± 0.25	73.47 ± 0.43	20.83 ± 0.08	11.31 ± 0.89	75.57 ± 2.16	17.34 ± 1.60	2.10 ± 0.35	72.23 ± 0.49	20.56 ± 0.37
8	1.45 ± 0.24	87.20 ± 0.28	10.42 ± 0.04	4.95 ± 1.48	52.20 ± 7.29	6.55 ± 3.01	2.85 ± 0.49	54.39 ± 3.81	6.16 ± 1.40
9	3.22 ± 0.52	61.89 ± 0.34	27.76 ± 0.06	13.63 ± 3.29	41.85 ± 5.80	28.29 ± 5.00	5.63 ± 1.61	50.90 ± 5.67	21.83 ± 4.31
10	2.97 ± 0.25	62.15 ± 0.19	27.59 ± 0.12	12.29 ± 1.25	50.81 ± 4.00	24.76 ± 4.45	4.60 ± 1.31	50.67 ± 3.65	24.37 ± 3.03
11	2.41 ± 0.61	80.37 ± 1.11	10.92 ± 0.15	9.29 ± 0.68	82.04 ± 4.22	11.91 ± 2.77	1.44 ± 0.33	65.92 ± 1.70	9.03 ± 0.38
12	2.05 ± 0.62	77.14 ± 0.38	19.78 ± 0.07	12.72 ± 1.51	73.26 ± 4.44	21.51 ± 4.13	2.63 ± 0.40	71.76 ± 1.61	18.97 ± 1.05
13	2.70 ± 0.27	54.87 ± 0.23	38.19 ± 0.07	16.87 ± 0.76	42.87 ± 6.30	43.89 ± 6.23	4.51 ± 0.75	49.43 ± 2.11	36.43 ± 2.12
14	3.42 ± 0.36	54.28 ± 0.39	43.48 ± 0.08	19.16 ± 1.41	49.45 ± 5.90	39.29 ± 6.64	3.75 ± 0.81	50.60 ± 2.22	39.56 ± 2.06
15	2.37 ± 0.34	47.74 ± 0.15	39.49 ± 0.24	16.95 ± 1.31	52.02 ± 1.91	38.58 ± 1.96	3.67 ± 0.92	46.28 ± 1.19	36.83 ± 1.55
16	1.69 ± 0.31	91.64 ± 0.15	6.89 ± 0.09	5.59 ± 0.76	89.56 ± 2.11	6.79 ± 1.40	1.73 ± 0.34	89.64 ± 0.80	7.21 ± 0.42
17	2.24 ± 0.33	73.82 ± 0.26	19.26 ± 0.09	15.74 ± 0.32	64.14 ± 0.99	19.92 ± 0.97	3.86 ± 1.29	64.76 ± 4.25	17.59 ± 2.93
18	1.92 ± 0.37	80.92 ± 0.15	15.41 ± 0.14	9.63 ± 1.13	70.76 ± 3.99	14.27 ± 3.05	3.08 ± 0.99	72.07 ± 2.49	15.49 ± 0.95
19	1.44 ± 0.32	86.08 ± 0.27	12.65 ± 0.12	9.99 ± 1.09	83.60 ± 3.35	13.59 ± 2.87	2.11 ± 0.15	86.71 ± 0.43	11.15 ± 0.47
20	1.68 ± 0.33	91.94 ± 0.23	6.23 ± 0.09	4.52 ± 0.68	89.21 ± 1.18	4.93 ± 1.02	1.34 ± 0.28	92.06 ± 0.39	5.53 ± 0.33
21	2.37 ± 0.54	88.61 ± 0.18	9.83 ± 0.14	6.52 ± 0.41	86.20 ± 3.27	6.45 ± 2.00	1.81 ± 0.17	87.93 ± 0.48	8.56 ± 0.60
22	1.97 ± 0.22	72.53 ± 0.21	16.97 ± 0.13	10.18 ± 0.78	75.37 ± 4.74	14.55 ± 3.48	2.22 ± 0.65	71.68 ± 1.20	14.94 ± 1.05
23	2.13 ± 0.40	54.41 ± 0.79	21.87 ± 0.39	11.90 ± 0.92	61.64 ± 5.01	18.32 ± 3.81	2.56 ± 0.92	53.61 ± 1.61	18.67 ± 1.00
24	2.07 ± 0.26	65.40 ± 0.70	18.50 ± 0.22	12.62 ± 0.80	71.73 ± 4.30	16.52 ± 2.98	3.01 ± 0.48	62.07 ± 1.01	17.19 ± 1.02
25	2.04 ± 0.44	63.56 ± 0.95	14.04 ± 0.27	10.41 ± 1.08	72.14 ± 3.23	12.75 ± 1.89	2.67 ± 0.29	59.90 ± 1.37	12.83 ± 0.63
Average	2.18 ± 0.35	71.75 ± 0.36	19.82 ± 0.13	11.26 ± 1.05	68.04 ± 3.96	18.87 ± 3.09	2.90 ± 0.71	66.06 ± 2.03	17.87 ± 1.43

Table C.10: **Adaptive ECE for Multimodal Models.** Aggregate and class-stratified Adaptive Expected Calibration Error (AECE) are reported for standard multimodal architectures. The positive class exhibits substantially higher AECE than the negative class across MedFuse, DrFuse, and MeTra, reinforcing that multimodal performance gains do not eliminate class-dependent calibration errors. (Dark-bold: $p < 0.05$, Wilcoxon signed-rank test, 5 seeds; Light-bold: highest mean, not significant)

Clinical Condition	MedFuse			DrFuse			MeTra		
	$\widehat{\text{AECE}} \downarrow$	$\widehat{\text{AECE}}_{c=1} \downarrow$	$\widehat{\text{AECE}}_{c=0} \downarrow$	$\widehat{\text{AECE}} \downarrow$	$\widehat{\text{AECE}}_{c=1} \downarrow$	$\widehat{\text{AECE}}_{c=0} \downarrow$	$\widehat{\text{AECE}} \downarrow$	$\widehat{\text{AECE}}_{c=1} \downarrow$	$\widehat{\text{AECE}}_{c=0} \downarrow$
1	3.22 \pm 0.91	54.09 \pm 2.26	25.83 \pm 2.03	5.16 \pm 1.20	55.57 \pm 3.44	26.30 \pm 2.71	3.01 \pm 0.39	59.40 \pm 1.02	26.08 \pm 1.02
2	1.29 \pm 0.26	67.96 \pm 1.40	5.40 \pm 0.18	1.92 \pm 1.05	60.49 \pm 4.83	6.21 \pm 1.11	2.02 \pm 0.62	59.96 \pm 5.55	6.44 \pm 1.01
3	1.46 \pm 0.30	86.93 \pm 0.14	8.05 \pm 0.25	2.37 \pm 0.35	88.15 \pm 2.27	7.83 \pm 1.71	2.11 \pm 1.13	84.75 \pm 2.92	9.34 \pm 1.95
4	3.93 \pm 1.67	55.81 \pm 3.90	32.44 \pm 3.51	4.82 \pm 2.29	58.41 \pm 5.60	34.07 \pm 4.56	5.16 \pm 1.72	59.83 \pm 4.55	34.46 \pm 4.44
5	3.64 \pm 1.62	61.94 \pm 4.35	18.95 \pm 3.07	4.64 \pm 2.03	66.72 \pm 4.71	19.40 \pm 3.56	3.36 \pm 1.01	70.00 \pm 1.94	20.06 \pm 1.60
6	3.30 \pm 0.47	72.08 \pm 2.20	13.20 \pm 1.03	2.09 \pm 0.41	80.69 \pm 1.41	13.62 \pm 1.68	3.04 \pm 0.72	80.71 \pm 0.80	15.64 \pm 1.15
7	2.10 \pm 0.35	72.23 \pm 0.49	20.56 \pm 0.37	2.59 \pm 0.81	71.65 \pm 2.04	19.77 \pm 1.74	2.65 \pm 0.56	68.49 \pm 2.80	20.47 \pm 1.49
8	2.85 \pm 0.49	54.39 \pm 3.81	6.16 \pm 1.40	2.95 \pm 0.72	83.42 \pm 2.79	9.94 \pm 1.79	2.86 \pm 0.62	83.64 \pm 2.86	10.67 \pm 1.87
9	5.63 \pm 1.61	50.90 \pm 5.67	21.83 \pm 4.31	5.96 \pm 1.98	59.53 \pm 5.53	23.29 \pm 3.96	3.77 \pm 2.08	60.69 \pm 3.90	26.55 \pm 3.21
10	4.60 \pm 1.31	50.67 \pm 3.65	24.37 \pm 3.03	5.60 \pm 4.01	58.56 \pm 6.49	25.02 \pm 4.83	3.66 \pm 0.54	60.09 \pm 2.47	27.23 \pm 1.50
11	1.44 \pm 0.33	65.92 \pm 1.70	9.03 \pm 0.38	3.56 \pm 1.41	56.76 \pm 3.60	10.96 \pm 1.96	3.97 \pm 1.48	66.96 \pm 7.52	9.91 \pm 3.44
12	2.63 \pm 0.40	71.76 \pm 1.61	18.97 \pm 1.05	5.34 \pm 1.63	66.72 \pm 4.76	21.70 \pm 4.08	3.05 \pm 0.36	70.49 \pm 2.58	19.90 \pm 1.71
13	4.51 \pm 0.75	49.43 \pm 2.11	36.43 \pm 2.12	6.29 \pm 2.27	51.18 \pm 5.76	37.77 \pm 5.41	4.51 \pm 1.15	54.64 \pm 3.47	37.09 \pm 3.68
14	3.75 \pm 0.81	50.60 \pm 2.22	39.56 \pm 2.06	4.26 \pm 1.22	52.29 \pm 3.74	41.47 \pm 3.68	5.42 \pm 1.12	52.00 \pm 3.93	42.60 \pm 4.47
15	3.67 \pm 0.92	46.28 \pm 1.19	36.83 \pm 1.55	4.52 \pm 0.97	45.05 \pm 1.60	37.45 \pm 1.67	4.54 \pm 1.00	48.14 \pm 2.66	37.85 \pm 2.71
16	1.73 \pm 0.34	89.64 \pm 0.80	7.21 \pm 0.42	1.41 \pm 0.13	88.91 \pm 0.80	6.84 \pm 0.74	1.58 \pm 0.32	89.01 \pm 2.17	6.91 \pm 0.99
17	3.86 \pm 1.29	64.76 \pm 4.25	17.59 \pm 2.93	5.09 \pm 2.14	69.27 \pm 5.22	18.16 \pm 3.98	2.79 \pm 0.72	72.40 \pm 1.46	18.41 \pm 1.27
18	3.08 \pm 0.99	72.07 \pm 2.49	15.49 \pm 0.95	2.85 \pm 0.60	75.23 \pm 2.04	15.81 \pm 1.36	3.03 \pm 0.75	79.54 \pm 2.29	14.56 \pm 1.57
19	2.11 \pm 0.15	86.71 \pm 0.43	11.15 \pm 0.47	2.09 \pm 0.34	86.99 \pm 1.08	11.54 \pm 1.14	2.38 \pm 0.61	86.28 \pm 0.82	12.37 \pm 0.55
20	1.34 \pm 0.28	92.06 \pm 0.39	5.53 \pm 0.33	1.36 \pm 0.27	91.18 \pm 0.85	4.63 \pm 0.81	1.77 \pm 0.78	91.13 \pm 2.14	5.37 \pm 1.20
21	1.81 \pm 0.17	87.93 \pm 0.48	8.56 \pm 0.60	1.79 \pm 0.57	89.23 \pm 0.93	8.20 \pm 0.81	2.69 \pm 0.33	88.34 \pm 1.42	9.54 \pm 1.33
22	2.22 \pm 0.65	71.68 \pm 1.20	14.94 \pm 1.05	2.70 \pm 0.92	71.49 \pm 2.65	14.08 \pm 1.39	2.60 \pm 0.98	72.13 \pm 2.90	15.42 \pm 1.77
23	2.56 \pm 0.92	53.61 \pm 1.61	18.67 \pm 1.00	3.39 \pm 0.88	51.35 \pm 2.24	19.06 \pm 1.41	3.45 \pm 1.10	50.94 \pm 4.49	20.52 \pm 2.38
24	3.01 \pm 0.48	62.07 \pm 1.01	17.19 \pm 1.02	3.14 \pm 1.03	58.51 \pm 3.24	17.23 \pm 1.22	3.76 \pm 0.83	63.73 \pm 4.49	16.19 \pm 1.87
25	2.67 \pm 0.29	59.90 \pm 1.37	12.83 \pm 0.63	3.39 \pm 0.75	54.13 \pm 2.21	12.96 \pm 0.46	3.28 \pm 0.86	61.72 \pm 4.80	12.07 \pm 2.19
Average	2.90 \pm 0.71	66.06 \pm 2.03	17.87 \pm 1.43	3.57 \pm 1.20	67.66 \pm 3.19	18.53 \pm 2.31	3.22 \pm 0.87	69.40 \pm 3.04	19.03 \pm 2.01

Table C.11: **Adaptive ECE for Multimodal Models under Loss-Upweighting.** Aggregate and class-stratified Adaptive Expected Calibration Error (AECE) are reported for multimodal architectures trained with loss upweighting. Loss upweighting reduces positive-class AECE relative to standard multimodal models, but it redistributes calibration error toward the negative class, indicating that this mitigation does not fully resolve class-dependent miscalibration. (Dark-bold: $p < 0.05$, Wilcoxon signed-rank test, 5 seeds; Light-bold: highest mean, not significant)

Clinical Condition	MedFuse (Loss-Upweighting)			DrFuse (Loss-Upweighting)			MeTra (Loss-Upweighting)		
	AECE \downarrow	AECE _{c=1} \downarrow	AECE _{c=0} \downarrow	AECE \downarrow	AECE _{c=1} \downarrow	AECE _{c=0} \downarrow	AECE \downarrow	AECE _{c=1} \downarrow	AECE _{c=0} \downarrow
1	14.47 \pm 1.33	38.49 \pm 1.01	39.54 \pm 1.49	14.91 \pm 1.51	40.19 \pm 1.19	40.42 \pm 1.72	13.67 \pm 3.13	43.47 \pm 3.99	40.70 \pm 2.76
2	15.85 \pm 1.07	28.20 \pm 1.19	19.61 \pm 1.08	15.46 \pm 3.11	26.76 \pm 3.93	18.51 \pm 3.20	21.47 \pm 3.66	22.96 \pm 1.76	25.26 \pm 3.93
3	30.73 \pm 2.15	43.34 \pm 2.27	38.31 \pm 2.15	35.58 \pm 5.53	41.28 \pm 5.99	43.24 \pm 5.51	31.64 \pm 7.15	46.78 \pm 8.88	39.66 \pm 6.98
4	11.02 \pm 2.99	43.65 \pm 2.86	44.38 \pm 3.15	12.31 \pm 1.80	46.40 \pm 1.75	48.18 \pm 1.84	9.66 \pm 2.90	48.86 \pm 3.10	45.35 \pm 2.94
5	19.10 \pm 2.90	39.06 \pm 2.45	37.57 \pm 3.07	22.74 \pm 3.28	41.06 \pm 3.14	42.39 \pm 3.37	21.17 \pm 5.49	45.53 \pm 6.87	42.35 \pm 5.08
6	22.38 \pm 5.69	43.41 \pm 6.81	33.87 \pm 5.52	32.35 \pm 1.77	45.29 \pm 1.73	46.18 \pm 1.78	29.49 \pm 1.48	50.92 \pm 1.53	43.53 \pm 1.50
7	26.14 \pm 0.83	43.50 \pm 0.79	46.45 \pm 0.84	26.42 \pm 1.30	43.32 \pm 1.62	45.61 \pm 1.26	22.74 \pm 3.04	45.18 \pm 3.38	42.55 \pm 3.00
8	19.58 \pm 4.81	32.96 \pm 4.22	26.28 \pm 5.02	32.57 \pm 3.77	44.21 \pm 4.29	42.06 \pm 3.71	29.97 \pm 5.85	48.38 \pm 6.71	40.15 \pm 5.74
9	14.09 \pm 4.56	36.02 \pm 4.54	34.69 \pm 4.87	19.13 \pm 2.87	38.66 \pm 2.72	43.25 \pm 2.94	15.60 \pm 4.46	46.09 \pm 5.79	41.32 \pm 4.08
10	12.76 \pm 3.80	36.72 \pm 4.26	36.87 \pm 4.31	16.44 \pm 4.15	39.35 \pm 3.87	44.52 \pm 4.30	11.47 \pm 5.38	47.35 \pm 6.92	40.48 \pm 5.98
11	19.46 \pm 1.80	30.43 \pm 1.15	26.26 \pm 1.90	22.79 \pm 4.17	26.38 \pm 3.88	29.41 \pm 4.23	26.63 \pm 6.84	32.92 \pm 6.52	34.76 \pm 7.06
12	23.30 \pm 1.27	42.68 \pm 1.36	40.99 \pm 1.29	26.10 \pm 4.15	40.14 \pm 4.34	43.63 \pm 4.11	26.53 \pm 5.90	42.61 \pm 4.93	45.07 \pm 6.25
13	9.12 \pm 2.37	43.34 \pm 3.42	42.34 \pm 3.63	10.02 \pm 4.06	43.05 \pm 4.03	47.20 \pm 4.09	8.85 \pm 4.11	46.86 \pm 5.58	45.61 \pm 5.10
14	6.64 \pm 2.80	46.92 \pm 3.72	44.29 \pm 3.80	8.55 \pm 1.99	45.22 \pm 2.00	49.87 \pm 2.09	8.30 \pm 1.89	46.71 \pm 2.69	49.15 \pm 2.71
15	4.72 \pm 0.70	42.31 \pm 1.07	41.49 \pm 1.21	5.10 \pm 1.56	42.00 \pm 2.10	41.85 \pm 2.15	4.61 \pm 0.89	43.61 \pm 1.45	42.75 \pm 0.80
16	33.39 \pm 2.00	48.08 \pm 2.12	39.63 \pm 2.01	33.64 \pm 2.55	41.54 \pm 3.60	39.19 \pm 2.49	37.05 \pm 4.10	44.48 \pm 3.11	43.29 \pm 4.18
17	20.11 \pm 3.19	39.77 \pm 2.72	37.21 \pm 3.34	24.27 \pm 3.37	41.72 \pm 3.31	42.53 \pm 3.41	23.11 \pm 5.01	45.61 \pm 6.25	42.73 \pm 4.70
18	27.01 \pm 3.55	40.99 \pm 3.09	40.88 \pm 3.69	27.99 \pm 2.45	44.11 \pm 2.58	42.46 \pm 2.43	30.46 \pm 1.86	44.18 \pm 2.04	45.69 \pm 1.86
19	34.37 \pm 1.04	48.07 \pm 0.92	46.22 \pm 1.07	35.13 \pm 1.06	48.70 \pm 1.27	47.21 \pm 1.04	35.67 \pm 3.84	49.40 \pm 4.29	47.88 \pm 3.78
20	31.26 \pm 1.58	46.86 \pm 1.13	35.72 \pm 1.61	28.76 \pm 2.18	50.16 \pm 2.28	33.12 \pm 2.24	32.60 \pm 3.85	52.47 \pm 4.25	37.46 \pm 3.83
21	35.91 \pm 3.05	41.29 \pm 2.14	44.02 \pm 3.16	34.58 \pm 2.43	48.66 \pm 2.44	43.56 \pm 2.44	32.90 \pm 2.58	53.28 \pm 2.44	41.96 \pm 2.60
22	24.67 \pm 1.45	38.51 \pm 1.15	38.98 \pm 1.52	23.37 \pm 2.08	41.12 \pm 2.30	37.91 \pm 2.11	22.88 \pm 0.85	43.97 \pm 1.61	38.03 \pm 0.73
23	13.66 \pm 1.51	34.59 \pm 1.06	32.58 \pm 1.73	14.17 \pm 2.79	35.16 \pm 3.00	32.85 \pm 2.84	12.24 \pm 2.32	37.30 \pm 3.63	31.61 \pm 1.89
24	19.28 \pm 1.54	37.39 \pm 0.93	35.92 \pm 1.78	19.27 \pm 0.98	36.12 \pm 1.00	35.24 \pm 1.04	17.68 \pm 1.55	40.68 \pm 1.94	34.84 \pm 1.54
25	19.30 \pm 1.10	33.08 \pm 0.87	30.33 \pm 1.18	19.43 \pm 1.96	30.19 \pm 2.23	29.27 \pm 2.05	16.98 \pm 3.07	37.24 \pm 3.80	28.33 \pm 3.03
Average	20.33 \pm 2.36	39.99 \pm 2.29	37.38 \pm 2.58	22.44 \pm 2.68	40.83 \pm 2.82	40.39 \pm 2.74	21.74 \pm 3.65	44.27 \pm 4.14	40.42 \pm 3.68

C.4.3. BRIER CALIBRATION LOSS

Finally, we report Brier score as an additional proper scoring rule for probabilistic predictions. For binary labels, the Brier score measures the squared error between the predicted probability and the observed outcome. We compute both aggregate Brier score and class-stratified Brier scores to evaluate whether the calibration trends observed with ECE and AECE persist under a metric that does not rely on binning.

The Brier score results reinforce the calibration trends observed with ECE and AECE. For standard unimodal and multimodal models (Table C.12, Table C.13), the positive class exhibits substantially higher Brier loss than the negative class, indicating that errors in predicted confidence are concentrated among positive labels. Loss upweighting reduces this gap by lowering positive-class Brier loss, but it also increases negative-class and aggregate Brier loss (Table C.14). Together, these results suggest that simple reweighting shifts calibration error across classes rather than fully resolving the class-dependent miscalibration that undermines selective prediction reliability.

Table C.12: **Brier Score for Unimodal Models.** Aggregate and class-stratified Brier scores are reported for unimodal baselines. Positive-class Brier loss remains substantially higher than negative-class Brier loss, showing that class-dependent reliability failures persist under a proper scoring rule that does not rely on binning. (Dark-bold: $p < 0.05$, Wilcoxon signed-rank test, 5 seeds; Light-bold: highest mean, not significant)

Clinical Condition	EHR			CXR			MedFuse		
	BS ↓	BS _{c=1} ↓	BS _{c=0} ↓	BS ↓	BS _{c=1} ↓	BS _{c=0} ↓	BS ↓	BS _{c=1} ↓	BS _{c=0} ↓
1	18.75±0.07	37.06±0.30	10.08±0.03	23.54±0.61	40.52±3.54	15.50±2.35	17.90±0.21	34.00±2.46	10.28±1.20
2	5.23±0.03	48.19±0.43	1.54±0.05	8.04±0.30	78.77±4.03	1.98±0.64	5.30±0.12	50.55±1.85	1.43±0.06
3	8.10±0.05	78.15±0.58	0.93±0.02	9.35±0.59	76.08±4.96	2.52±1.13	7.99±0.02	76.07±0.21	1.02±0.03
4	22.91±0.07	37.48±0.14	14.00±0.10	24.46±0.50	36.38±4.50	17.17±3.53	21.21±0.20	34.26±4.39	13.23±2.42
5	16.75±0.11	51.69±0.41	5.66±0.02	21.83±0.63	48.04±2.19	13.51±1.21	15.29±0.18	43.13±5.36	6.45±1.57
6	12.48±0.04	66.83±0.33	2.98±0.04	13.29±0.67	63.54±10.93	4.51±2.60	11.56±0.06	56.58±2.58	3.70±0.46
7	16.47±0.10	55.29±0.52	5.15±0.06	18.89±0.42	62.12±2.77	6.29±1.09	16.24±0.02	53.80±0.66	5.29±0.18
8	9.98±0.04	76.34±0.45	1.36±0.03	7.00±0.54	43.27±6.68	2.29±1.31	6.37±0.17	44.73±2.77	1.39±0.49
9	19.06±0.13	40.56±0.46	10.04±0.06	20.44±1.65	28.45±4.99	17.08±4.28	16.53±0.29	33.18±6.06	9.55±2.69
10	20.00±0.10	40.70±0.18	9.50±0.09	20.84±0.87	36.28±4.18	13.01±3.25	17.67±0.21	32.40±3.60	10.20±2.02
11	9.53±0.13	65.54±1.60	1.89±0.07	12.28±0.44	71.55±6.08	4.20±1.26	8.03±0.05	48.81±1.84	2.47±0.22
12	16.19±0.10	60.04±0.56	4.43±0.04	19.32±0.89	58.21±5.90	8.90±2.52	15.26±0.09	52.94±2.22	5.16±0.51
13	22.47±0.06	31.82±0.17	16.09±0.04	26.58±0.59	25.34±6.30	27.43±5.23	21.50±0.13	27.77±2.30	17.21±1.54
14	23.90±0.12	29.94±0.33	19.29±0.07	28.35±0.93	33.54±5.82	24.38±5.90	22.33±0.14	27.85±2.13	18.10±1.76
15	21.64±0.13	25.92±0.09	18.08±0.18	27.66±0.68	34.63±2.42	21.88±1.41	21.31±0.15	25.98±1.42	17.42±0.93
16	6.52±0.01	84.09±0.26	0.58±0.02	7.31±0.32	82.45±3.10	1.56±0.56	6.49±0.03	80.77±1.32	0.81±0.13
17	15.97±0.08	55.57±0.35	4.67±0.02	20.46±0.33	52.09±0.91	11.43±0.40	14.74±0.16	46.38±5.42	5.71±1.42
18	13.50±0.04	66.02±0.22	2.79±0.05	14.47±0.58	58.61±4.66	5.47±1.64	12.99±0.17	55.93±2.78	4.24±0.75
19	10.85±0.02	74.26±0.43	1.74±0.04	12.98±0.55	73.31±5.12	4.32±1.31	10.76±0.02	75.48±0.76	1.47±0.10
20	5.10±0.02	84.78±0.40	0.55±0.03	5.74±0.27	82.97±1.64	1.33±0.37	5.05±0.01	85.03±0.72	0.48±0.04
21	8.56±0.03	78.76±0.31	1.18±0.04	9.24±0.37	78.18±4.57	1.99±0.86	8.37±0.02	77.82±0.83	1.07±0.10
22	13.59±0.05	54.71±0.34	4.28±0.05	16.18±0.45	62.20±6.11	5.75±1.90	13.24±0.07	54.17±1.79	3.97±0.34
23	15.23±0.27	34.40±0.88	7.72±0.15	19.61±0.36	47.47±5.88	8.69±2.38	14.87±0.18	35.14±1.96	6.92±0.53
24	14.90±0.15	46.48±0.64	5.62±0.06	18.92±0.35	58.83±5.22	7.19±1.87	14.42±0.09	44.08±1.21	5.70±0.44
25	11.27±0.21	45.32±0.81	4.09±0.11	15.12±0.54	60.44±3.38	5.57±1.36	10.94±0.09	43.20±1.32	4.14±0.28
Average	14.36±0.09	54.80±0.45	6.17±0.06	16.88±0.58	55.73±4.64	9.36±2.01	13.45±0.12	49.60±2.32	6.30±0.81

Table C.13: **Brier Score for Multimodal Models.** Aggregate and class-stratified Brier scores are reported for standard multimodal architectures. Across MedFuse, DrFuse, and MeTra, the positive class has substantially higher Brier loss than the negative class, indicating that multimodal fusion does not remove minority-class calibration error despite improving aggregate discrimination. (Dark-bold: $p < 0.05$, Wilcoxon signed-rank test, 5 seeds; Light-bold: highest mean, not significant)

Clinical Condition	MedFuse			DrFuse			MeTra		
	BS ↓	BS _{c=1} ↓	BS _{c=0} ↓	BS ↓	BS _{c=1} ↓	BS _{c=0} ↓	BS ↓	BS _{c=1} ↓	BS _{c=0} ↓
1	17.90±0.21	34.00±2.46	10.28±1.20	18.89±0.31	35.95±3.57	10.97±1.95	18.99±0.20	38.85±0.98	9.58±0.36
2	5.30±0.12	50.55±1.85	1.43±0.06	4.99±0.11	43.53±5.07	2.04±0.49	5.10±0.04	41.10±5.99	2.02±0.51
3	7.99±0.02	76.07±0.21	1.02±0.03	8.02±0.05	78.18±3.84	1.01±0.41	8.05±0.11	72.62±4.76	1.45±0.54
4	21.21±0.20	34.26±4.39	13.23±2.42	22.28±0.14	36.31±5.85	13.72±3.69	22.72±0.28	37.69±5.09	13.56±3.43
5	15.29±0.18	43.13±5.36	6.45±1.57	16.14±0.31	48.24±6.35	6.25±1.66	16.65±0.23	51.22±2.59	5.67±0.83
6	11.56±0.06	56.58±2.58	3.70±0.46	12.21±0.10	66.25±2.26	2.58±0.46	12.50±0.08	66.02±1.37	3.16±0.33
7	16.24±0.02	53.80±0.66	5.29±0.18	15.40±0.12	53.42±2.78	4.93±0.72	15.80±0.09	50.01±3.24	5.83±0.97
8	6.37±0.17	44.73±2.77	1.39±0.49	9.36±0.08	70.99±4.30	1.75±0.57	9.85±0.09	71.25±4.31	1.87±0.64
9	16.53±0.29	33.18±6.06	9.55±2.69	18.17±0.26	40.32±6.49	8.92±2.36	18.68±0.40	40.14±4.29	9.68±2.23
10	17.67±0.21	32.40±3.60	10.20±2.02	19.34±0.77	38.86±7.62	9.52±2.76	19.79±0.20	39.57±2.56	9.76±1.19
11	8.03±0.05	48.81±1.84	2.47±0.22	8.16±0.28	40.21±4.02	3.84±0.84	8.81±0.33	51.87±8.92	2.93±1.57
12	15.26±0.09	52.94±2.22	5.16±0.51	15.27±0.35	46.76±6.28	6.94±2.00	15.42±0.28	51.66±3.16	5.70±0.83
13	21.50±0.13	27.77±2.30	17.21±1.54	22.29±0.40	29.45±6.14	17.28±4.05	22.53±0.17	32.46±4.03	15.75±2.49
14	22.33±0.14	27.85±2.13	18.10±1.76	23.14±0.24	28.89±4.01	18.69±2.96	23.56±0.22	28.67±4.28	19.66±3.58
15	21.31±0.15	25.98±1.42	17.42±0.93	21.27±0.18	25.10±1.49	18.20±1.33	21.94±0.17	27.02±2.63	17.71±2.33
16	6.49±0.03	80.77±1.32	0.81±0.13	6.14±0.02	79.44±1.44	0.72±0.12	6.39±0.04	79.79±3.52	0.78±0.26
17	14.74±0.16	46.38±5.42	5.71±1.42	15.55±0.29	51.52±7.32	5.60±1.71	15.87±0.21	54.52±1.97	4.84±0.61
18	12.99±0.17	55.93±2.78	4.24±0.75	13.00±0.09	58.60±2.93	3.85±0.54	13.39±0.13	64.43±3.39	2.98±0.64
19	10.76±0.02	75.48±0.76	1.47±0.10	10.89±0.03	75.88±1.86	1.52±0.28	10.97±0.04	74.75±1.39	1.81±0.16
20	5.05±0.01	85.03±0.72	0.48±0.04	4.82±0.03	83.89±1.62	0.45±0.09	5.07±0.08	83.66±3.67	0.58±0.26
21	8.37±0.02	77.82±0.83	1.07±0.10	8.59±0.03	79.92±1.63	0.89±0.14	8.63±0.06	78.48±2.44	1.29±0.32
22	13.24±0.07	54.17±1.79	3.97±0.34	12.95±0.11	54.14±3.30	3.67±0.65	13.40±0.20	54.80±3.82	4.03±0.78
23	14.87±0.18	35.14±1.96	6.92±0.53	14.87±0.17	33.10±2.15	7.96±0.82	15.05±0.22	31.60±4.13	8.56±1.66
24	14.42±0.09	44.08±1.21	5.70±0.44	13.93±0.11	40.38±3.02	6.30±0.89	14.34±0.22	45.76±4.60	5.11±1.31
25	10.94±0.09	43.20±1.32	4.14±0.28	10.36±0.22	37.42±1.51	4.93±0.50	10.92±0.30	44.53±5.15	3.84±1.19
Average	13.45±0.12	49.60±2.32	6.30±0.81	13.84±0.19	51.07±3.87	6.50±1.28	14.18±0.18	52.50±3.69	6.33±1.16

Table C.14: **Brier Score for Multimodal Models under Loss-Upweighting.** Aggregate and class-stratified Brier scores are reported for multimodal architectures trained with loss upweighting. Loss upweighting reduces positive-class Brier loss but increases negative-class and aggregate Brier loss, suggesting that the intervention shifts calibration error across classes rather than fully resolving the reliability gap. (Dark-bold: $p < 0.05$, Wilcoxon signed-rank test, 5 seeds; Light-bold: highest mean, not significant)

Clinical Condition	MedFuse (Loss-Upweighting)			DrFuse (Loss-Upweighting)			MeTra (Loss-Upweighting)		
	BS ↓	BS _{c=1} ↓	BS _{c=0} ↓	BS ↓	BS _{c=1} ↓	BS _{c=0} ↓	BS ↓	BS _{c=1} ↓	BS _{c=0} ↓
1	20.22±0.47	19.19±1.02	20.71±1.16	21.30±0.55	20.82±1.21	21.52±1.34	21.35±0.81	22.66±3.41	20.73±2.78
2	10.68±0.51	17.41±0.75	10.10±0.61	10.46±1.44	16.47±2.92	10.00±1.77	12.87±1.95	11.96±1.37	12.94±2.23
3	20.79±1.45	23.15±2.02	20.55±1.80	24.24±4.12	21.09±5.20	24.56±5.03	20.63±4.35	25.50±8.98	20.14±5.71
4	22.49±0.73	21.98±2.67	22.81±2.78	24.15±0.56	22.95±1.58	24.88±1.85	23.48±0.51	25.52±3.00	22.23±2.64
5	20.05±1.22	20.14±2.46	20.02±2.35	22.08±1.49	20.60±2.84	22.54±2.81	21.79±2.29	23.77±6.22	21.17±4.94
6	18.68±3.02	25.18±5.83	17.55±4.56	23.65±1.18	23.04±1.62	23.76±1.67	21.94±0.97	27.74±1.44	20.92±1.39
7	23.33±0.38	21.81±0.78	23.78±0.71	22.71±0.78	21.33±1.23	23.09±1.33	21.41±1.24	23.51±3.18	20.80±2.50
8	13.27±2.93	22.82±3.49	12.03±3.76	22.72±2.70	24.31±3.81	22.53±3.49	20.68±3.72	27.08±6.30	19.85±5.01
9	19.19±1.74	19.15±3.82	19.21±4.04	22.59±1.14	19.14±2.39	24.03±2.59	21.79±1.54	24.53±4.91	20.64±4.17
10	19.96±1.16	19.70±3.53	20.09±3.49	22.19±1.43	19.26±3.37	23.66±3.80	21.59±0.98	25.94±6.96	19.38±4.93
11	14.69±0.96	16.26±0.92	14.48±1.20	16.54±2.52	13.18±3.00	17.00±3.27	18.50±4.04	16.60±5.45	18.76±5.28
12	21.60±0.45	21.43±1.30	21.65±0.87	22.71±2.39	19.03±3.36	23.69±3.90	23.51±3.10	20.98±5.14	24.19±5.27
13	22.50±0.44	22.32±3.24	22.63±2.94	23.21±0.93	21.06±3.53	24.72±4.00	23.34±0.78	24.26±5.28	22.71±4.80
14	22.91±0.42	24.05±3.45	22.04±3.30	23.92±0.47	21.48±1.75	25.81±2.17	24.09±0.46	22.84±2.49	25.05±2.69
15	21.58±0.10	21.95±1.02	21.27±0.96	21.36±0.25	21.58±1.93	21.18±1.82	22.09±0.24	22.54±1.02	21.71±1.01
16	21.06±1.59	27.14±2.09	20.59±1.87	21.18±2.09	21.66±2.93	21.14±2.46	23.25±2.74	22.94±3.54	23.28±3.21
17	20.22±1.38	20.99±2.71	20.00±2.53	22.39±1.61	21.33±3.03	22.68±2.88	21.99±2.30	23.99±5.58	21.41±4.52
18	21.62±2.15	22.05±2.89	21.53±3.18	22.31±1.42	23.29±2.37	22.12±2.16	23.48±1.04	21.65±1.95	23.86±1.64
19	23.46±0.72	25.00±0.88	23.24±0.95	23.69±0.84	24.90±1.11	23.52±1.12	24.68±2.72	25.95±4.21	24.50±3.71
20	17.87±0.90	28.31±1.16	17.27±1.00	16.35±1.54	31.72±1.72	15.50±1.72	18.25±2.74	31.37±4.72	17.50±3.11
21	23.58±2.10	20.74±2.31	23.87±2.56	22.06±1.62	26.16±2.61	21.61±2.05	20.65±1.95	30.12±2.48	19.66±2.40
22	20.27±0.67	18.95±1.18	20.57±1.08	19.66±1.06	21.47±1.84	19.26±1.71	19.73±0.57	23.25±1.21	18.93±0.97
23	16.97±0.61	18.47±0.78	16.38±1.14	17.26±0.90	18.08±2.21	16.94±2.07	17.22±0.56	19.39±2.84	16.37±1.83
24	19.49±0.74	19.72±1.00	19.42±1.24	18.83±0.45	18.54±0.80	18.91±0.79	18.21±0.56	21.43±1.53	17.26±1.12
25	16.82±0.53	18.23±0.97	16.53±0.81	16.31±0.98	16.05±1.63	16.36±1.51	15.45±1.38	20.60±2.97	14.36±2.30
Average	19.73±1.09	21.45±2.09	19.53±2.04	20.96±1.38	21.14±2.40	21.24±2.37	20.88±1.74	23.45±3.85	20.33±3.21

C.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 (Figure C.1), 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 (Figure C.2) shows stratified calibration curves closely aligned with the multimodal results from MedFuse (Figure C.3), though with consistently higher positive-class ECE for several low-prevalence conditions.

Extending this analysis to DrFuse (Figure C.4) and MeTRA (Figure C.5) 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

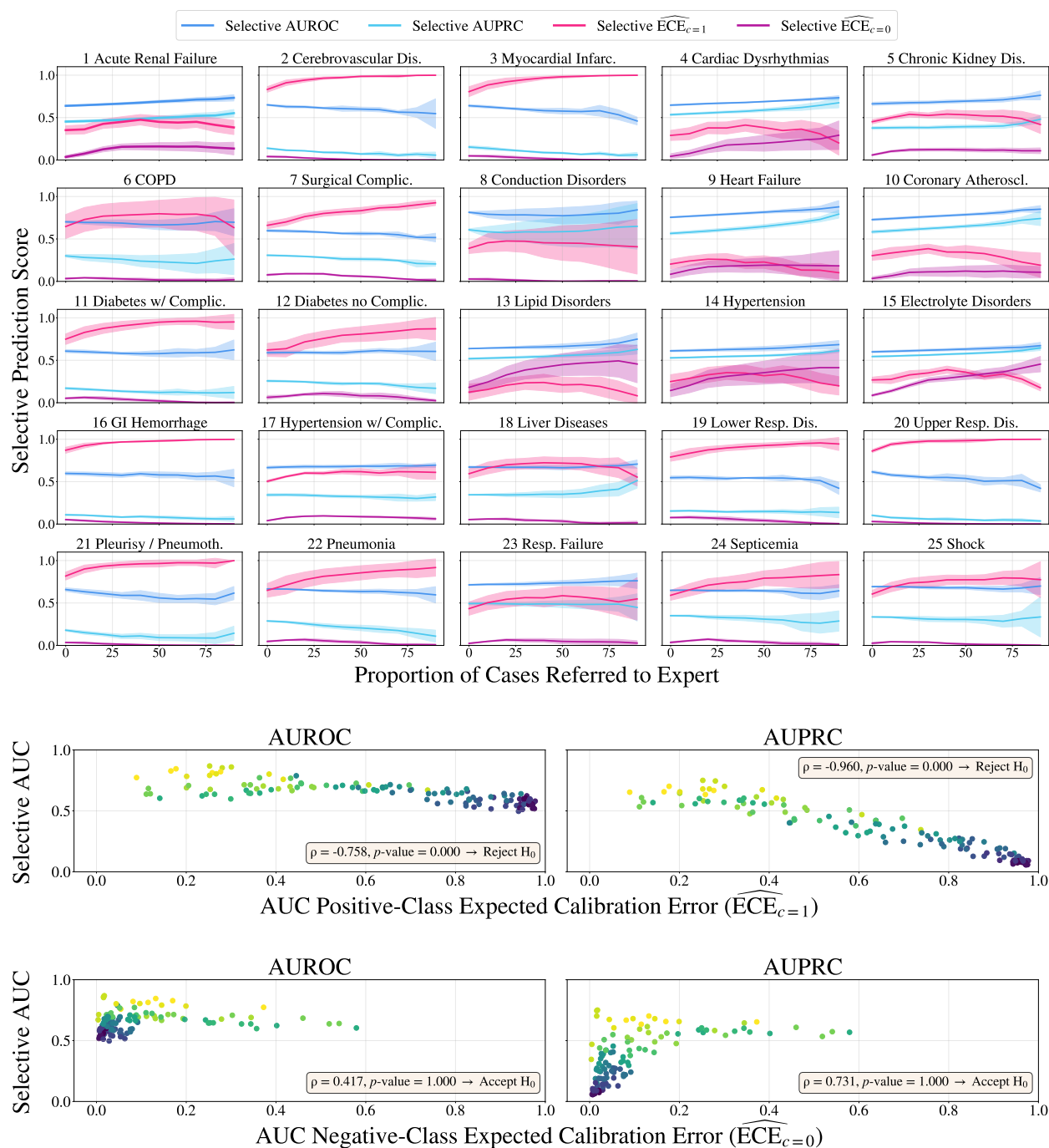


Figure C.1: **Unimodal CXR Stratified Calibration and Selective Prediction Across Clinical Conditions.** For unimodal CXR, stratified positive-class ECE exhibits substantial variability across seeds and conditions (top), and AUROC/AUPRC remain below multimodal MedFuse performance. Across conditions, higher positive-class ECE AUC is associated with lower selective AUC (middle; Spearman's rank $p < 0.05$). Negative-class ECE AUC shows no clear monotonic relationship with selective AUC (bottom).

Unimodal EHR

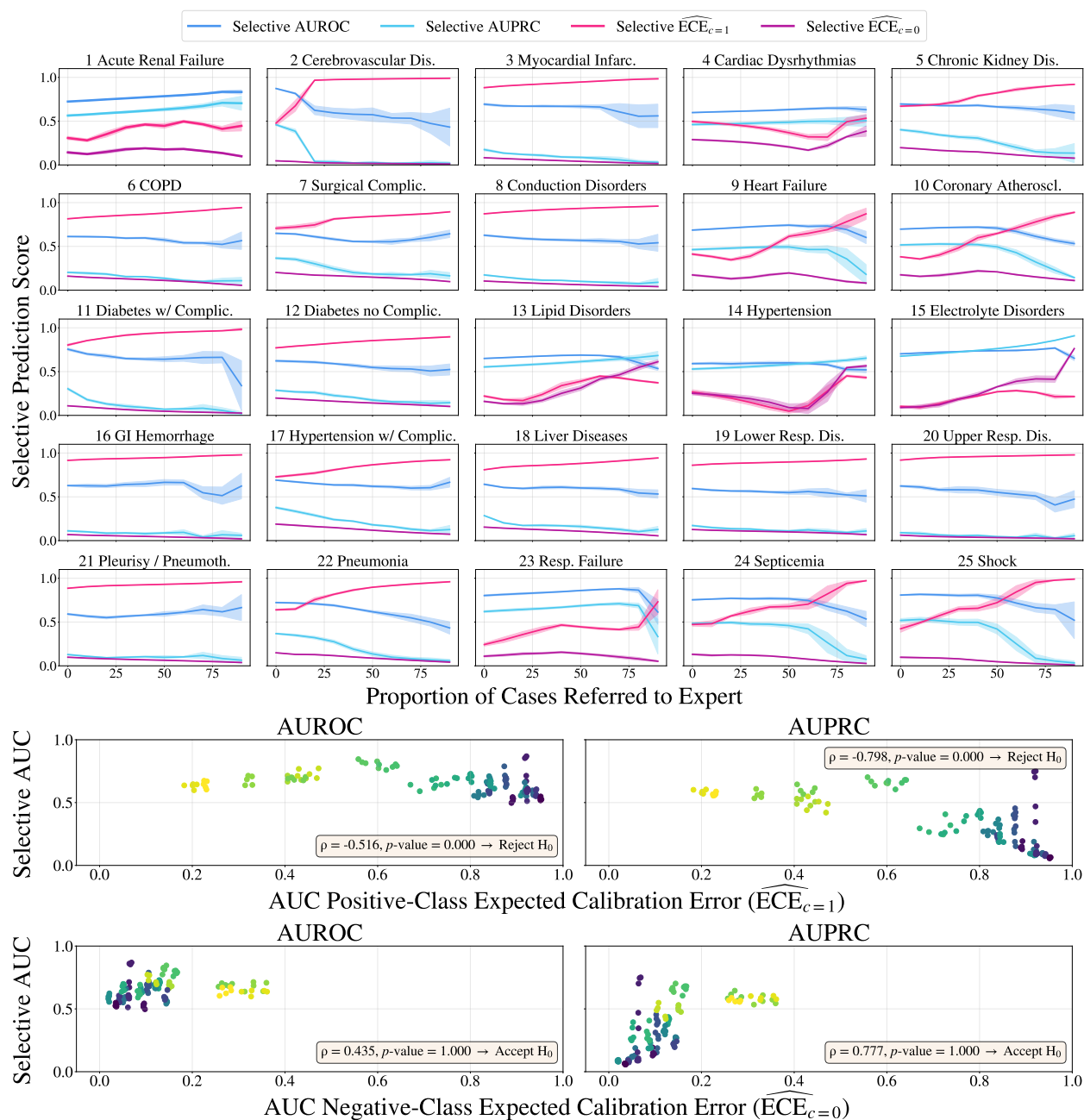


Figure C.2: **Unimodal EHR Stratified Calibration and Selective Prediction Across Clinical Conditions.** Positive-class ECE is elevated for several conditions, indicating overconfidence on underrepresented positive cases. Compared to CXR, stratified ECE and AUROC/AUPRC show lower variability across seeds and conditions, yielding more stable predictions that closely mirror MedFuse behavior. Positive-class ECE AUC remains negatively associated with selective AUC, while negative-class ECE shows no clear monotonic relationship with selective performance.

MedFuse

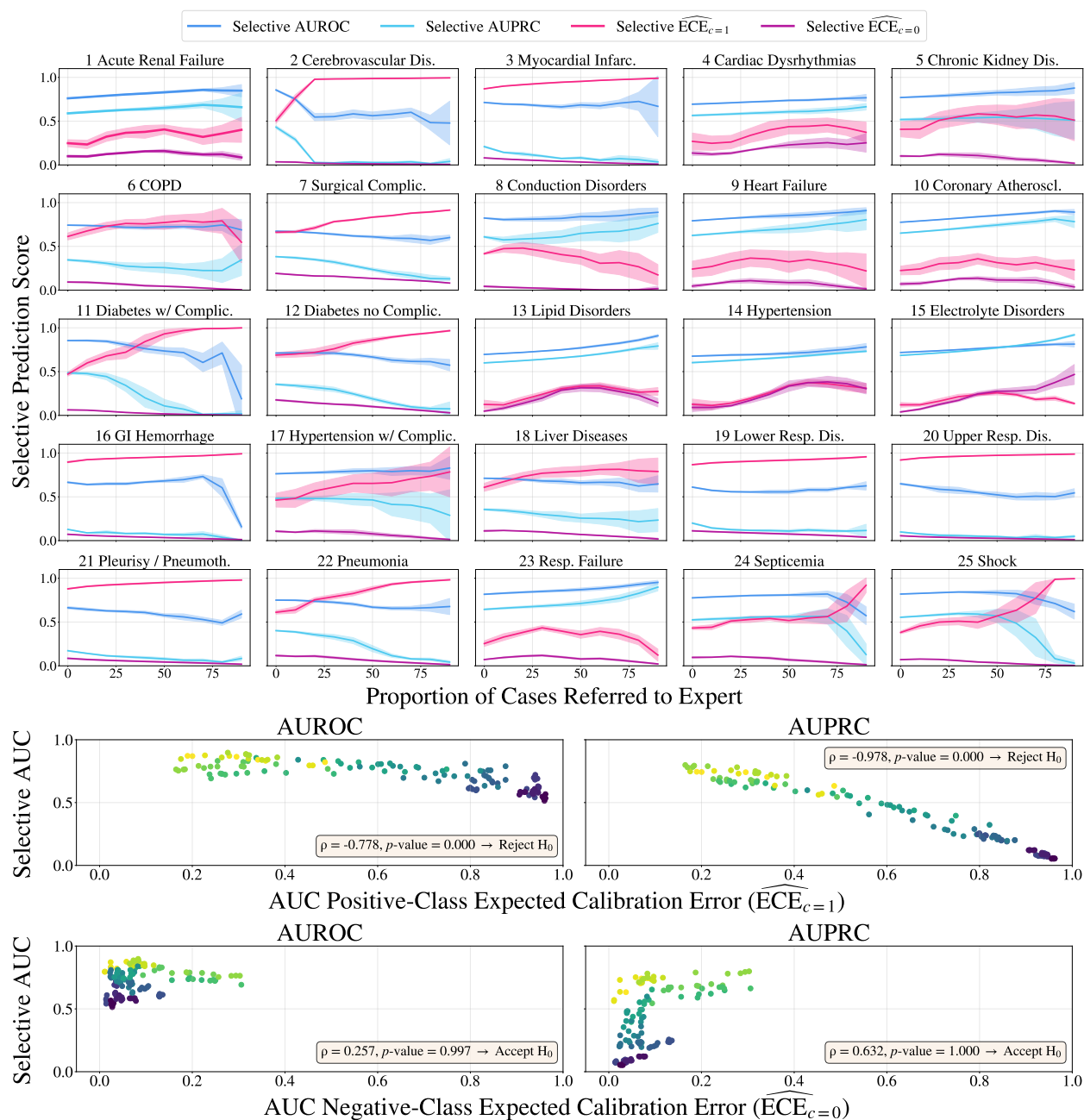


Figure C.3: **MedFuse Stratified Calibration and Selective Prediction Across Clinical Conditions.** Across conditions, higher positive-class ECE AUC consistently predicts lower selective AUC, indicating that overconfidence in positive predictions drives instability in selective evaluation. In contrast, negative-class ECE AUC remains low for most conditions and shows no clear monotonic relationship with selective AUC, suggesting that miscalibration is primarily driven by overconfident positive predictions rather than negative-class errors.

DrFuse

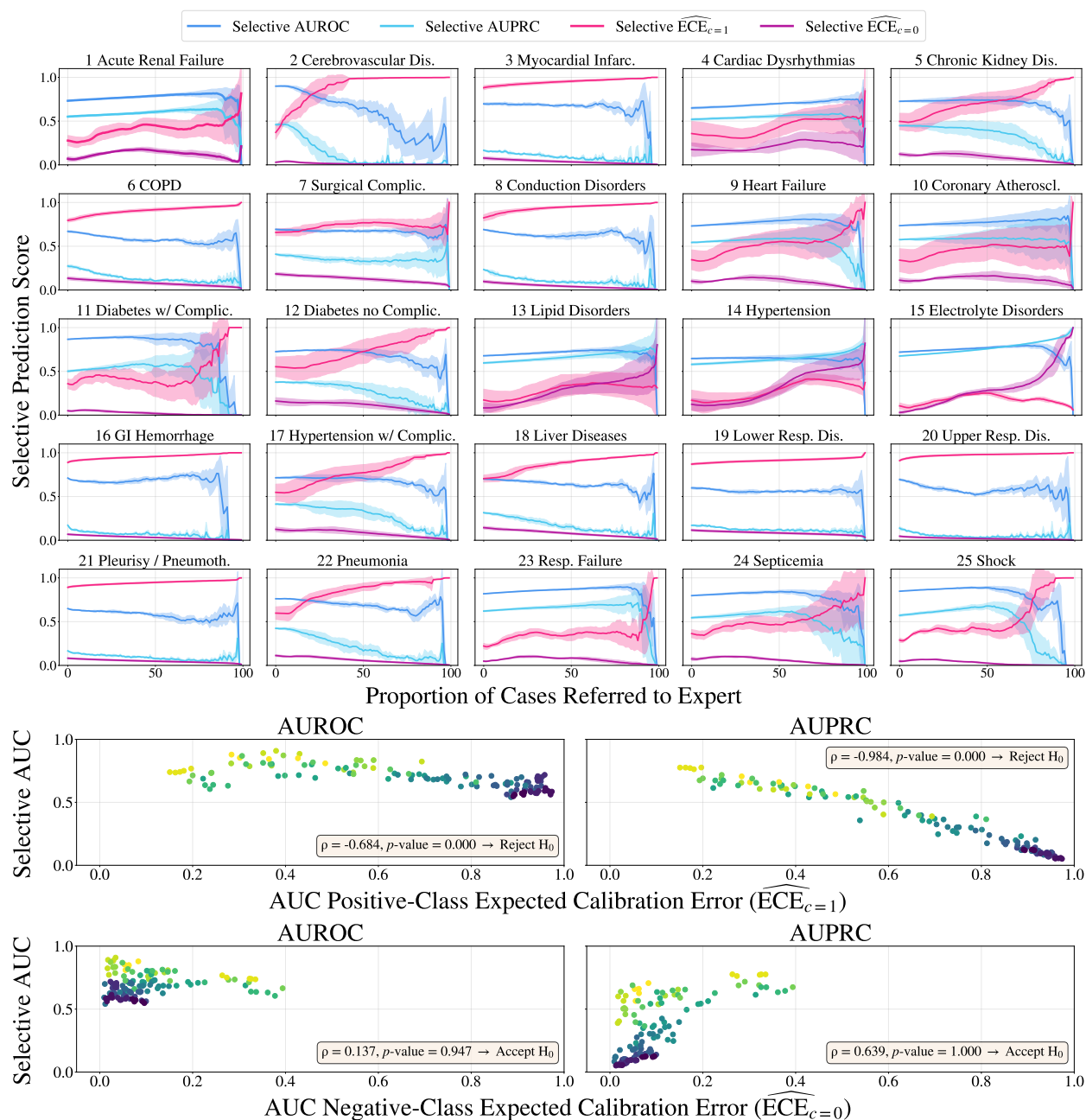


Figure C.4: **DrFuse Stratified Calibration and Selective Prediction Across Clinical Conditions.** DrFuse exhibits elevated positive-class ECE across several conditions, indicating persistent overconfidence on underrepresented positive cases. Stratified ECE and selective AUROC/AUPRC closely mirror MedFuse trends, with higher positive-class ECE AUC associated with degraded selective performance. As in other models, negative-class ECE shows no clear monotonic relationship with selective AUC, suggesting that increased fusion complexity alone does not resolve calibration-driven reliability failures.

MeTRA

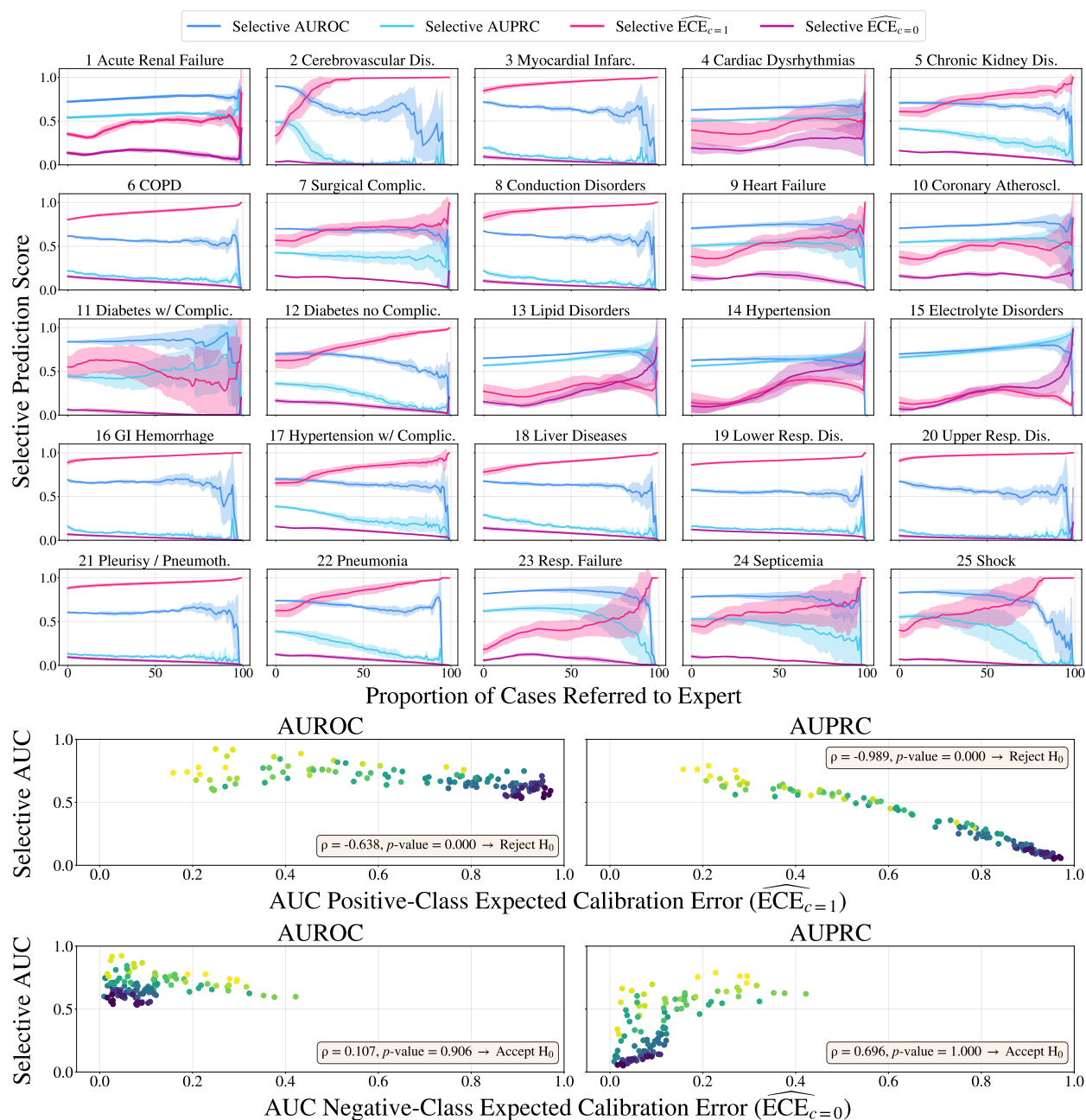


Figure C.5: **MeTRA Stratified Calibration and Selective Prediction Across Clinical Conditions.** MeTRA shows class-dependent calibration behavior similar to MedFuse and DrFuse, with elevated positive-class ECE indicating overconfidence on underrepresented positives. Positive-class ECE AUC is negatively associated with selective AUC, while selective AUROC/AUPRC exhibit higher variance across seeds, consistent with other multimodal models, indicating that architectural complexity alone does not resolve calibration-driven selective instability.

Appendix D. Per-Condition Comparative Analysis

To provide qualitative insight into selective prediction behavior across models, we examine three representative clinical conditions: Acute Cerebrovascular Disease (2) (Figure D1), Conduction Disorders (8) (Figure D2), and Respiratory Shock (25) (Figure D3), across all model variants (EHR, CXR, MedFuse, MedFuse + Loss-Upweight, DrFuse and MeTra). These examples highlight how calibration and selective prediction patterns vary across modalities and fusion strategies, and how simple loss upweighting affects positive-class calibration without fundamentally altering the selective trends observed across models.

D.1. Label 2: Acute Cerebrovascular Disease

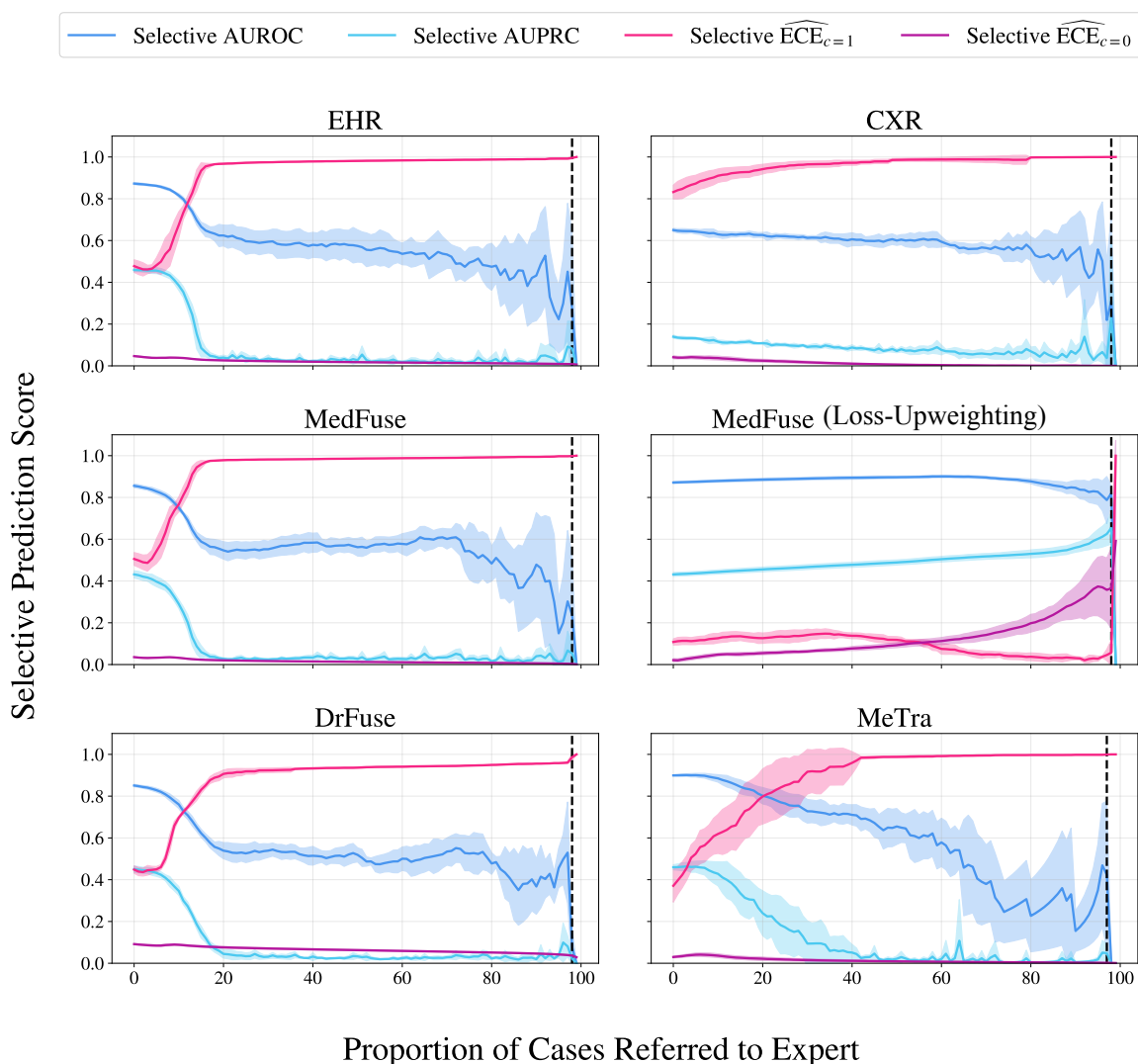


Figure D1: **Comparative analysis for Acute Cerebrovascular Disease.** For this condition, EHR predominantly drives the discriminative performance of MedFuse, while CXR shows poor positive-class calibration and contributes to the oscillatory behavior of selective AUROC. Applying label-weighted loss reduces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

D.2. Label 8: Conduction Disorders

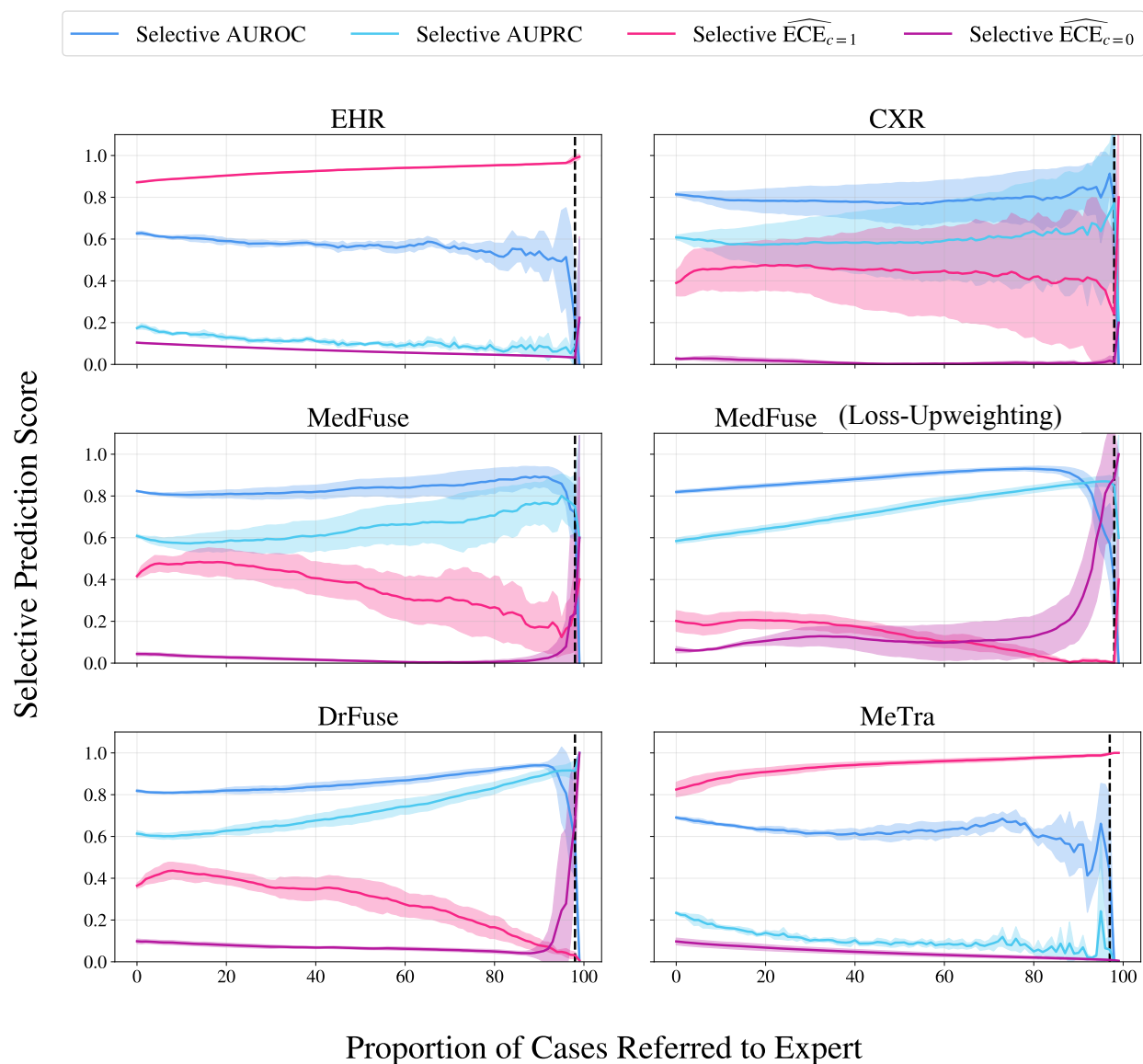


Figure D2: **Comparative analysis for Conduction Disorders.** For this condition, the CXR model shows substantial seed-dependent variability, achieving higher mean performance but poorer calibration consistency than EHR. MedFuse integrates both modalities to improve selective AUROC and stabilize performance, 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.

D.3. Label 25: Respiratory Shock

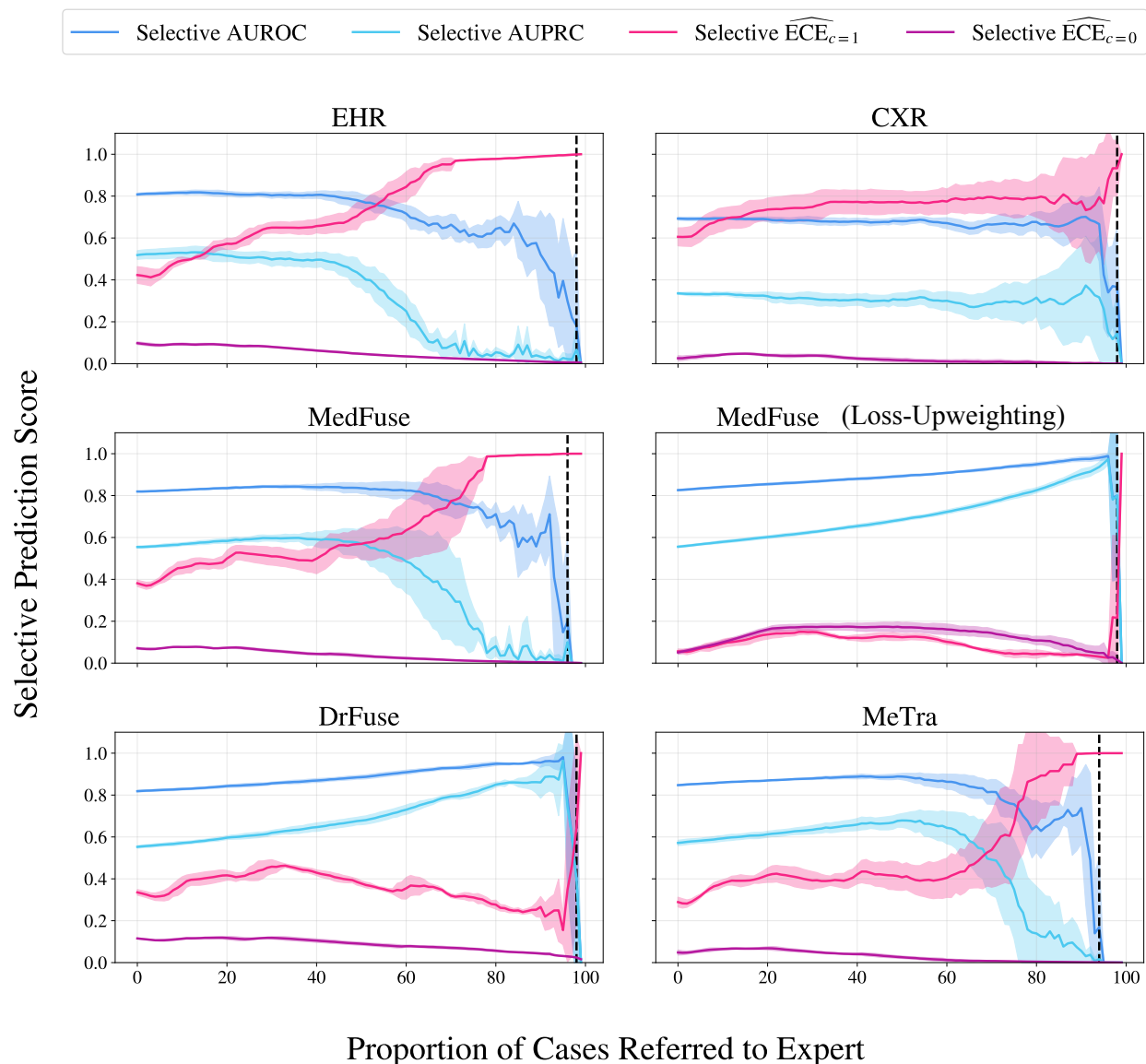


Figure D3: **Comparative analysis for Respiratory Shock.** For this condition, unimodal EHR and CXR achieve comparable baseline performance, with EHR showing slightly lower positive-class ECE at early thresholds. As rejection increases, however, EHR’s AUROC quickly deteriorates while CXR remains uniformly 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.

Appendix E. 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. We plot the details of the regression curves for all 25 conditions using two independent predictors for ECE: AUROC (Figure D4) and AUPRC (Figure D5).

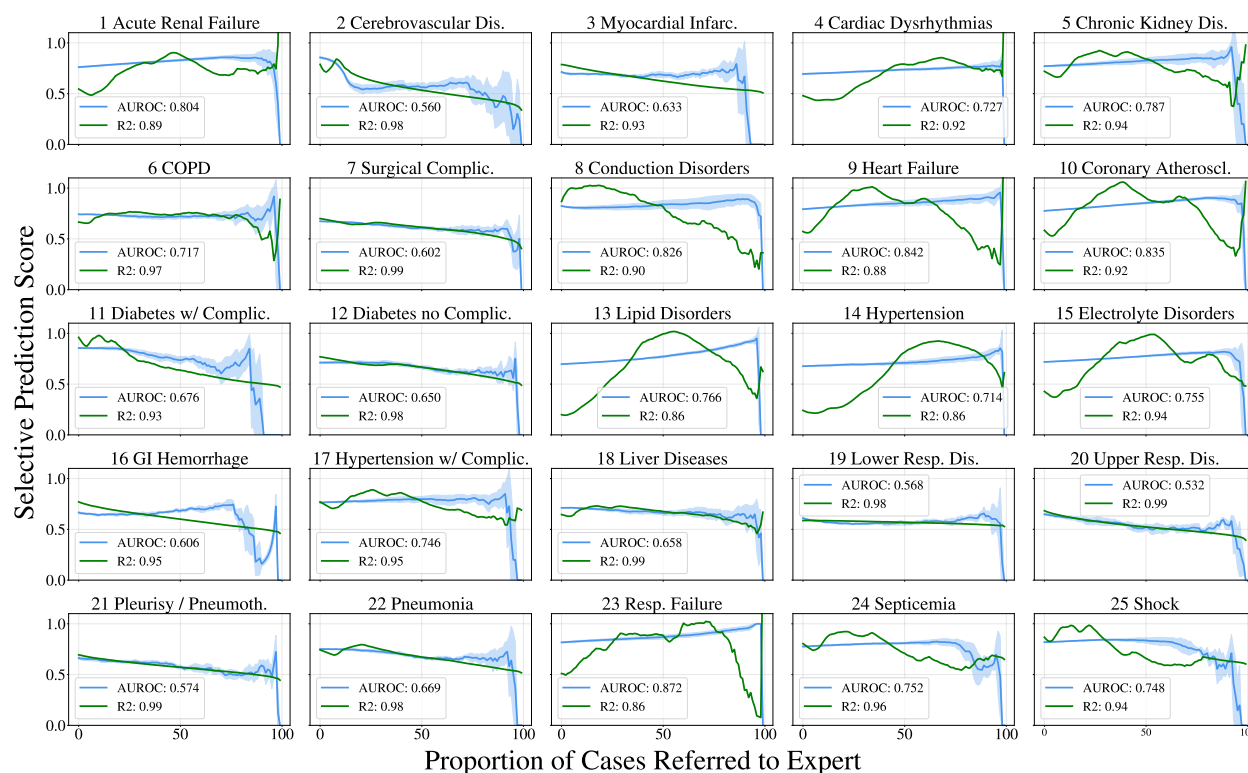


Figure D4: **Linear Regression of AUROC Against Stratified ECE.** Relationship between AUROC and both positive- and negative-class ECE across 25 conditions. Regression slopes demonstrate a consistent, albeit not conclusive, negative association, indicating that poorer calibration modestly reduces discriminative performance in terms of AUROC.

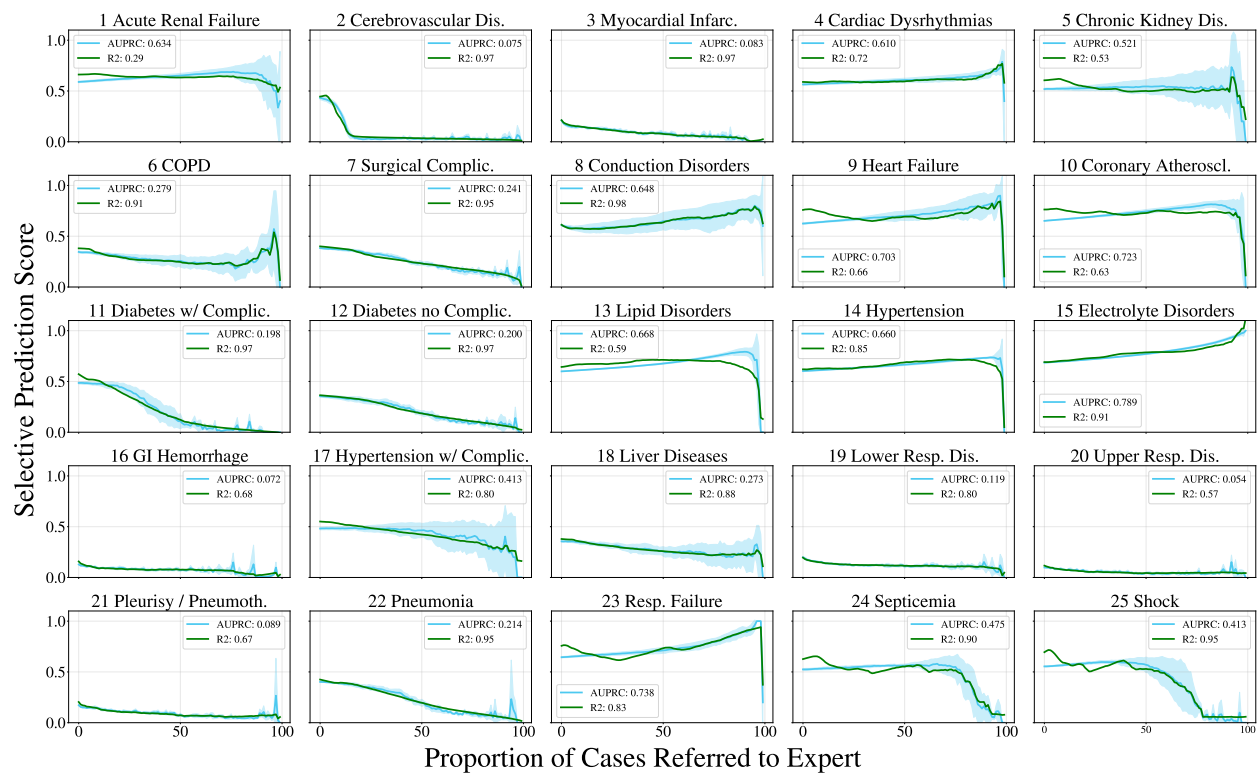


Figure D5: **Linear Regression of AUPRC Against Stratified ECE.** Relationship between AUPRC and class-stratified ECE across 25 conditions. The correlation between ECE and AUPRC exhibits a clear, monotonic decline, confirming via another approach, that high positive-class miscalibration strongly predicts lower precision-recall performance under selective prediction.