EMORY COLLEGE OF ARTS AND SCIENCES

Emory College of Arts and Sciences Honors Program
Department of Data and Decision Sciences

# Causal Representation Learning under Informative Missingness for Clinical Multimodal Prediction and Offline Decision-Making

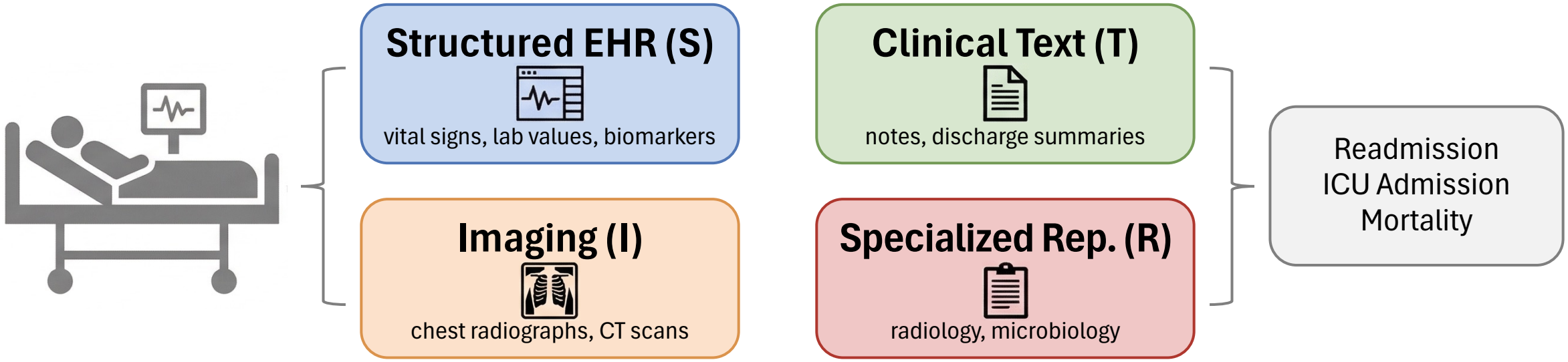
Zihan Liang
Thesis Advisor: Prof. Ruoxuan Xiong
Thesis Committee: Prof. Ali Emami, Prof. Manuela Manetta

# Big Picture & Motivation

From clinical observations to a modeling challenge

# Promise: AI Meets the Modern Hospital

***A hospitalized patient generates data from multiple streams simultaneously.***




Each modality encodes a **different view** of the patient's physiological state and disease trajectory.

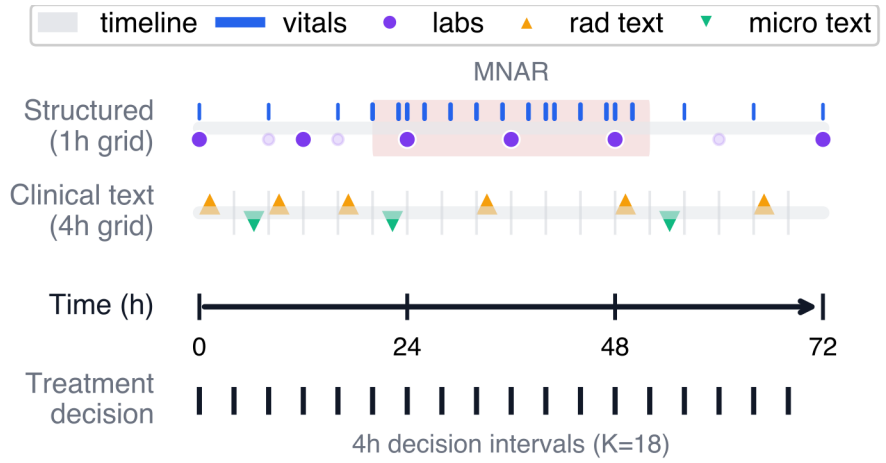
**If AI can integrate these streams** → Predict 30-day readmission · ICU admission · In-hospital mortality · Guide treatment decisions

# Reality: Clinical Data Is Systematically Incomplete

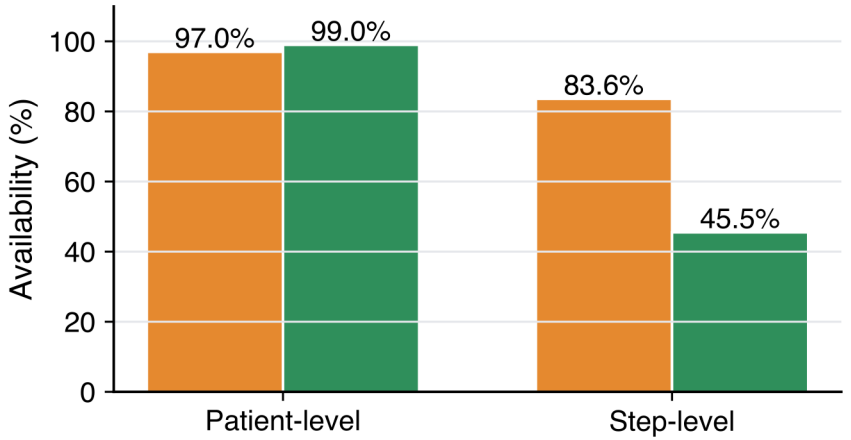
- ✓ Structured vitals & labs → recorded on a 1-hour grid
- ✓ Clinical text (notes, summaries) → sparse, on a 4-hour decision grid
- ✓ In MIMIC-IV: 24.5% of hospitalized patients have no discharge summary
- ✓ Microbiology reports available at only 45.5% of ICU decision steps
- ✓ Text coverage: 97–99% at the patient level, but only 45.5–83.6% at the step level

*These gaps are not random artifacts of data collection. They reflect deliberate clinical decisions.*

**The data was not lost — it was never collected.**




**Multimodal Clinical Trajectory**

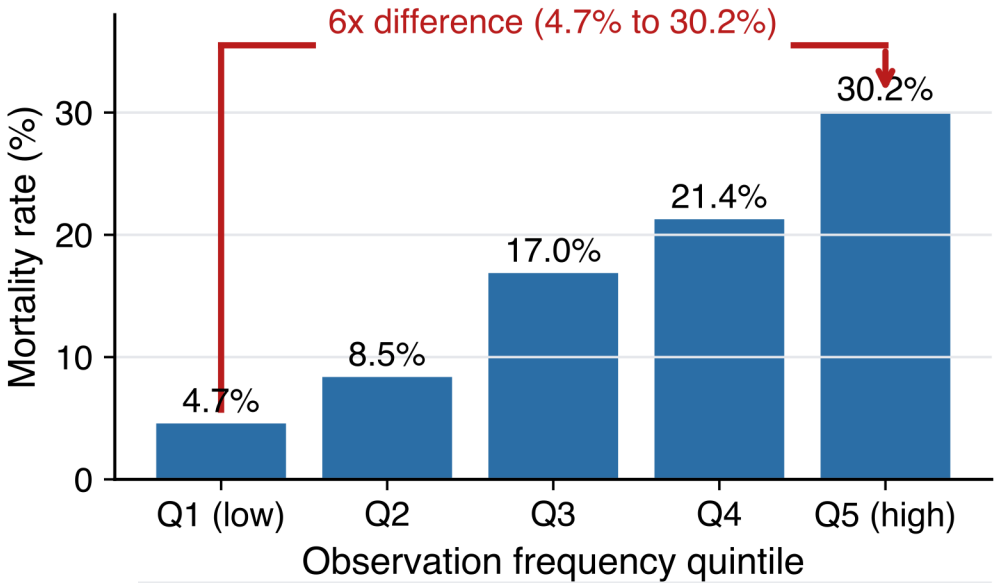



**Text Availability: Patient vs. Step Level**

# The Key Intuition: Absence Is a Signal

**A concrete example:**

A clinician orders a lactate measurement when **sepsis is suspected** — but not for patients who appear hemodynamically stable. The decision to observe depends on the very quantity the measurement would characterize.




**Mortality by Observation Frequency**

This is **Missing Not At Random (MNAR)**: the probability of observing a value depends on the value itself.

**~6× difference**. Sicker patients are monitored more intensively.

*"How patients are observed is at least as informative as what is observed in many clinical scenarios."*

# Where the Literature Falls Short

| Area | Best existing work | Limitation |
|---|---|---|
| **Multimodal fusion** | SMIL, MUSE+, DrFuse | Assume MAR; don't model *why* data is missing |
| **Temporal EHR models** | GRU-D, BRITS, Raindrop | Step-level only; no cross-scale unification |
| **Clinical RL** | AI Clinician, CQL | Don't address MNAR in state learning |
| **Causal missing data** | Standard MNAR methods | No connection to policy learning or multimodal fusion |

**Four gaps this thesis addresses:**

- Gap 1: Static vs. temporal missingness — never unified
- Gap 2: Observation process ($\delta \rightarrow y$) — causally unmodeled
- Gap 3: Prediction and decision-making under MNAR — developed independently
- Gap 4: No systematic diagnostic framework for MNAR failure types

# The Core Problem: Informative Missingness

From missing data to informative observation processes

# Why Standard Approaches Fail under MNAR

Standard missing data methods treat missingness as **noise to remove** — not signal to model.

## Missing = Signal

observation pattern carries information

| Patient ID | Date | Heart Rate | Blood Pressure | Lab Result |
|---|---|---|---|---|
| 303101 | 21-03-25 | | | B.0 |
| 302002 | 21-03-26 | 84 | | A.4 |
| 302003 | 21-03-27 | | 190 | B.3 |
| 203104 | 21-03-25 | 86 | | B.0 |
| 303105 | 21-03-25 | 70 | | B.4 |
| 302006 | 21-03-26 | | 120 | A.3 |

Illustrative example

**Mean Imputation**
fill with averages

wrong method
Loses pattern info

**Complete-Case Analysis**
drop incomplete records

wrong method
Selection Bias

**Multiple Imputation**
sample from assumed model

wrong method
Assumes MAR

**MNAR-aware Modeling**
Treat $\delta$ as a **causal variable**
Model observation patterns explicitly

# MCAR, MAR, and MNAR: Why the Distinction Matters

Three mechanisms by which data may be missing.

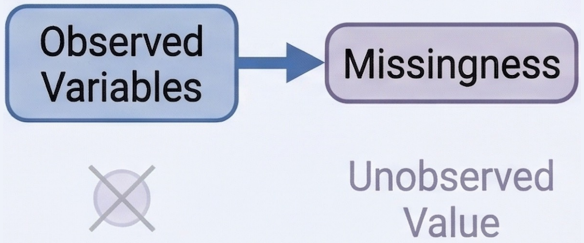
## MCAR

"independent of observed and unobserved data"


**Clinical Intuition:**
"Lab instrument randomly breaks"

## MAR

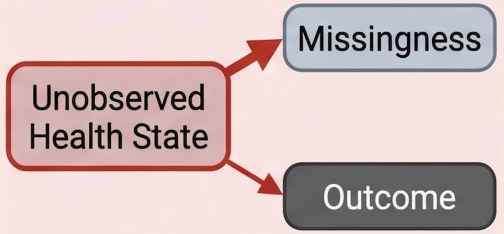
"depends on observed values only"




**Clinical Intuition:**
" Doctor skips a test based on known chart history"

## MNAR

“depends on the unobserved value”

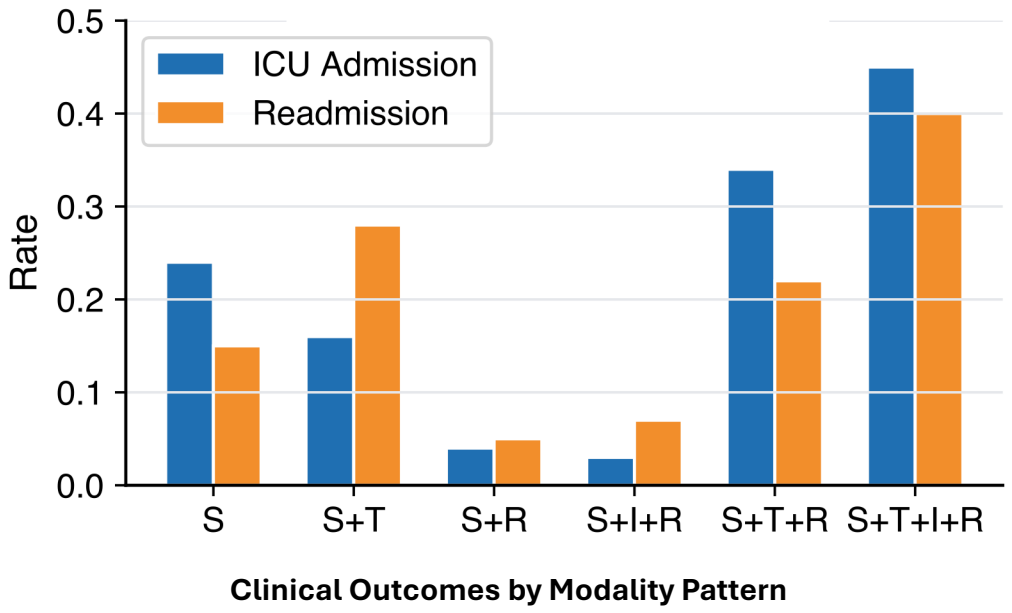



**Clinical Intuition:**
" Doctor orders lactate because sepsis is suspected"

Clinical missingness is predominantly MNAR.

**Key implication:** The observation process correlates with the underlying clinical state — making it a *signal*, not noise.

# Evidence: Which Tests Were Run Predicts What Happens Next




**Clinical Outcomes by Modality Pattern**

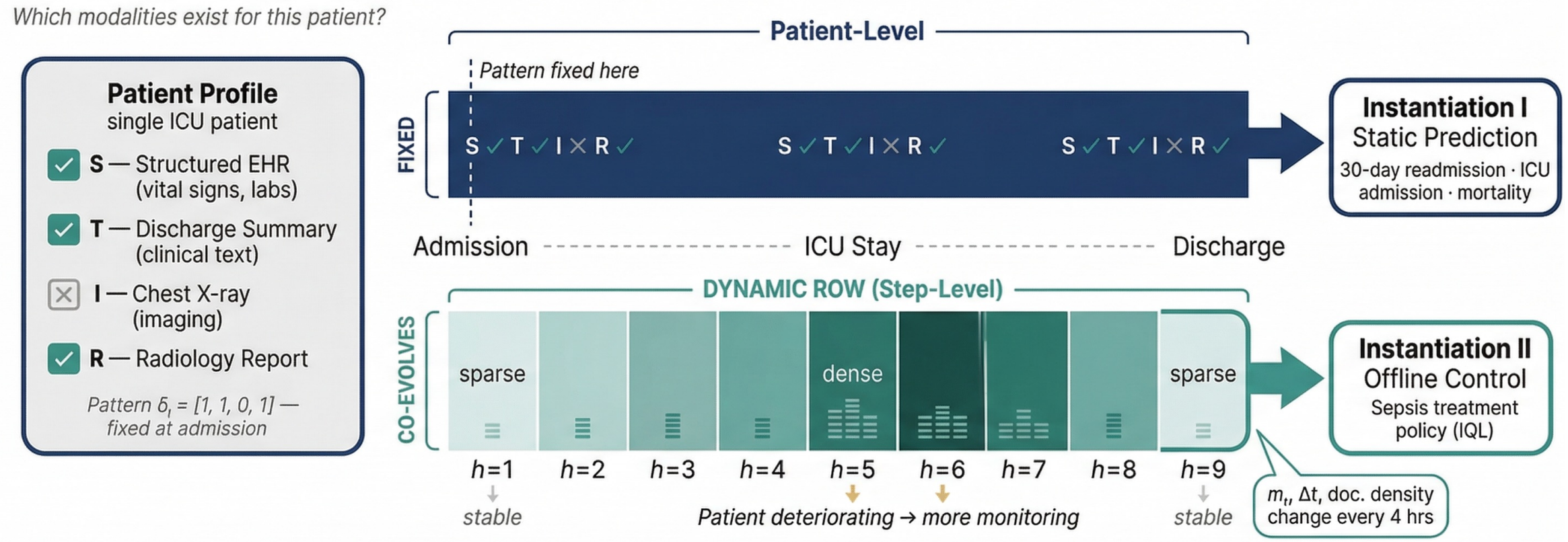
Patients with **complete multimodal records (S+I+T+R)** have **substantially higher** rates of ICU admission and 30-day readmission than those with only structured data (S only).

**Conclusion:**

Which tests a clinician chose to run is itself a predictor of outcome — independent of the test results.

# Missingness Operates at Two Distinct Scales

This thesis identifies two complementary scales of informative missingness in clinical data:



# My Thesis Question and Overall Framework

From problem formulation to a unified modeling framework

# The Thesis Question

*"How should observation patterns in multimodal clinical data — at both patient and step levels — be modeled to learn representations that are sufficient for prediction and control, and what methodological interventions correct the distinct failure modes that arise when such patterns are ignored?"*

**In plain language:**

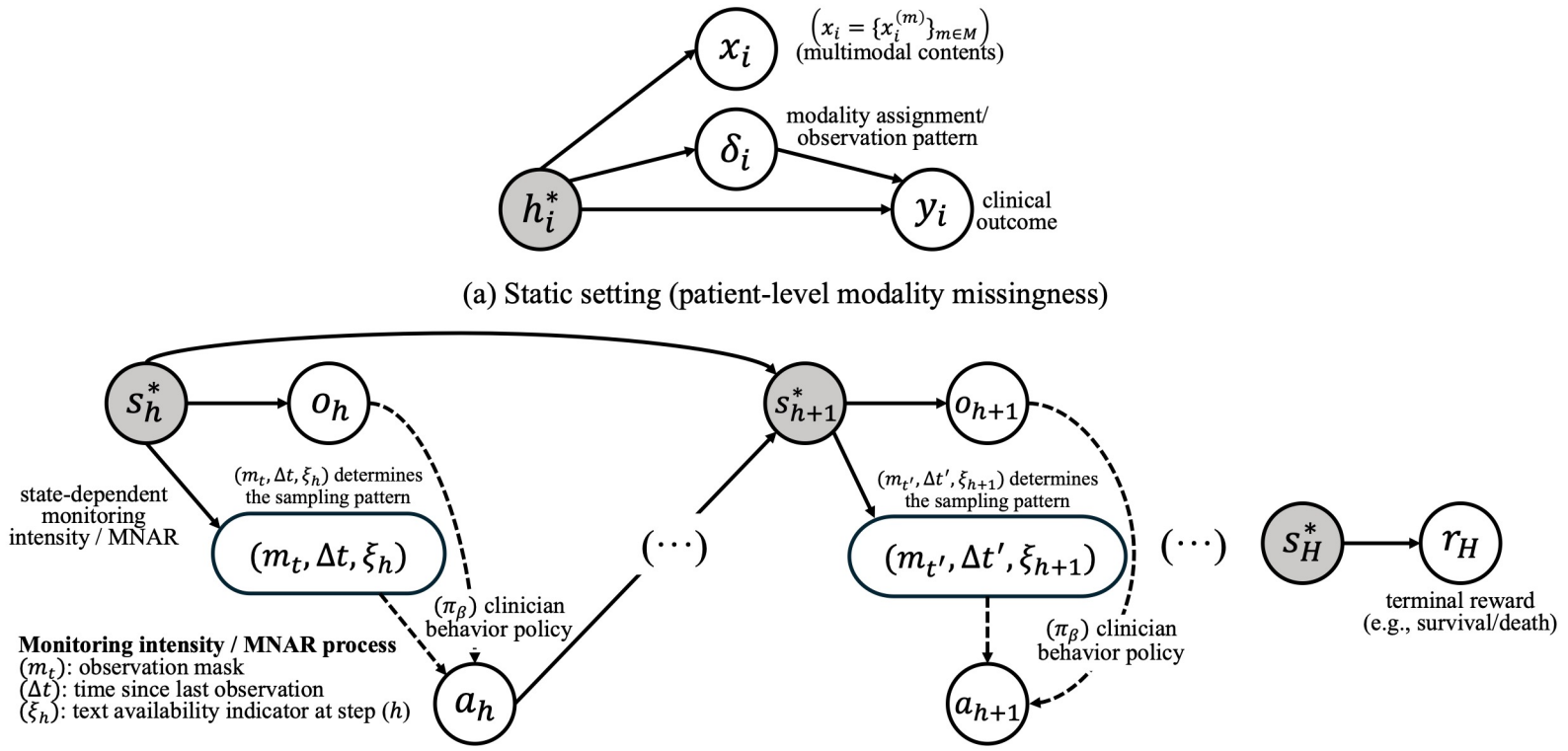
- **Q1 (Modeling):** How do we treat *who was tested, and how often* as signal — not noise?
- **Q2 (Correction):** When we ignore this, what breaks — and how do we fix it?

**Three verifiable objectives:**

- **O1** — Representation sufficiency under MNAR
- **O2** — Observation process as explicit causal signal
- **O3** — Unified error decomposition & diagnostics

# The Unified Causal Framework

## Static Setting (patient-level modality missingness)




(a) Static setting (patient-level modality missingness)

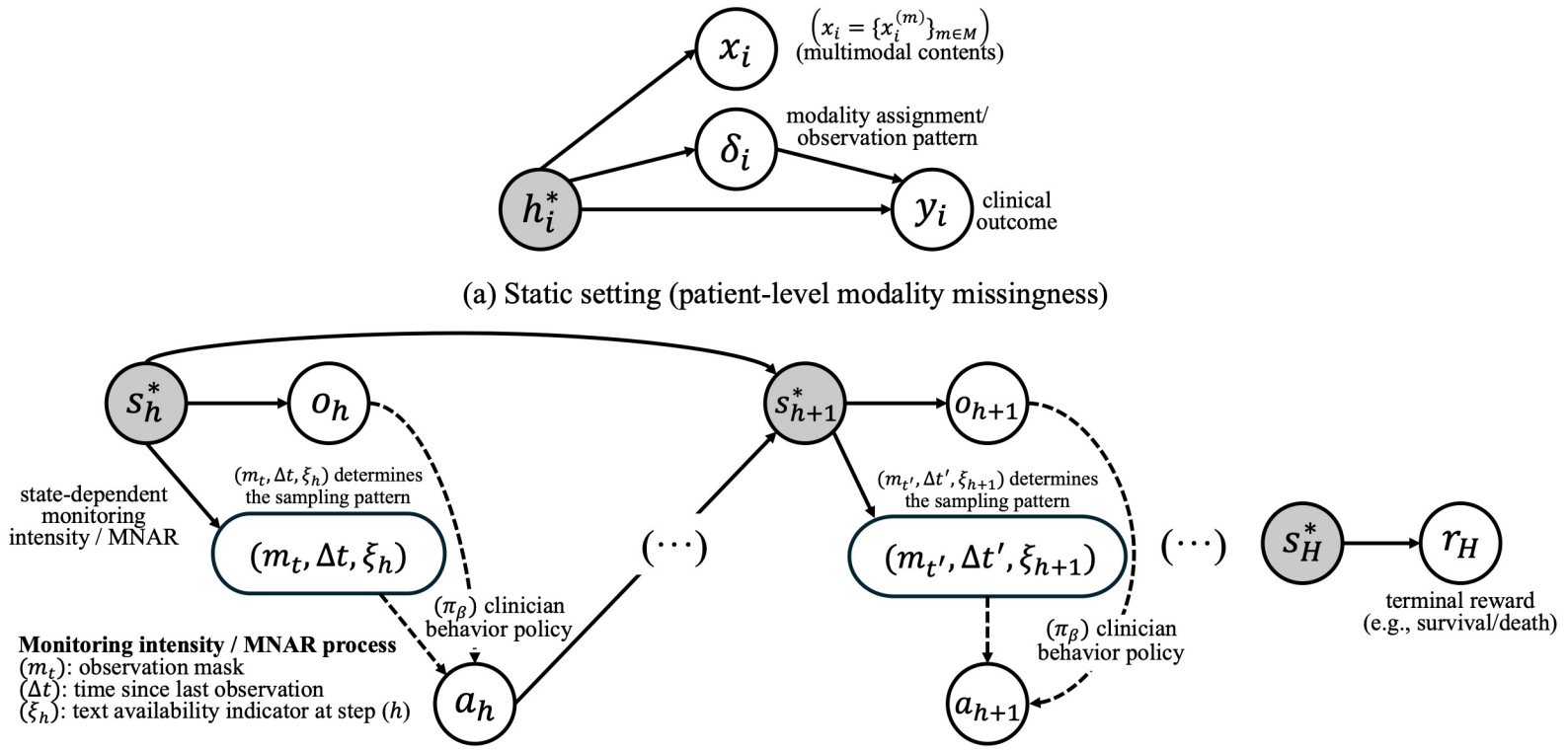
(b) Dynamic setting (step-level monitoring intensity variation)

**The Unified Causal Framework**

- Latent health $h_i^*$ → multimodal contents $x_i$ (what data is generated)
- Latent health $h_i^*$ → observation pattern $\delta_i$ (which modalities the clinician ordered)
- Latent health $h_i^*$ → clinical outcome $y_i$ (the standard path)
- Critical extra edge: $\delta_i \rightarrow y_i$ directly (more tests → earlier intervention → changed outcome)

# The Unified Causal Framework

## Dynamic Setting (step-level monitoring intensity)




(a) Static setting (patient-level modality missingness)

(b) Dynamic setting (step-level monitoring intensity variation)

**The Unified Causal Framework**

- Latent state $s_h^*$ → observations $o_h$ and monitoring intensity $(m_t, \Delta t, \xi_h)$
- $(s_h^*, a_h)$ → $s_{h+1}^*$ *(treatment changes the patient's state)*
- Terminal state $s_H^*$ → reward $r_H$ *(survival or death)*

Overall, we treat observation process variables as first-class modeling objects, not nuisance factors to be imputed away.

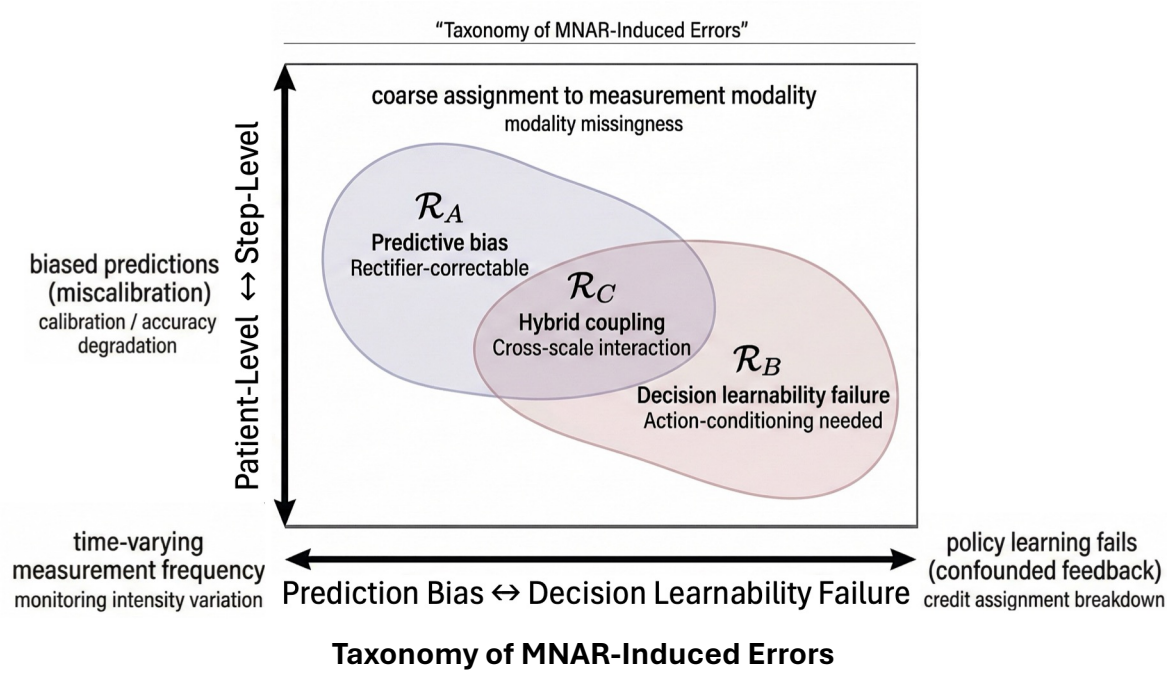
# Two Tasks, Two Different Requirements

| | **Prediction** | **Decision-Making** |
|---|---|---|
| Goal | Estimate outcome probability | Optimize treatment policy |
| Representation need | Prediction-sufficient: $Y \perp X \mid h$ | Control-sufficient: state must encode how actions shape future states |
| MNAR failure mode | Pattern-specific calibration error | Credit assignment breakdown |
| Key Challenge | Residual bias even after good representation | Policy gradient cannot reach early actions |

*A representation that is prediction-sufficient may fail to be control-sufficient.*

# Three Ways MNAR Can Break Your Model

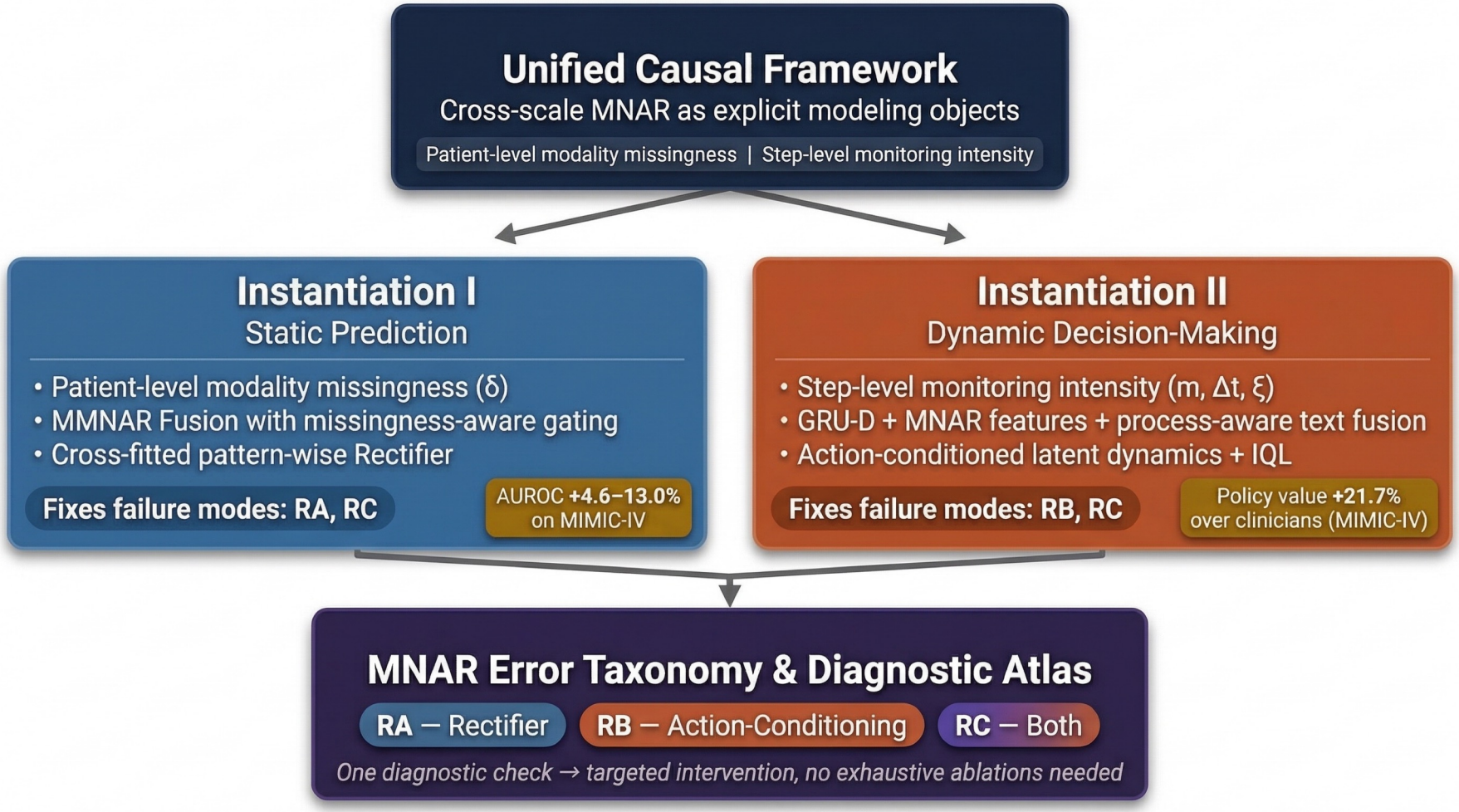
| Regime | What breaks | Diagnostic signal | Fix |
|---|---|---|---|
| $R_A$ — Predictive Bias | Miscalibrated predictions | Elevated ECE per pattern | Rectifier |
| $R_B$ — Decision Learnability Failure | Policy fails; prediction looks fine | Near-zero ActSens | Action-conditioned dynamics |
| $R_C$ — Hybrid | Both degrade | Elevated ECE + high Δdissoc | both |



**Taxonomy of MNAR-Induced Errors**

***Why this taxonomy matters:*** *One diagnostic check before deployment tells you which fix to apply — no exhaustive ablations needed.*

# The Roadmap: One Framework, Two Paths



# Two Instantiations: Prediction and Decision-Making

From representation learning to prediction and decision-making

# Instantiation I: Static Prediction under Patient-Level Modality Missingness

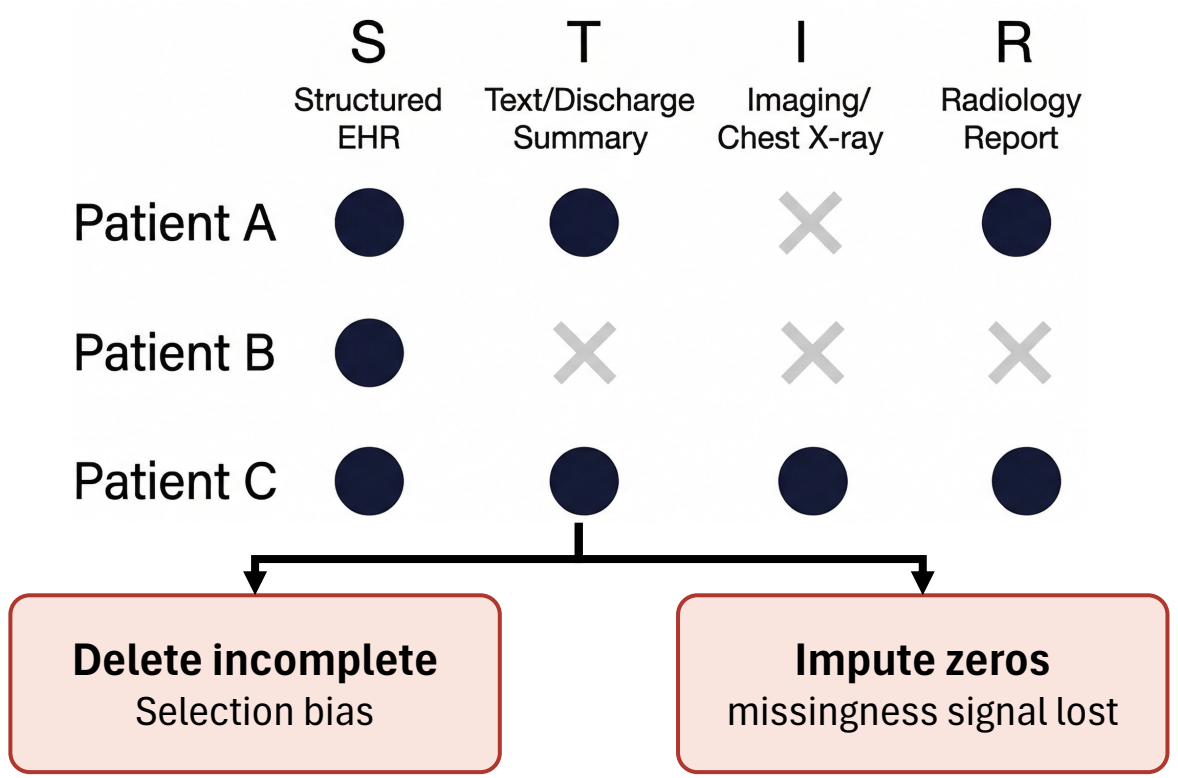
**Setting:** Episode-level prediction from multimodal ICU records — modality availability varies across patients.

**Four modalities:**

- S — Structured EHR (labs, vitals, demographics) · always present
- T — Discharge summaries · 75.5% of patients
- I — Chest X-rays · 26.1% of patients
- R — Radiology reports · 85.0% of patients

**Three prediction targets** (binary classification):

- 30-day unplanned hospital readmission
- Post-discharge ICU admission (within 90 days)
- In-hospital mortality



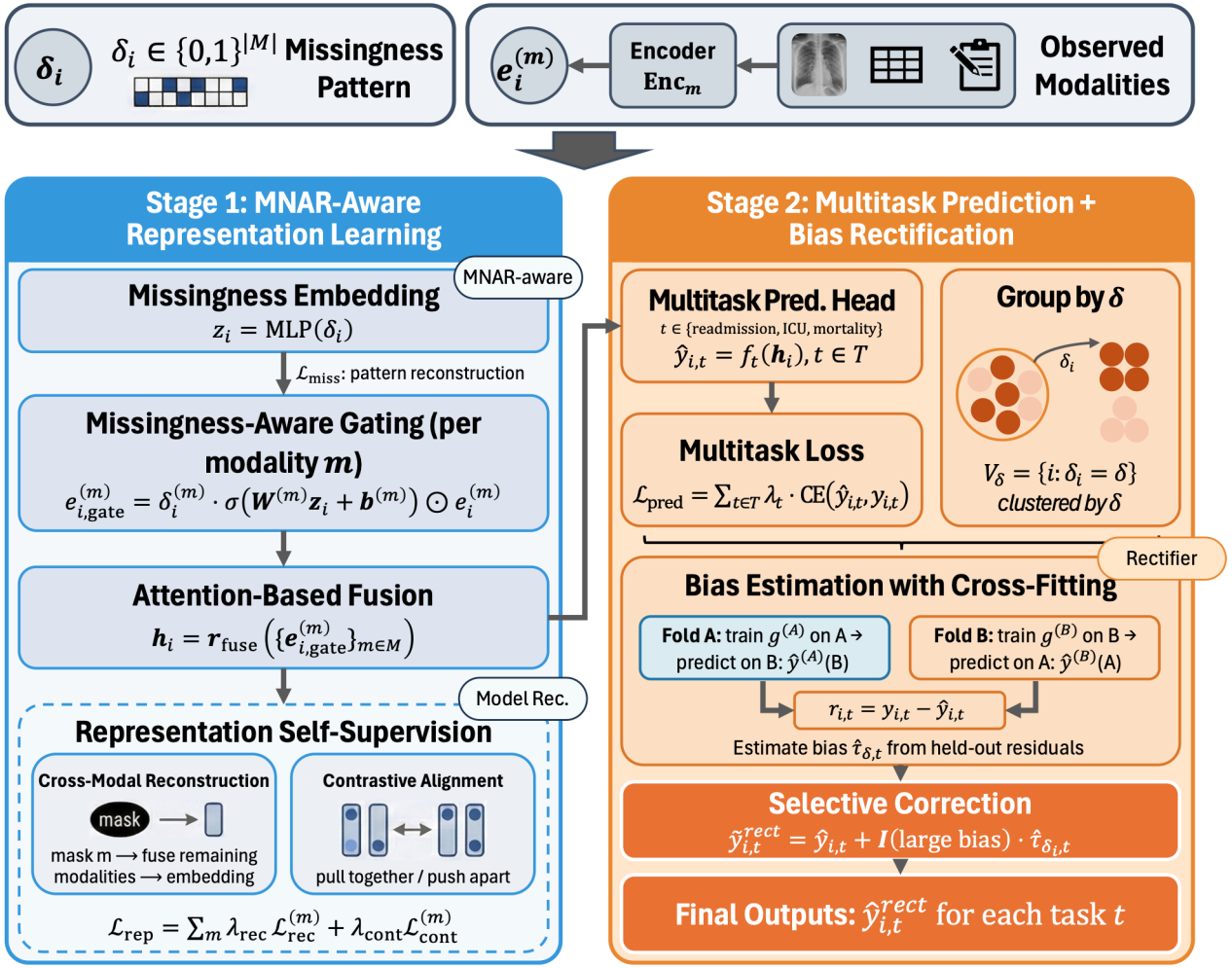
# Instantiation I: Two-Stage MMNAR Pipeline

## Stage 1 — Learn a Missingness-Aware Representation

- Encode who is missing ($\delta_i$) as an explicit feature vector — not a hole to fill
- Gate each modality's contribution by the overall missingness pattern
- Self-supervision: reconstruct each modality from the others — forces the representation to be informationally complete

## Stage 2 — Predict & Correct

- Predict all three outcomes from the shared representation $h_i$
- Rectifier: estimate residual bias per missingness pattern on held-out data; correct at inference time

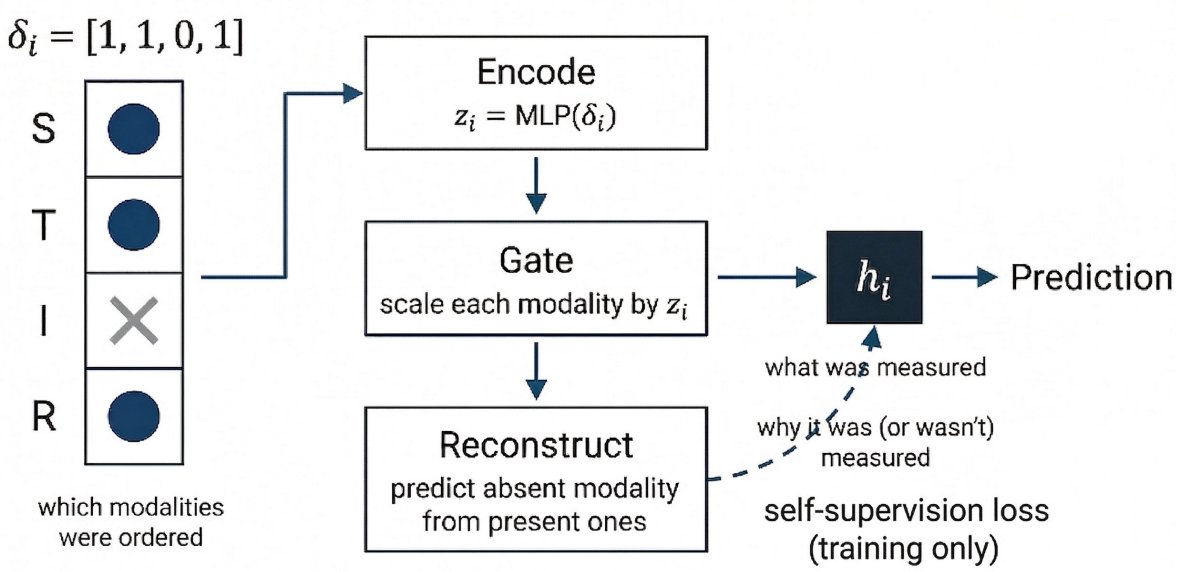


# How Static Missingness Becomes Signal: The $\delta_i$ Pathway

The pattern of which modalities are present is itself a clinical signal — sicker patients get more tests.

- **Encode the pattern** — $\delta_i \in \{0,1\}^4$ is embedded directly: $z_i = \mathrm{MLP}(\delta_i)$, giving the model an explicit representation of who was tested
- **Gate each modality** — each modality's contribution is weighted by $z_i$, so the representation knows why a modality is present or absent
- **Self-supervise completeness** — cross-modal reconstruction forces $h_i$ to retain the information that missingness carries, beyond just the observed values

**Result:** $h_i$ encodes not just what was measured, but what the measurement pattern reveals about the patient's underlying health state.



# The Rectifier: Correcting Residual Missingness Bias

Even after learning a good representation, one causal path remains unblocked: the observation pattern $\delta_i$ directly affects outcomes beyond what $h_i$ captures.

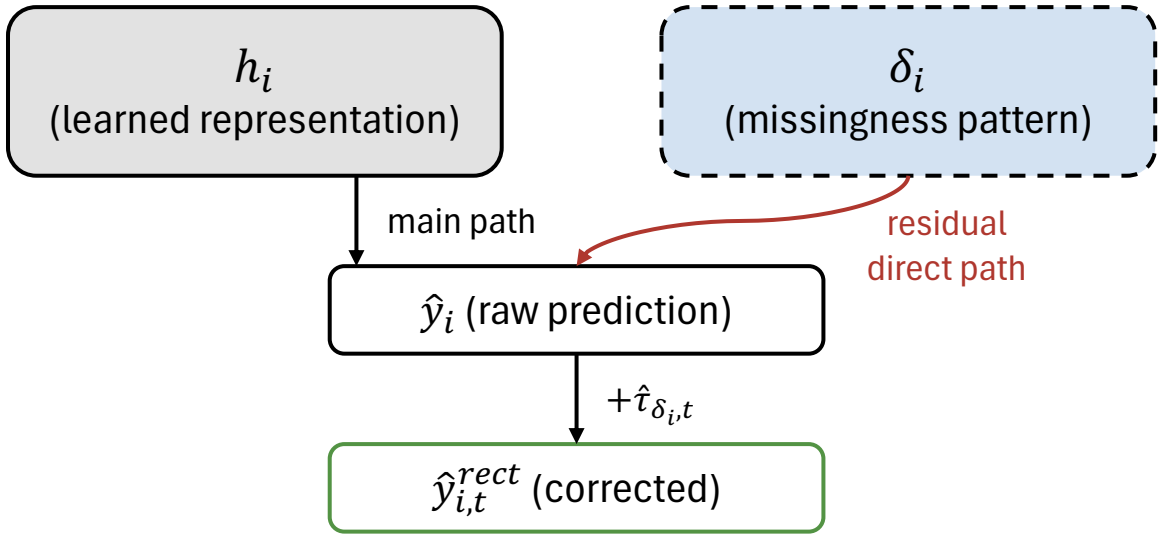
*Example: ordering a chest X-ray is itself a clinical act — it may trigger further intervention, independent of what the image shows.*

## How it works (3 steps):

- Group patients by missingness pattern δ
- Estimate residual bias per pattern on held-out data
- Apply correction at inference if bias exceeds threshold κ

## Key properties:

- ICU admission AUROC: +9.3% (MIMIC-IV ablation)
- Lightweight and overfitting-proof via cross-fitting



# Instantiation II: Offline Decision-Making for Sepsis under Step-Level MNAR

**Setting:** Learn a treatment policy for sepsis management from historical ICU trajectories — no randomized trial, no online interaction.

**Data (MIMIC-IV):**

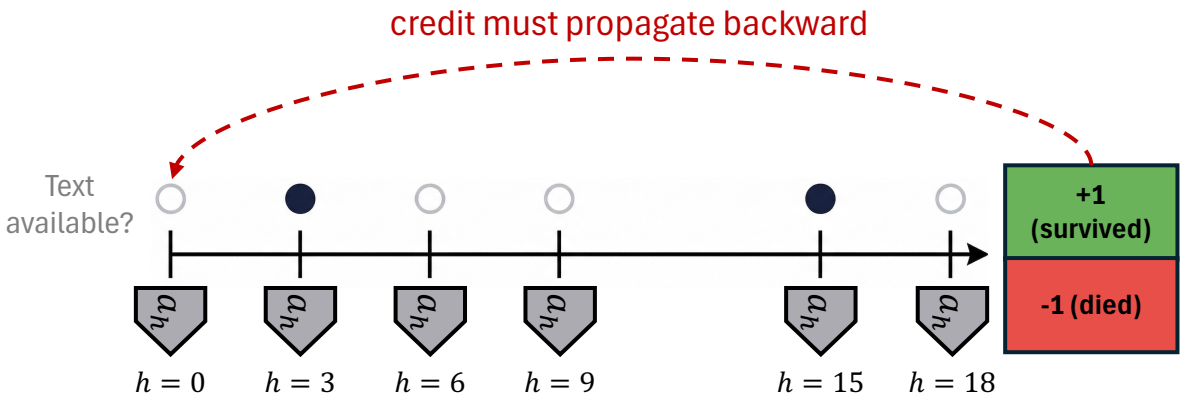
- 32,837 ICU stays · 28,419 unique patients · 11.8% in-hospital mortality
- 72-hour window → **H = 18 decision steps** at 4-hour intervals
- Text is step-level sparse: radiology available at only **83.6%** of steps; microbiology at **45.5%**

**Action space:** 3 × 3 = 9 discrete actions (IV fluids × vasopressors)

**Reward:** Sparse and terminal — +1 survival / −1 death; zero at all intermediate steps

**Challenge:**

*Reward only arrives at step 18. But treatment decisions start at step 1. How does the model know which early action deserved credit for the final outcome?*



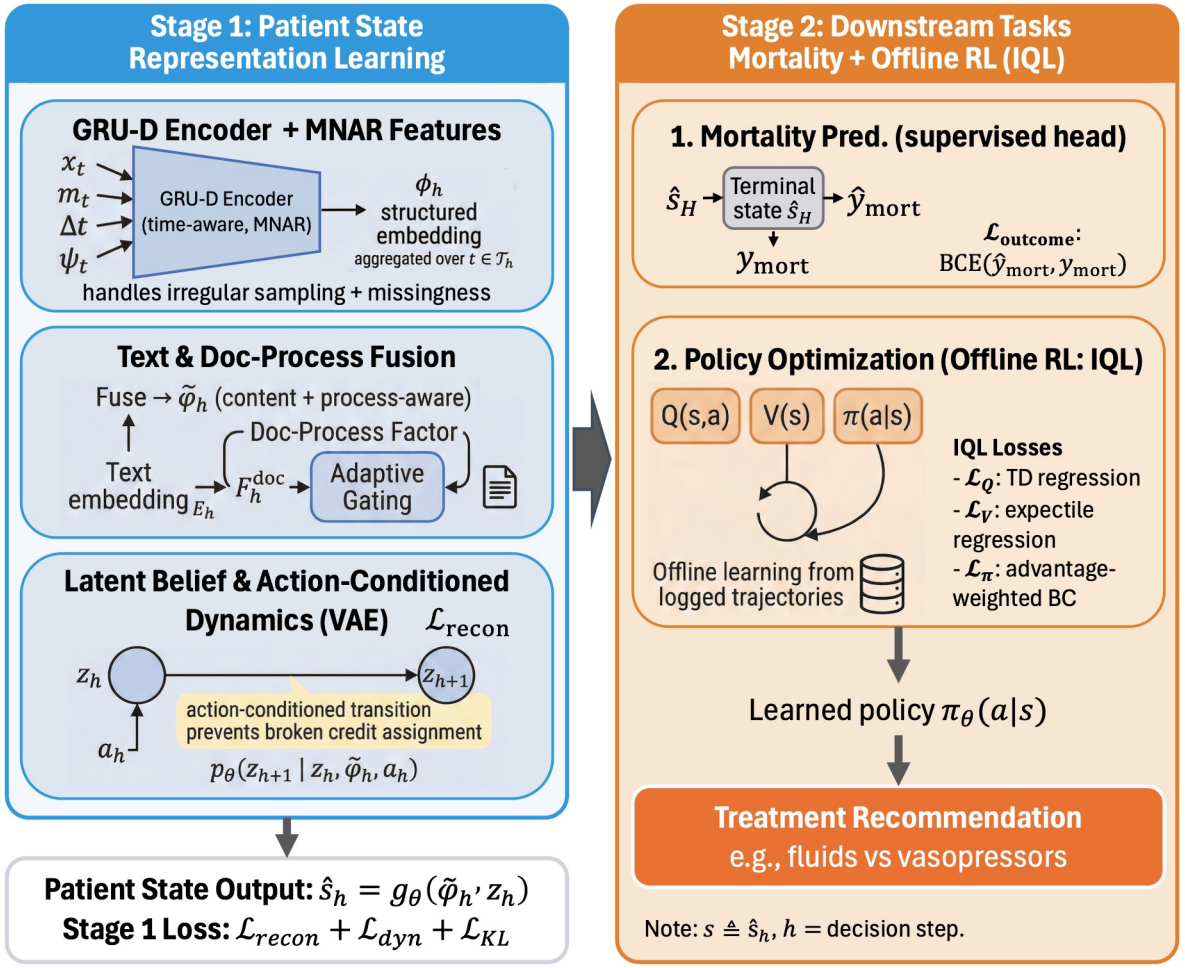
# Instantiation II: MNAR-Aware State Learning for Offline Control

## Stage 1 — Learn Decision-Step Patient States

- **Structured encoding (GRU-D):** Encodes not just values, but when and how often each variable was measured
- **Process-aware text fusion:** Stale or sparse documentation is discounted even when text is present
- **Action-conditioned dynamics:** Actions are wired into state transitions — enabling credit to flow backward to early decisions

## Stage 2 — Optimize an Offline Policy

- Learns value functions from historical data without querying actions outside the dataset → outputs $\pi(a|s)$



# How Temporal Missingness Becomes Signal: The $(m_t, \Delta t, \xi_h)$ Pathway

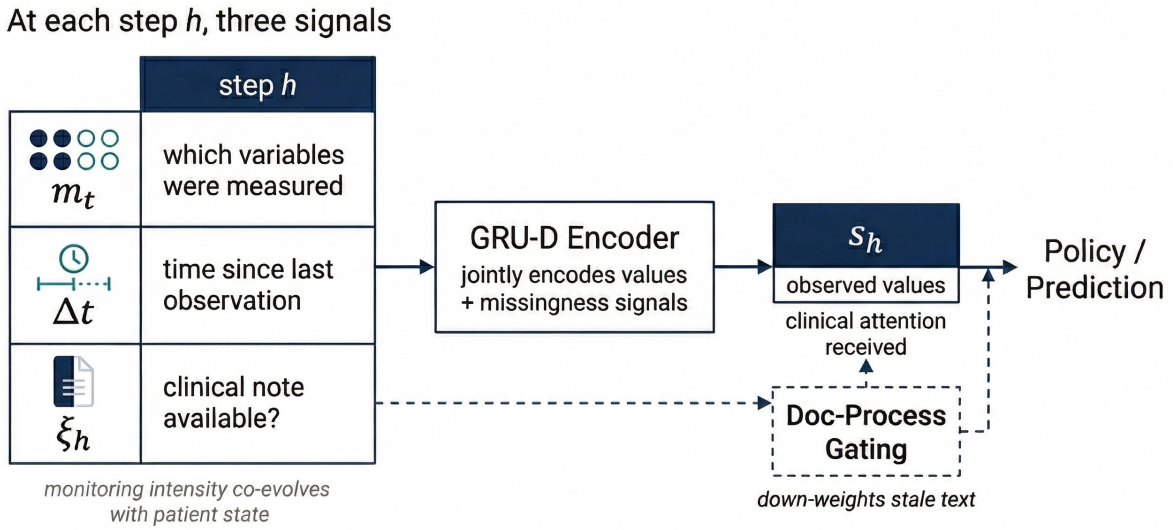
At each decision step, how a patient is monitored reveals as much as what is measured.

- $m_t$ **(observation mask)** — which variables were measured at this step; sparse monitoring signals stability
- $\Delta t$ **(time since last observation)** — longer gaps indicate the clinician judged the patient stable enough to skip
- $\xi_h$ **(text availability)** — whether clinical notes exist at this step; absence is itself informative

GRU-D encodes all three jointly with the observed values — monitoring intensity co-evolves with patient state.

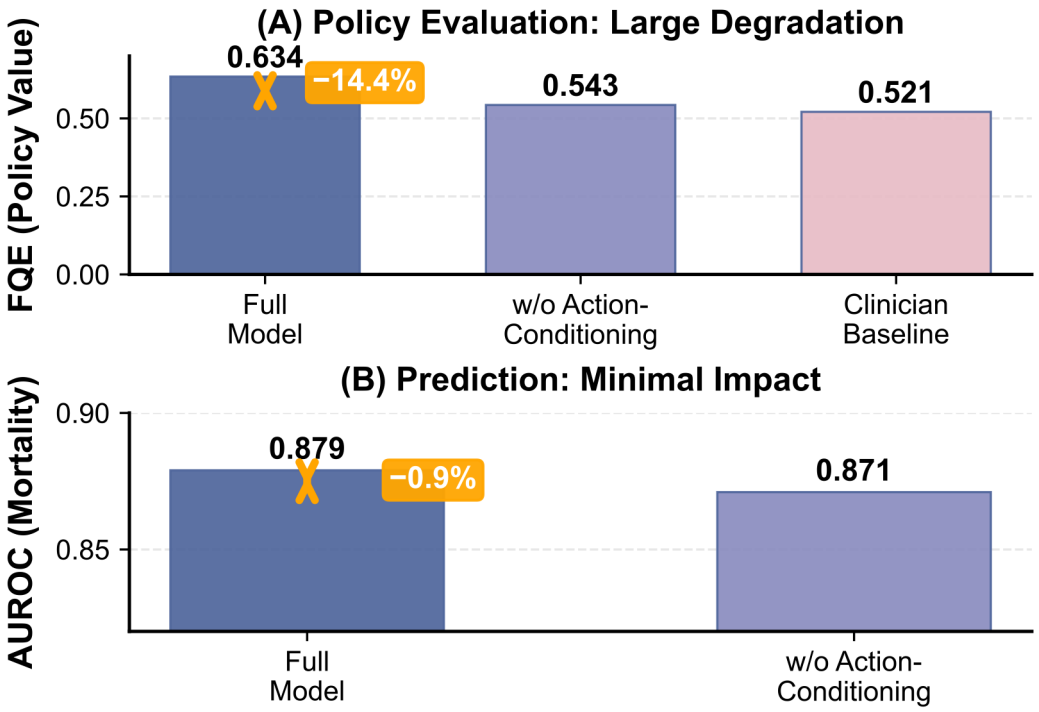
**Doc-Process Gating:** stale or sparse documentation is down-weighted at inference, so the model is not misled by the timing of text.

**Result:** $s_h$ encodes not just vitals and labs, but the full clinical attention the patient received.



# Why Action-Conditioning Is Non-Negotiable: The RB Dissociation

**The experiment:** Remove action-conditioning from latent dynamics. What breaks?




*Prediction barely changes. Policy collapses.*

**Why?** Without action-conditioning, the reward gradient cannot reach early actions — credit assignment is severed by design.

Action Sensitivity (ActSens) confirms:

- Without action-conditioning: 0.023 (near zero)
- With action-conditioning: 0.847

***Good Prediction ≠ Good Policy. This is not an engineering quirk — it is a structural property.***

# Results and Takeaways

From modeling ideas to empirical evidence

# Results I: Static Prediction under MMNAR

**Our method outperforms all 13 baselines on both MIMIC-IV and eICU.**

| Task | Best Baseline | Ours | Gain |
|---|---|---|---|
| ICU Admission (MIMIC-IV) | 0.869 (DrFuse) | **0.982** | **+13.0%** |
| 30-Day Readmission (MIMIC-IV) | 0.799 (MUSE+) | **0.866** | **+8.4%** |
| In-Hospital Mortality (MIMIC-IV) | 0.905 (MUSE+) | **0.947** | **+4.6%** |

**Some insights:**

- **Robustness under distribution shift:** Under severe missingness shift (KL = 1.0), the AUROC gap for readmission widens from **+6.7% to +12.5%** — MMNAR-aware models become more advantageous when patterns shift.
- **It means** if the previous best model caught 87 out of 100 ICU admissions, our model catches 98.

# Results II: Offline Decision-Making under Dynamic MNAR

**Our policy outperforms clinician behavior and all RL baselines on both datasets.**

| Method | MIMIC-IV FQE | vs. Clinician | eICU FQE | vs. Clinician |
|---|---|---|---|---|
| Clinician (BC) | 0.521 | — | 0.534 | — |
| CQL | 0.418 | −19.8% | 0.408 | — |
| Tabular Q (AI Clinician) | 0.491 | −5.8% | 0.478 | — |
| **Ours (IQL)** | **0.634** | **+21.7%** | **0.604** | **+13.1%** |

Outperforms AI Clinician by +29.1%, CQL by +51.7%

All four OPE estimators (FQE, WIS, PDIS, DR) confirm improvement; even the most conservative lower bound (WIS: 0.571) exceeds the clinician baseline (0.521)

# Results II: Offline Decision-Making under Dynamic MNAR

**Gains concentrate where it matters most — the sickest patients:**

| SOFA Severity | Clinician FQE | Ours FQE | Gain |
|---|---|---|---|
| Low (0–5) | 0.684 | 0.756 | +10.5% |
| Medium (6–10) | 0.423 | 0.548 | +29.6% |
| **High (>10)** | **0.187** | **0.319** | **+70.6%** |

The AI helps most where clinicians struggle most — the most critically ill patients in the ICU.

# Robustness and Cross-Institutional Generalization

### Part 1 — Robustness to Missingness Pattern Shift (Static Setting)

| Shift Severity | Ours | MUSE+ | Gap |
|---|---|---|---|
| None (original) | 0.866 | 0.799 | +6.7% |
| Mild (KL = 0.2) | 0.854 | 0.771 | +8.3% |
| Moderate (KL = 0.5) | 0.838 | 0.734 | +10.4% |
| **Severe (KL = 1.0)** | **0.812** | **0.687** | **+12.5%** |

### Part 2 — Cross-Institutional Generalization (Dynamic Setting)

| Protocol | Train → Test | FQE | vs. Clinician |
|---|---|---|---|
| In-distribution | MIMIC-IV → MIMIC-IV | 0.634 | +21.7% |
| **Zero-shot transfer** | **MIMIC-IV → eICU** | **0.568** | **+6.4%** |
| Fine-tuned transfer | MIMIC-IV + eICU FT → eICU | 0.594 | +11.2% |

# 3 Practical Takeaways for Clinical AI Developers

## Construct MNAR features explicitly

- Encode missingness as signal, not noise
- Removing MNAR features: FQE ↓13.4%, AUROC ↓3.2%

## Apply the diagnostic atlas before deployment

- ECE elevated → Rectifier
- High $\Delta_{\text{dissoc}}$ + low ActSens → action-conditioned dynamics

## Distinguish prediction objectives from decision-support objectives

- Good AUROC ≠ good policy
- If decision support is the goal: use FQE, not just AUROC

# Conclusion

From specific methods to broader implications

# Contributions at a Glance

**1. Unified Cross-Scale MNAR Framework.** Formalized patient-level modality assignment and step-level monitoring intensity as explicit causal variables within a structural causal model.

**2. Instantiation I — Static Multimodal Prediction.** Missingness-aware fusion + cross-fitted rectifier → AUROC gains of **+4.6% to +13.0%** over 13 baselines (MIMIC-IV); **+0.5% to +13.7%** on eICU.

**3. Instantiation II — Dynamic Sequential Decision-Making.** MNAR-aware encoder + action-conditioned latent dynamics + IQL → **+21.7%** over clinician behavior (MIMIC-IV); **+70.6%** for high-acuity patients (SOFA > 10).

**4. MNAR Error Taxonomy & Diagnostic Atlas.** Three failure regimes mapped to observable diagnostic signals and targeted interventions — enabling prospective deployment monitoring.

# Future Directions

**1. Sharper Causal Identification.** Current framework corrects for δ → y but does not identify the mechanism. Instrumental variable methods could enable counterfactual reasoning about the observation process itself.

**2. Active Observation Strategies.** Move from passively handling missingness to learning which measurements to acquire at each step — jointly optimizing observation cost against treatment benefit.

**3. Multi-Site Federated Learning & Prospective Trials.** Zero-shot transfer recovers only 72–81% of in-distribution performance. Federated learning + clinician-in-the-loop trials are needed to close the gap between OPE estimates and real-world impact.

# Acknowledgments


- Honors Thesis Advisor: Prof. Ruoxuan Xiong
- Honors Committee: Prof. Ali Emami, Prof. Manuela Manetta
- Collaborator: Ziwen Pan
- Faculty: Dr. Frank Gaertner, Prof. Azra Ismail, Prof. Talea L. Mayo
- Institutional Support
  - The Department of Data and Decision Sciences (𝔻Sci)
  - The Undergraduate Research Programs (URP)
- Family & Friends
- Thank you for 3 years of guidance, challenge, and encouragement.

*Names listed alphabetically within each category.*

# Thank You

*"The incompleteness of clinical data is not merely a technical inconvenience to be overcome before the real modeling can begin.*

*It is a reflection of the human process through which knowledge about patients is gathered, filtered, and recorded — a process that is itself a rich source of signal about patient health and clinical priorities."*

**Models that model missingness outperform models that merely tolerate it.**

# Questions & Discussion

Thank you for listening!