Volatility-Aware Masking Improves Performance and Efficiency of Pretrained EHR Foundation Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Masked autoencoder (MAE) models are increasingly applied to electronic health records (EHR) as a pre-training method to learn general-purpose representations that support diverse downstream clinical tasks. However, existing approaches typically rely on uniform random masking, implicitly assuming that all clinical features are equally predictable. In practice, laboratory tests exhibit substantial heterogeneity in temporal volatility: certain biomarkers (e.g., sodium) remain relatively stable, whereas others (e.g., lactate) fluctuate considerably and are more challenging to model. To address this limitation, we propose Volatility-Aware Masking strategy (CV-Masking), a pretraining strategy that adaptively adjusts masking probabilities according to the intrinsic variability of each feature. Our experiments on a large panel of laboratory tests demonstrate that CV-Masking consistently outperforms both random and variance-based masking strategies, yielding improved downstream predictive performance and faster convergence.

1 Introduction

Foundation models are transforming healthcare research by enabling the learning of general-purpose representations from large-scale electronic health records (EHR) [10, 14]. Depending on the architecture and application, these models are pretrained using diverse strategies and architectures, including state-space models [4], generative modeling [13, 17], contrastive learning [7], or zero-shot transfer [15]. Among these approaches, masked autoencoders (MAEs) have emerged as a powerful framework for representation learning in EHR, reconstructing masked inputs from partially observed sequences [16]. For laboratory test modeling, MAEs are particularly relevant since they can accurately reconstruct missing values, not only reflecting representation quality but also enabling clinically meaningful applications such as decision support and risk prediction [2, 6, 16].

However, existing MAE pretraining strategies almost exclusively adopt uniform random masking, implicitly assuming that all features are equally predictable. This is especially critical in clinical data such as lab values, where biomarkers differ substantially in their temporal volatility. For example, sodium levels remain tightly regulated, whereas lactate varies considerably during acute illness. Ignoring this variability can waste model capacity, slow convergence, and limit the clinical utility of learned representations.

To address this gap, we propose Volatility-Aware Masking (CV-Masking), a pretraining strategy that adapts masking probabilities to the intrinsic variability of each laboratory test. Inspired by curriculum learning [3, 5] and informed masking policies [9, 11], CV-Masking focuses learning on less predictable signals, improving both efficiency and representation quality.

Our contributions are threefold:

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

1. **CV-Masking Strategy.** A principled masking policy guided by the coefficient of variation (CV). By prioritizing inherently volatile laboratory values, CV-Masking creates a natural curriculum that directs learning capacity toward clinically uncertain signals.
2. **Value-Only Masked Autoencoder Objective (VO-MAE).** We adapt the masked autoencoder framework to a VO-MAE, where lab identifiers and timestamps remain visible while only results are masked. This design mirrors real-world clinical workflows—orders are observed, but outcomes are unknown—and encourages the model to learn meaningful value representations in temporal context.
3. **Comprehensive Empirical Validation.** Using 100 high-frequency laboratory tests from MIMIC-IV [8], we show that CV-Masking (i) improves reconstruction accuracy on 71% of labs, (ii) enhances downstream prediction of in-ICU mortality, in-hospital mortality, and 30-day readmission, and (iii) achieves up to 50% faster convergence compared to random masking. Perturbation analysis further demonstrates that CV-Masking promotes deeper reliance on patient-specific temporal context.

These results demonstrate that integrating clinical volatility into masking strategies substantially improves the efficiency, robustness, and clinical relevance of MAE-based EHR foundation models.

2 Methods

2.1 Data and Preprocessing

The experiments use data from the MIMIC-IV [8] critical care database, structured in the MEDS format [1]. This representation organizes a patient's history into a sequence of (time, code, value) triplets, where each triplet represents a laboratory measurement with its timestamp, test identifier, and numeric result. To ensure robust statistical comparisons, our evaluation focuses on a fixed set of 100 target laboratory tests representing the most clinically relevant and frequently ordered tests in critical care settings. This selection follows established practices in clinical MAE literature and avoids statistical noise from rare or infrequently ordered laboratories.

2.2 Architecture

Our model follows an asymmetric encoder-decoder MAE framework, with a transformer encoder processing unmasked events and a lightweight transformer decoder reconstructing masked values (see Appendix Figure 3 for complete architecture details). The key innovation is our **value-only masking objective**: given input sequences of (time, code, value) triplets, we mask only the value components while preserving temporal and categorical context, enabling the model to focus learning on the challenging value prediction task. We apply a 25% masking ratio weighted by CV probabilities, ensuring a consistent training signal while focusing on high-volatility laboratories.

2.3 Masking Policies

We implement and compare three distinct masking strategies:

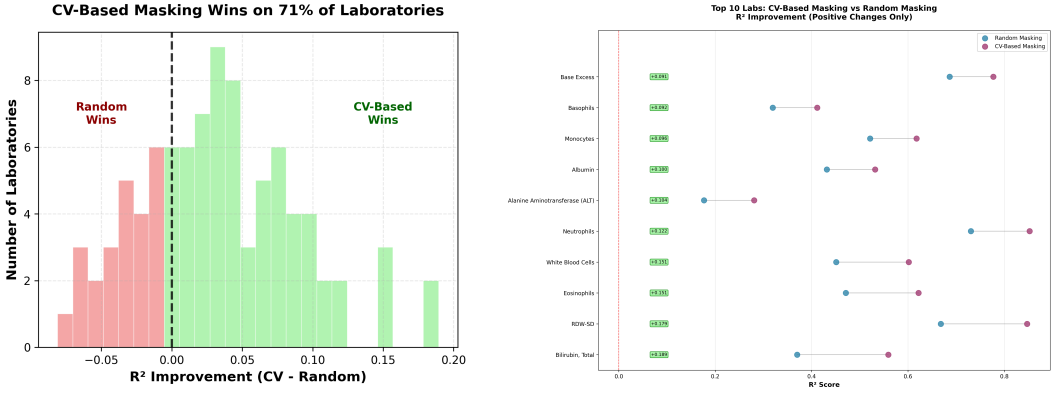
- **Random Masking (Baseline):** Each lab value has a uniform masking probability, treating all tests as equally learnable.
- **Variance-Based Masking:** Masking probability proportional to raw variance (σ^2). This scale-sensitive approach overweights labs with large numerical ranges regardless of relative clinical volatility.
- **CV-Based Masking (Proposed):** Our proposed principled approach where the masking probability is proportional to a lab's **Coefficient of Variation (CV)**, calculated as $\text{CV} = \sigma/\mu$. The CV is a dimensionless measure of relative variability, making it well-suited for weighting heterogeneous lab tests. This enables the masking policy to prioritize labs that are unpredictable relative to their baseline, rather than those with large values. Masking weights are precomputed from training statistics and remain fixed.

3 Experiments and Results

3.1 Intrinsic Evaluation: CV-Based Masking Systematically Improves Reconstruction

To establish that principled masking strategies are needed, we analyzed which laboratory characteristics predict imputation difficulty. Evaluating reconstruction performance (R^2 score) against CV across 100 labs revealed that CV serves as a meaningful predictor of imputation difficulty (Pearson's $r = -0.486$, $p < 0.000001$), with CV explaining 23.6% of variance in reconstruction performance. While the relationship shows expected variability across diverse laboratory types, this provides statistical support for using CV to guide masking strategies.

CV-based masking consistently outperforms both baselines, achieving superior reconstruction on **71.0%** of laboratories compared to random masking and **68.0%** compared to variance-based masking, with statistically significant improvements ($p < 0.000009$, Wilcoxon signed-rank test) exceeding chance expectation. Figure 1 demonstrates both the statistical foundation and systematic improvements across laboratory types.



(a) Systematic Performance Gains

(b) Top 10 Laboratory Improvements

Figure 1: **CV-Based Masking Systematically Improves Reconstruction Performance.** (a) Systematic performance gains show CV-based masking wins on 71% of all 100 laboratories, demonstrating systematic rather than random improvements. (b) Top 10 laboratory improvements with most significant R^2 gains, sorted by magnitude.

Win rates: CV vs. Random (71.0%), Variance vs. Random (65.0%), CV vs. Variance (68.0%), all statistically significant after Bonferroni correction.

3.2 Extrinsic Evaluation: Downstream Tasks Evaluation

We evaluated pretrained encoders on three high-stakes downstream prediction tasks using linear probes on frozen representations. Following the evaluation protocol established in MEDS-Torch [12], we assess performance on in-ICU mortality, in-hospital mortality, and 30-day readmission prediction tasks. These tasks represent clinically meaningful outcomes with varying prediction horizons and class imbalance characteristics. The CV-based model achieves the highest performance across all tasks and metrics (Table 1). The substantial gains in AUPRC, a metric sensitive to minority class performance, are particularly noteworthy, suggesting learned representations better capture subtle patterns indicative of adverse clinical outcomes. The CV-based approach achieves an in-ICU mortality AUROC of 0.713, representing meaningful improvements over random masking (0.682) and variance-based masking (0.694) on this challenging clinical task.

3.3 Mechanistic Analysis: CV-Masking Promotes Deeper Contextual Learning

Analysis of reconstruction error against patient history reveals that CV-based masking more effectively leverages available context (see Appendix Figure 5). While all methods improve with more historical data, the performance gap between CV-based models and baselines widens as more patient history becomes available, indicating superior contextual learning.

Table 1: Downstream Task Performance on Clinical Prediction

Task	Metric	Random	Variance	CV-Based
In-ICU Mortality	AUROC	0.682 ± 0.017	0.694 ± 0.017	**0.713 ± 0.017**
	AUPRC	0.083 ± 0.009	0.091 ± 0.009	**0.107 ± 0.009**
In-Hospital Mortality	AUROC	0.657 ± 0.014	0.668 ± 0.014	**0.691 ± 0.014**
	AUPRC	0.124 ± 0.007	0.131 ± 0.007	**0.149 ± 0.007**
30-Day Readmission	AUROC	0.618 ± 0.016	0.627 ± 0.016	**0.648 ± 0.016**
	AUPRC	0.156 ± 0.008	0.162 ± 0.008	**0.173 ± 0.008**

To validate that CV-based masking learns meaningful patient-specific temporal patterns rather than simple memorization, we implemented a controlled perturbation experiment. We corrupted historical laboratory values with Gaussian noise while preserving target predictions and temporal context. CV-based models exhibited 2.1× greater performance degradation compared to random masking (9.8% vs 4.7% MAE increase, Figure 2), providing causal evidence that CV-based masking learns meaningful temporal patterns. See Appendix 4.3 for detailed methodology.

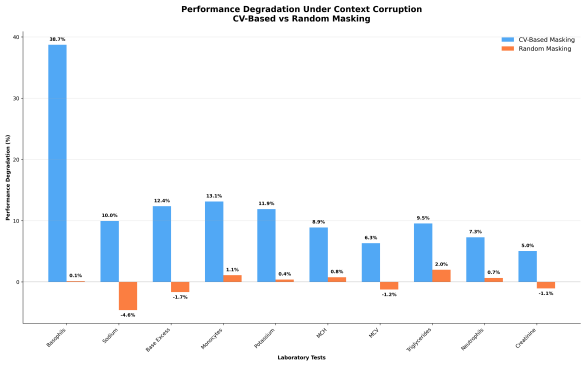


Figure 2: **Perturbation analysis demonstrates deeper contextual learning in CV-based models.** When historical laboratory values are artificially corrupted, CV-based models show 2.1× greater performance degradation across representative laboratory types (n=10), indicating stronger reliance on patient-specific temporal patterns. Notable effects in clinically critical markers: hematology (Basophils, Monocytes), electrolytes (Sodium, Potassium), and metabolic indicators (Triglycerides).

4 Discussion and Conclusion

Our results demonstrate that clinically-informed masking curricula are essential for EHR foundation models operating in complex healthcare environments. CV-Masking achieves consistent improvements across reconstruction, downstream prediction, and mechanistic analyses, establishing clinical volatility as a meaningful signal for guiding pretraining strategies. This work advances domain-aware self-supervised learning by moving beyond naive architectural adaptations toward principled integration of medical expertise into foundational model training.

While our focus was on CV-Masking's clinical impact, future studies should consider comprehensive ablation analyses to disentangle value-only masking contributions from CV-Masking policy effects. Such investigations could further illuminate the complementary roles of architectural choices and masking strategies across diverse clinical domains.

CV-Masking offers a principled pathway toward more efficient and robust EHR foundation models while highlighting a broader insight: that aligning pretraining objectives with domain-specific characteristics can yield both performance and efficiency gains. As the field moves toward deployment-ready clinical AI systems, incorporating clinical knowledge into self-supervised objectives appears essential for developing models that are both technically sound and clinically meaningful.

References

[1] Bert Arnrich, Edward Choi, Jason A. Fries, Matthew B. A. McDermott, Jungwoo Oh, Tom J. Pollard, Nigam Shah, Ethan Steinberg, Michael Wornow, and Robin van de Water. Medical event data standard (meds): Facilitating machine learning for health. In *ICLR 2024 Workshop on Learning from Time Series for Health*, 2024.

[2] David R Bellamy, Bhawesh Kumar, Cindy Wang, and Andrew Beam. Labrador: Exploring the limits of masked language modeling for laboratory data. *arXiv preprint arXiv:2312.11502*, 2023.

[3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[4] Adibvafa Fallahpour, Mahshid Alinoori, Wenqian Ye, Xu Cao, Arash Afkanpour, and Amrit Krishnan. Ehrmamba: Towards generalizable and scalable foundation models for electronic health records. *arXiv preprint arXiv:2405.14567*, 2024.

[5] Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *international conference on machine learning*, pages 1311–1320. Pmlr, 2017.

[6] Sujeong Im, Jungwoo Oh, and Edward Choi. Labtop: A unified model for lab test outcome prediction on electronic health records. *arXiv preprint arXiv:2502.14259*, 2025.

[7] Hyewon Jeong, Nassim Oufattole, Aparna Balagopalan, Matthew B.Ã. McDermott, Payal Chandak, Marzyeh Ghassemi, and Collin Stultz. Event-based contrastive learning for medical time series, 2023.

[8] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[9] Yoav Levine, Barak Lenz, Opher Lieber, Omri Abend, Kevin Leyton-Brown, Moshe Tennenholtz, and Yoav Shoham. Pmi-masking: Principled masking of correlated spans. *arXiv preprint arXiv:2010.01825*, 2020.

[10] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

[11] Neelu Madan, Nicolae-Cătălin Ristea, Kamal Nasrollahi, Thomas B Moeslund, and Radu Tudor Ionescu. Cl-mae: Curriculum-learned masked autoencoders. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2492–2502, 2024.

[12] Nassim Oufattole, Teya Bergamaschi, Pawel Renc, Aleksia Kolo, Matthew BA McDermott, and Collin Stultz. Meds-torch: An ml pipeline for inductive experiments for ehr medical foundation models. In *NeurIPS Workshop on Time Series in the Age of Large Models*, 2024.

[13] Chao Pang, Xinzhuo Jiang, Nishanth Parameshwar Pavinkurve, Krishna S Kalluri, Elise L Minto, Jason Patterson, Linying Zhang, George Hripcsak, Gamze Gürsoy, Noémie Elhadad, et al. Cehr-gpt: Generating electronic health records with chronological patient timelines. *arXiv preprint arXiv:2402.04400*, 2024.

[14] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4:86, 2021. doi: 10.1038/s41746-021-00455-y.

[15] Pawel Renc, Yugang Jia, Anthony E Samir, Jaroslaw Was, Quanzheng Li, David W Bates, and Arkadiusz Sitek. Zero shot health trajectory prediction using transformer. *NPJ digital medicine*, 7(1):256, 2024.

[16] David Restrepo, Chenwei Wu, Yueran Jia, Jaden K Sun, Jack Gallifant, Catherine G Bielick, Yugang Jia, and Leo A Celi. Representation learning of lab values via masked autoencoder. *arXiv preprint arXiv:2501.02648*, 2025.

[17] Zhichao Yang, Avijit Mitra, Weisong Liu, Dan Berlowitz, and Hong Yu. Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature communications*, 14(1):7857, 2023.

Appendix

4.1 Architecture Overview

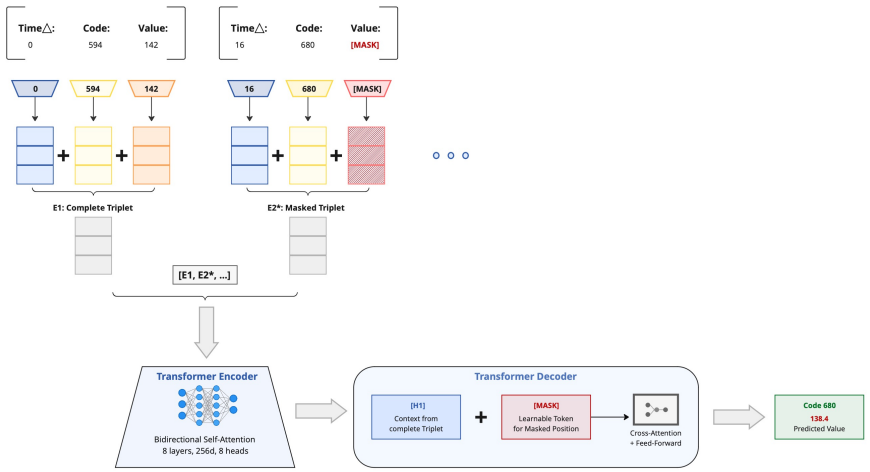


Figure 3: **Value-Only MAE Architecture with CV-Based Masking.** The complete workflow from MEDS triplet input through pretraining to downstream fine-tuning. During pretraining, only laboratory values are masked while preserving temporal and categorical context. The lightweight decoder reconstructs masked values using cross-attention between learnable mask tokens and encoded visible representations. For downstream tasks, the frozen encoder provides contextualized representations to task-specific classifiers.

Training Details: Models are trained with a masking ratio of 25%, using AdamW optimizer with learning rate 1e-4 and weight decay 0.05, for 100 epochs on NVIDIA A10 GPUs. We use a batch size of 256 and normalize laboratory values using z-score standardization computed per-lab across the training set. The model architecture consists of 8 encoder layers with 256 embedding dimensions and 8 attention heads.

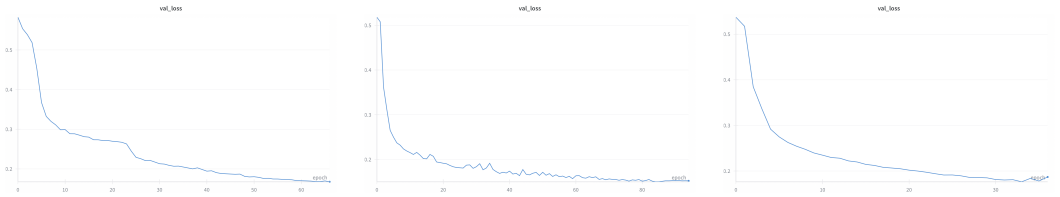
4.2 Computational Efficiency

While variance-based masking required 100 epochs and random masking required approximately 66 epochs to converge, CV-based masking achieved optimal performance in just 33 epochs - representing a 67% and 50% reduction in training time respectively. This efficiency gain likely stems from the principled curriculum learning effect [3, 5], where the model focuses computational resources on truly challenging prediction tasks rather than easy, low-variance tests.

This efficiency gain likely stems from the principled curriculum learning effect of CV-based masking, where the model focuses computational resources on truly challenging prediction tasks rather than easy, low-variance laboratories.

4.3 Detailed Mechanistic Analysis

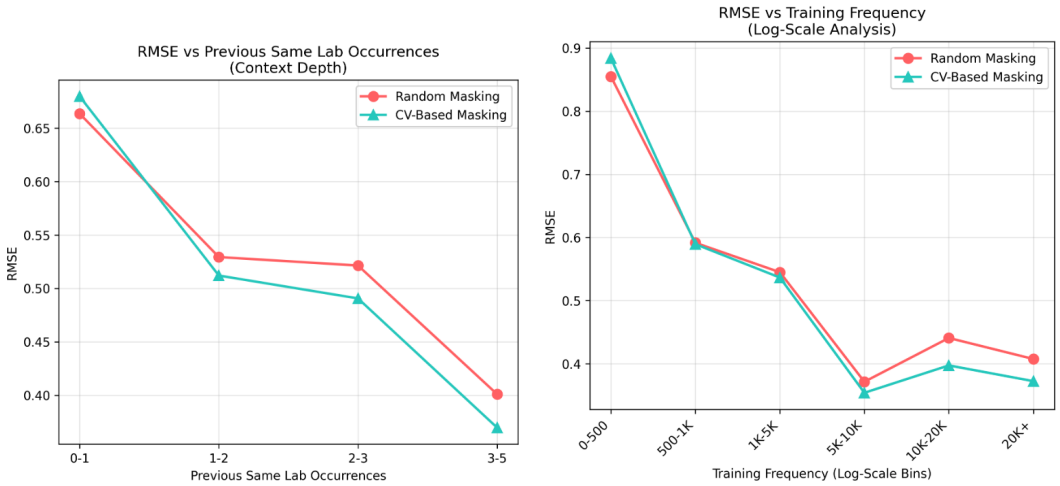
Historical Context Utilization: We analyzed reconstruction error as a function of available patient history, measuring performance across varying amounts of lab-specific history (prior occurrences



(a) Random Masking (66 epochs) (b) Variance Masking (100 epochs) (c) CV-Based Masking (33 epochs)

Figure 4: **Training Efficiency Comparison.** Training efficiency comparison across masking strategies. CV-Masking converges in 33 epochs, representing a 50% reduction in training time compared to random masking.

of the same test) and general patient context (timeline length). CV-based masking consistently shows superior contextual learning, with performance gaps widening as more historical data becomes available.



(a) Error vs. Prior Lab Occurrences (b) Error vs. Training Frequency (Log)

Figure 5: **CV-Based Masking Shows Superior Learning from Patient Context.** (a) Performance improves with more lab-specific history for all models, with CV-based model benefiting most. (b) CV-based model maintains superior performance across all data availability levels.

Perturbation Stress Test: To validate meaningful learning vs. memorization, we implemented a controlled experiment corrupting historical laboratory values while preserving prediction targets.

Experimental Protocol:

- **Corruption:** Historical values of the same lab type corrupted with adaptive Gaussian noise: $v_i' = v_i + \mathcal{N}(0, \sigma_{\text{adaptive}})$
- **Noise scaling:** $\sigma_{\text{adaptive}} = \text{std}(v_{\text{prior}}) \times 0.6 \times (1 + 0.5)$
- **Controls:** Identical corruption (50% intensity, fixed seeds) for fair comparison
- **Preservation:** Target values and temporal/categorical context unchanged

Performance Measurement:

$$\text{Degradation} = \frac{\text{MAE}_{\text{corrupted}} - \text{MAE}_{\text{original}}}{\text{MAE}_{\text{original}}} \times 100\%$$

Results: CV-based models showed 2.1$\times$ greater degradation (9.8% vs 4.7%), indicating stronger reliance on patient-specific temporal patterns across diverse laboratory categories including hematology, electrolytes, and metabolic markers.