PAMF: PRIOR-AWARE MULTIMODAL FUSION UNDER TIME-SERIES INCOMPLETENESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal time-series tasks often operate on incomplete observations in practice. For example, in healthcare, ECG segments are lost because electrodes detach or an entire respiratory channel is unavailable during overnight monitoring. Such missingness typically appears in two structurally distinct patterns: within-modality missing, where values are absent within an otherwise observed modality, and modality-level missing, where an entire modality is unavailable. Existing methods typically represent unobserved data implicitly through masks or missing embeddings, without learning instance-specific missing information, and most are designed for only one missingness pattern. A natural approach is to explicitly estimate the missing data; however, existing imputation methods treat missingness uniformly despite their different structural priors, and the imputation process is often isolated from downstream tasks, preventing downstream tasks from guiding imputation toward more informative representations. To address these limitations, we present PAMF, a multimodal time-series framework that explicitly handles different missingness patterns while coupling imputation with downstream prediction through prior-aware flow matching and task-informed representation transfer. Specifically, the method initializes the flow-matching source state with type-specific priors to distinguish the two missing types. It further connects imputation and classification through task-informed representation transfer between architecturally matched encoders. Experiments on multiple multimodal healthcare time-series benchmarks, real native-missingness clinical datasets, and explicit reconstruction comparisons show that the proposed method achieves the strongest overall downstream performance across the evaluated datasets and missingness settings compared with existing baselines.

1 INTRODUCTION

Multimodal time series are used in many applications, where heterogeneous sensors jointly capture different trends and periodic patterns (Zhang et al., 2023; Shukla & Marlin, 2021; Soenksen et al., 2022; Tsai et al., 2019; Hoffmann et al., 2020; Liu et al., 2025; Liang et al., 2024). However, in real-world deployment, these benefits are often limited by pervasive missing data. For example, in healthcare, during sleep staging from overnight polysomnography, it is often difficult to obtain complete physiological recordings for every subject-night. Missing data takes two structurally distinct forms: *within-modality missing*, where individual time points are absent within an otherwise recorded modality, for example when a sensor goes offline for a short period of time, and *modality-level missing*, where an entire modality goes unrecorded (Little & Rubin, 2019; Wu et al., 2024a; Han et al., 2024), for example when a patient is not in a condition to undergo a particular measurement, such as electrocardiography (ECG) or respiratory monitoring. If not handled properly, both types of missingness can seriously degrade downstream performance. How to effectively handle both types of missingness therefore remains a critical open problem in multimodal time series analysis (Wu et al., 2024a; Liu et al., 2025; Wang et al., 2025b).

Early multimodal time series studies assume all modalities are fully observed and readily aligned (Tsai et al., 2019; Zhang et al., 2023; Zadeh et al., 2017; Rahman et al., 2020), limiting their applicability under real-world missingness. Recent methods that account for missing data typically take a conservative approach: substituting absent modalities with learned representations (Mohapatra et al., 2025), routing around missing inputs via expert gating and masks (Han et al., 2024), or using memory

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Capabilities for incomplete multimodal time series. "Within" and "Modality" denote incomplete-input handling, "Imputation" signal reconstruction, and "Task-informed" prediction-to-imputation representation transfer. ✓/× indicate whether each capability is explicitly supported.

Method	Within	Modality	Imputation	Task-informed
MulT, TFN, MAG	×	×	×	×
FuseMoE	✓	✓	×	×
Flex-MoE	×	✓	×	×
Maestro	✓	✓	×	×
MIRA	✓	×	×	×
CSDI, SSSD	✓	×	✓	×
PAMF	✓	✓	✓	✓

banks (Yun et al., 2024), without learning instance-specific representations for the missing content. Table 1 summarizes how they fall short across the key capabilities required.

A natural way to close this gap is to explicitly reconstruct missing data. Yet despite this need, no direct imputation solution has been proposed for multimodal time-series downstream tasks with both within-modality missing and modality-level missing, which we attribute to two independent technical challenges. First, existing diffusion-based generative imputation methods (Tashiro et al., 2021; Lopez Alcaraz & Strodthoff, 2023) typically start from the same generic source distribution for all missing cases, without distinguishing the different structures of within-modality missing and modality-level missing. As a result, the generative process may not be tailored to the specific missing type it must handle. Second, imputation and downstream prediction have largely been treated as independent, sequential tasks (Wang et al., 2025b): the imputation module is trained separately from the downstream task, so task-relevant representations learned later for fusion and downstream task cannot be used to guide imputation, leaving the imputation process largely task-agnostic. Consequently, the recovered representations may not be adapted to the needs of the downstream task.

To address these challenges, we propose PAMF, a unified framework for multimodal time series analysis under missingness. For the first challenge, PAMF features a newly designed flow-matching-based (Lipman et al., 2022; Liu et al., 2022) generative module with *prior-aware initialization*, so that within-modality missing and modality-level missing can be handled with different initial states that reflect their distinct structural information. For within-modality missing, the initialization is constructed from temporal neighbor and mean values to keep the starting point within a plausible local observation range, following the intuition of classical carry-forward imputation (Molenberghs & Kenward, 2007). For modality-level missing, the source state is initialized from observed-data statistics, while the conditional flow uses the available modalities as cross-modal context for sample-specific refinement. This design also reduces inference overhead relative to diffusion-based alternatives (Tashiro et al., 2021; Lopez Alcaraz & Strodthoff, 2023). For the second challenge, task-informed representation transfer connects reconstruction with downstream prediction. The fusion encoder is first pretrained on incomplete inputs and then transferred once to initialize the matched imputation encoder. After imputation training, the predictor is adapted to the completed inputs with the imputer fixed. Our main contributions are as follows:

- We propose PAMF, a unified framework for multimodal time series analysis under missingness, with *prior-aware initialization* in a flow-matching-based generative module, so that within-modality missing and modality-level missing can be handled with different initial states that reflect their distinct structural information.
- We introduce task-informed representation transfer, which carries task-specifically learned representation structure into conditional imputation to guide the imputation model toward task-relevant recovery.
- We evaluate PAMF on multimodal healthcare time-series benchmarks and real native-missingness clinical datasets, and include explicit reconstruction comparisons with diffusion imputation baselines, alongside downstream, efficiency, and component analyses that characterize its benefits and limitations under the tested protocols.

2 Related Work

Multimodal Time Series Fusion. Multimodal time-series fusion methods typically assume complete modalities. Tensor fusion networks (Zadeh et al., 2017), multimodal transformers (Tsai et al., 2019), and multimodal adaptation gates (Rahman et al., 2020) demonstrate strong gains on sentiment and affective benchmarks; in the healthcare domain, MISTS (Zhang et al., 2023) extends this line to irregularly sampled electronic health records via attention-based modality alignment. Despite their promising results, their full-observation assumption limits their use in realistic clinical settings where missingness is ubiquitous (Little & Rubin, 2019).

More recent studies address this issue using masks, routing, or learned placeholders to represent missing inputs implicitly rather than explicitly recovering signal content (Wu et al., 2024a; Wang et al., 2023; Wu et al., 2024b; Reza et al., 2024; Dai et al., 2025). For modality-level missing, FuseMoE (Han et al., 2024) routes inputs through a mixture-of-experts gating mechanism, while its mTAN module also encodes irregularly sampled and partially observed signals; Flex-MoE (Yun et al., 2024) substitutes absent modalities with a learnable embedding bank, providing fixed representations that do not condition on the specific observed context; Maestro (Mohapatra et al., 2025) applies adaptive sparse attention over partially observed streams, yet similarly does not recover the missing signal content. For within-modality missing, MIRA (Li et al., 2025) handles irregular sampling and missing values through channel-independent sparse mixture-of-experts modeling, but remains prediction-oriented rather than an explicit cross-modal signal imputer. Related multimodal physiological and cross-modal generation studies also emphasize robust representations under incomplete observations (Jiang et al., 2025; Liu et al., 2023b), but they do not target explicit imputation of missing signals.

Time Series Imputation. General time-series imputation primarily targets within-modality missing within observed channels, as summarized in recent surveys (Wang et al., 2025b). Earlier approaches include neural imputers such as M-RNN (Yoon et al., 2018), BRITS (Cao et al., 2018), SAITS (Du et al., 2023), and MVI (Bansal et al., 2021), which propagate observations across missing positions through recurrent or attention-based mechanisms; and generative models such as GP-VAE (Fortuin et al., 2020), which places a Gaussian process prior over latent trajectories, and GAN-based models (Luo et al., 2018), which learn the data distribution adversarially. Motivated by denoising diffusion models (Ho et al., 2020), recent probabilistic imputers model conditional distributions over missing values more directly and achieve stronger imputation results: CSDI (Tashiro et al., 2021) performs conditional score generation by treating observed time steps as conditioning signals via two-channel attention; SSSD (Lopez Alcaraz & Strodthoff, 2023) further incorporates structured state space models for long-range dependency modeling when imputing missing values; and LSCD (Fons et al., 2025) conditions diffusion on frequency-domain structure for time-series imputation. Related probabilistic formulations include spatiotemporal diffusion in PriSTI (Liu et al., 2023a), Schrödinger-bridge-based imputation (Chen et al., 2023), consistent diffusion imputation (Zhou et al., 2024), and optimal-transport-based imputation (Wang et al., 2025a). These methods improve signal-level imputation within observed multivariate series, but do not directly address multimodal modality-level missing. Unlike diffusion-based generation, which simulates stochastic denoising dynamics, flow matching learns a velocity field parameterized by the transport variable τ that defines an ODE transporting samples from a source distribution to the data distribution.

Flow Matching for Imputation. Flow matching (Lipman et al., 2022; Liu et al., 2022) therefore offers a natural alternative for generative imputation. Compared with diffusion-based iterative denoising, its source distribution and probability path can be specified directly, making structured source initialization practical. Inference cost depends on the learned field and numerical solver, so we measure NFE and latency for the evaluated implementations. Recent imputation-oriented extensions follow the same direction: CLWF (Qian et al., 2024) studies Lagrangian Wasserstein flows for imputation; MIRI (Yu et al., 2025) frames imputation as minimizing data-mask mutual information under a rectified-flow objective; and CFMI (Simkus & Gutmann, 2025) studies conditional flow matching for general missing-data imputation. Although these methods differ in path design and conditioning objectives, they share the same practical premise: the source distribution can encode useful structure before transport rather than starting from fixed noise. This makes flow matching particularly attractive for imputation, where observed temporal context or distribution information can inform initialization. However, existing flow-matching imputation methods still focus on tabular

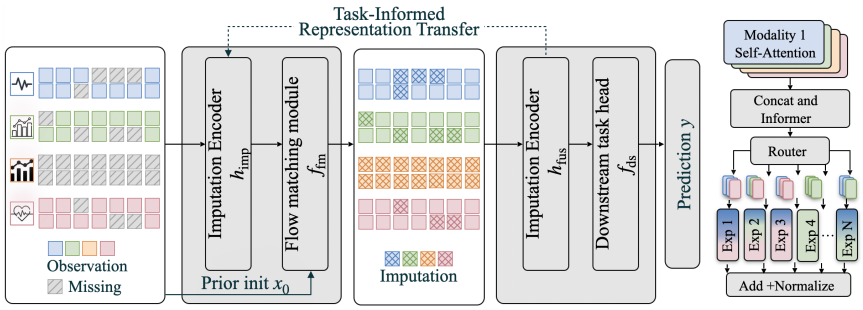


Figure 1: Overview of PAMF. Left: the prediction pipeline combines an imputation encoder Enc_{imp}, a fusion encoder Enc_{fus}, a flow matching module f_{fm}, and a downstream head f_{ds}. Enc_{imp} extracts the imputation context $\mathbf{c}_{\text{imp}}$ from the observed inputs. Conditioned on $\mathbf{c}_{\text{imp}}$ and a prior-aware source state x_0, f_{fm} imputes the missing signals to generate completed data, which are then fed into Enc_{fus} to produce the fused representation $\mathbf{c}_{\text{fus}}$ for downstream head f_{ds} to predict y. Right: the matched MoE backbone instantiated separately in each encoder. The transfer link denotes task-informed representation transfer, implemented through a one-time transfer of the task-specifically learned representation structure from Enc_{fus} to Enc_{imp}.

or unimodal missing-value recovery; they do not encode type-specific priors for within-modality missing and modality-level missing or couple signal-level imputation with downstream classification.

3 Method

3.1 Problem Formulation

For a sample with M modalities, modality m provides a time series $\mathbf{x}^{(m)} \in \mathbb{R}^{T \times C_m}$, where T is the shared sequence length and C_m is the number of channels. For each modality m, let $\mathbf{R}^{(m)} \in \{0,1\}^{T \times C_m}$ be the binary observation mask, with value 1 for observed entries and 0 for missing entries. Let Ω_w denote the set of within-modality missing positions, and let Ω_m denote the set of modality-level missing positions, where an entire modality is absent and thus $\mathbf{R}^{(m)} = \mathbf{0}_{T \times C_m}$. We write $\Omega = \Omega_w \cup \Omega_m$ for the full set of positions to impute. Routing follows the observation mask independently for each modality: $\mathbf{R}^{(m)} = \mathbf{0}$ selects the modality-level prior, while a partially observed branch uses the within-modality prior only at its gaps. Mixed cases apply both rules to their respective branches, without an external missing-type label or classifier. Our inputs use aligned time grids; the flow variable τ is distinct from physical observation time. For training, let $\mathcal{H}$ be the held-out positions with known target values. Under controlled masking of complete records, $\mathcal{H} = \Omega$. For natively incomplete records, $\mathcal{H}$ contains only additionally withheld observed entries, and native missing entries are excluded from the imputation loss (Appendix I). Given partially observed multimodal data and a downstream label y, PAMF initializes missing positions in Ω_w and Ω_m with type-specific priors, transports these states toward completed signals via flow matching, and then learns task-relevant multimodal representations from the imputed data for downstream prediction.

3.2 Model Overview

PAMF addresses imputation and downstream prediction in a single framework consisting of four components: an imputation encoder Enc_{imp}, a fusion encoder Enc_{fus}, a flow matching module f_{fm}, and a task-specific downstream head f_{ds}. Enc_{imp} and Enc_{fus} share the same encoder structure. Concretely, each modality is first linearly projected and encoded by self-attention; the resulting modality embeddings are then concatenated and passed to a cross-modal MoE block, following sparsely gated expert routing (Shazeer et al., 2017) and prior multimodal studies (Han et al., 2024), with Informer attention (Zhou et al., 2021) to produce the fused context. The encoders share an architecture but maintain separate parameters after task-informed representation transfer. Their roles and transfer procedure are specified in Section 3.4. In practice, Enc_{imp} provides the context representation $\mathbf{c}_{\text{imp}}$ used by the imputation branch, while Enc_{fus} provides the fused representation $\mathbf{c}_{\text{fus}}$ for downstream prediction. f_{fm} imputes missing time series values by transporting a structured

source state x_0 to the target data distribution along linear conditional paths (Lipman et al., 2022; Liu et al., 2022), with a conditional velocity network driven by the observed signal and $\mathbf{c}_{\text{imp}}$. Instead of fixing x_0 to a Gaussian distribution, we initialize it via *prior-aware initialization* that injects type-specific structural priors for within-modality missing and modality-level missing respectively (Section 3.3). f_{ds} is a task-specific downstream head applied to $\mathbf{c}_{\text{fus}}$ to predict the final target y. Implementation details are provided in Appendix G.

3.3 Prior-Aware Flow Matching

The intuition behind the type-specific priors is that the imputation source state should depend on the missingness type. For within-modality missing, nearby observed values in the same channel provide a local temporal prior. For modality-level missing, same-modality evidence is absent, so initialization may instead rely on observed-data statistics. Flow matching (Lipman et al., 2022; Liu et al., 2022) is well suited to encode this asymmetry by letting us specify a structured source state x_0. A continuous conditional vector field transports this state toward the target signal x_1; x_τ denotes an intermediate state. To let the generative path use information from the observed inputs, we condition the vector field on the imputation context $\mathbf{c}_{\text{imp}}$ produced by Enc_{imp}. Let $\Phi_\tau^\theta(\cdot; \mathbf{c}_{\text{imp}})$ denote the resulting flow map induced by v_θ. The conditional transport process can be written as

$$\begin{aligned} p_\theta(x_1 \mid \mathbf{c}_{\text{imp}}) &= \left[\Phi_1^\theta(\cdot; \mathbf{c}_{\text{imp}})\right]_{\#} p_0(x_0 \mid \mathbf{c}_{\text{imp}}, \Omega), \\ \frac{d\Phi_\tau^\theta(x_0; \mathbf{c}_{\text{imp}})}{d\tau} &= v_\theta\big(\Phi_\tau^\theta(x_0; \mathbf{c}_{\text{imp}}),\, \tau,\, \mathbf{c}_{\text{imp}}\big)\,. \end{aligned} \tag{1}$$

Here $p_0(x_0 \mid \mathbf{c}_{\text{imp}}, \Omega)$ denotes the source initialization law, and $[\cdot]_{\#}$ is the push-forward operator. Equivalently, the state $x_\tau = \Phi_\tau^\theta(x_0; \mathbf{c}_{\text{imp}})$ follows the conditional ODE

$$\frac{dx_\tau}{d\tau} = v_\theta(x_\tau,\, \tau,\, \mathbf{c}_{\text{imp}})\,, \quad \tau \in [0, 1], \tag{2}$$

with the boundary condition $x_{\tau=0} = x_0$ and $x_{\tau=1} \approx x_1$. Here $\mathbf{c}_{\text{imp}}$ plays the standard role of conditioning the vector field at each FM step.

Following conditional flow matching (CFM) (Lipman et al., 2022; Tong et al., 2024), we train this vector field by regressing to the conditional velocity along a chosen probability path. For the linear conditional bridge (Lipman et al., 2022; Liu et al., 2022), the path and target velocity u_τ are

$$x_\tau = (1-\tau)\, x_0 + \tau\, x_1, \qquad u_\tau(x_\tau \mid x_0, x_1, \mathbf{c}_{\text{imp}}) = x_1 - x_0, \qquad \tau \sim \mathcal{U}[0, 1]. \tag{3}$$

This yields the masked conditional FM regression objective

$$\mathcal{L}_{\text{fm}} = \mathbb{E}_{\tau,\, x_0,\, x_1}\left[\frac{1}{|\mathcal{H}|}\sum_{i \in \mathcal{H}} \left|[v_\theta(x_\tau, \tau, \mathbf{c}_{\text{imp}})]_i - (x_1[i] - x_0[i])\right|^2\right]. \tag{4}$$

Thus $\mathbf{c}_{\text{imp}}$ conditions the vector field at every FM step, while $\mathcal{H}$ restricts supervision to withheld entries whose values are known. The velocity network operates on the full state; the subscript i selects its output at a scored position. During ODE integration, observed entries are clamped to their input values at every step, so only missing positions are updated.

Our main contribution is to condition the source state itself on the missing type. Generalized CFM does not require the source distribution to be Gaussian (Tong et al., 2024), which lets us replace an uninformative Gaussian source with a structured imputation prior. In practice, we construct x_0 directly with missing-type-specific rules. For each missing position $i = (t, c)$, we define

$$x_0[i] = \begin{cases} x_0^{\text{wm}}[i], & i \in \Omega_w, \\ x_0^{\text{ml}}[i], & i \in \Omega_m, \end{cases}. \tag{5}$$

Thus, within-modality entries use temporal-neighbor information, while fully missing modalities use an aggregate channel statistic. The default batch-mean source is not itself an instance-conditioned cross-modal estimate. Instance-specific information enters through the observed context conditioning the learned vector field. The prior specifies a starting state, not the final completion; its downstream contribution is evaluated against a prior-only control. In our implementation, the default within-modality missing prior is a rule-based temporal neighbor interpolation,

$$x_0^{\text{wm}}[t, c] = \alpha_{t,c}\, x_{t,c}^{\text{nbr}} + (1 - \alpha_{t,c})\, \bar{x}_c^{\text{batch}}, \qquad \alpha_{t,c} = \exp(-\lambda \cdot |t^* - t|)\,, \tag{6}$$

where t^* denotes the time index of the nearest observed entry to t in channel c, so that $x_{t,c}^{\mathrm{nbr}} = x_{t^*,c}$, and $\bar{x}_c^{\mathrm{batch}}$ is the per-channel batch mean. Here $\lambda > 0$ is a fixed decay coefficient that controls how quickly the neighbor weight decreases as the temporal gap $|t^* - t|$ grows. We use this weighting because missing entries are more likely to follow their nearest temporal neighbor when the gap is short, whereas the global channel mean becomes a safer fallback as the gap widens. In the default implementation, we set $\lambda = 0.1$, yielding an exponential decay of neighbor confidence with temporal distance. The default modality-level missing prior is the per-channel batch mean:

$$x_0^{\mathrm{ml}} = \bar{x}^{\mathrm{batch}}. \tag{7}$$

This choice reflects the intuition that when an entire modality is absent, a conservative global statistic may provide a stable initialization that stays close to the typical value range of that channel before cross-modal conditioning refines it. We also evaluated richer variants, including a learned-neighbor weighting for x_0^{wm} that adapts the neighbor confidence, a batch-matching prior for x_0^{ml} that retrieves the most context-similar observed sample, and a memory-bank prior for x_0^{ml} that uses learnable cross-sample prototypes; Appendix C gives their full formulations, and Table 10 shows that they do not consistently improve over the simpler defaults above.

3.4 Task-Informed Representation Transfer

Although the flow-matching objective reconstructs missing signals, it does not identify which representations are most useful for downstream prediction. PAMF therefore first conducts task-specific pretraining of $\mathrm{Enc}_{\mathrm{fus}}$ and f_{ds} on incomplete inputs $\mathbf{x}_{\mathrm{obs}}$ using the downstream loss $\mathcal{L}_{\mathrm{ds}}$:

$$\mathcal{L}_{\mathrm{ds}}(\mathbf{x}, y) = \ell_{\mathrm{task}}(f_{\mathrm{ds}}(\mathrm{Enc}_{\mathrm{fus}}(\mathbf{x})), y)\,. \tag{8}$$

Here ℓ_{task} is cross-entropy for single-label tasks and binary cross-entropy for multi-label tasks. Training on incomplete inputs encourages representations that remain predictive under missingness. After pretraining, PAMF copies the parameters of $\mathrm{Enc}_{\mathrm{fus}}$ once to initialize the architecturally matched $\mathrm{Enc}_{\mathrm{imp}}$. Keeping the prediction branch fixed, PAMF then optimizes $\mathrm{Enc}_{\mathrm{imp}}$ and f_{fm} with the masked CFM objective in Equation (4). The encoders subsequently update independently, and downstream gradients are not propagated through the imputation branch. This carries task-informed representation structure into imputation without persistent weight tying or an alignment regularizer. After reconstruction training, PAMF conducts full-pipeline prediction adaptation by freezing $\mathrm{Enc}_{\mathrm{imp}}$ and f_{fm} and obtaining the reconstructed input $\mathbf{x}_{\mathrm{complete}}$. PAMF then optimizes $\mathrm{Enc}_{\mathrm{fus}}$ and f_{ds} with $\mathcal{L}_{\mathrm{ds}}(\mathbf{x}_{\mathrm{complete}}, y)$, allowing the predictor to adapt to completed inputs while the reconstruction and prediction objectives remain separate.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on four multimodal healthcare time-series benchmarks. **Sleep-EDF** (Kemp et al., 2018) is a polysomnography dataset for sleep staging, which we represent with four modalities, EEG, EOG, Resp and EMG, for five-class sleep stage classification. **PTB-XL** (Wagner et al., 2020) is a 12-lead ECG benchmark, which we partition into three lead-group modalities, Limb (I/II/III), Augmented (aVR/aVL/aVF), and Precordial (V1–V6), for multi-label prediction of five cardiac diagnostic superclasses. **PPG-DaLiA** (Reiss et al., 2019) is a wearable sensing dataset; in our benchmark-adapted setting, we use five sensor modalities, chest ACC, wrist ACC, wrist BVP, wrist EDA, and wrist TEMP, for nine-class activity recognition. **Chapman-Shaoxing** (Zheng et al., 2020) is a 12-lead arrhythmia corpus, which we represent as twelve single-lead modalities and use for 7-label diagnosis prediction. We further evaluate on two real-world clinical datasets with native observations. **PhysioNet Challenge 2012** (Silva et al., 2012) contains intensive-care records for in-hospital mortality prediction. **MIMIC-III** (Johnson et al., 2016) provides a clinical cohort for early Sepsis-3 onset prediction from native clinical observations. Detailed dataset statistics, split protocols, and preprocessing are provided in Appendix D and Appendix I.

Missing Data Protocol. We evaluate missingness under both basic and extended protocols. As basic settings, we consider within-modality missing and modality-level missing separately at a 20%

Table 2: Classification performance on four benchmarks under 20% missingness, with within-modality missing and modality-level missing evaluated separately. Means and standard deviations over three runs are rounded to three decimals. **Bold**: best; underlined: second-best.

	Sleep-EDF		PTB-XL		PPG-DaLiA		Chapman	
Method	F1	AUROC	F1	AUROC	F1	AUROC	F1	AUROC
within-modality missing (20%)								
FuseMoE	$0.552_{\pm 0.007}$	$0.912_{\pm 0.006}$	$0.243_{\pm 0.020}$	$0.683_{\pm 0.011}$	$0.602_{\pm 0.014}$	$0.918_{\pm 0.008}$	$0.311_{\pm 0.023}$	$0.630_{\pm 0.007}$
Flex-MoE	$\underline{0.668}_{\pm 0.007}$	$\underline{0.967}_{\pm 0.003}$	$0.422_{\pm 0.022}$	$0.731_{\pm 0.009}$	$0.591_{\pm 0.005}$	$0.917_{\pm 0.004}$	$0.343_{\pm 0.003}$	$0.708_{\pm 0.008}$
Maestro	$0.627_{\pm 0.003}$	$0.955_{\pm 0.000}$	$\underline{0.622}_{\pm 0.007}$	$\underline{0.876}_{\pm 0.001}$	$\underline{0.692}_{\pm 0.017}$	$\underline{0.943}_{\pm 0.001}$	$0.351_{\pm 0.024}$	$0.762_{\pm 0.009}$
MIRA	$0.582_{\pm 0.013}$	$0.936_{\pm 0.010}$	$0.531_{\pm 0.004}$	$0.839_{\pm 0.008}$	$0.672_{\pm 0.002}$	$0.936_{\pm 0.002}$	$\underline{0.403}_{\pm 0.013}$	$\underline{0.815}_{\pm 0.017}$
PAMF	$\mathbf{0.697}_{\pm 0.023}$	$\mathbf{0.977}_{\pm 0.001}$	$\mathbf{0.661}_{\pm 0.003}$	$\mathbf{0.884}_{\pm 0.003}$	$\mathbf{0.714}_{\pm 0.017}$	$\mathbf{0.948}_{\pm 0.006}$	$\mathbf{0.518}_{\pm 0.037}$	$\mathbf{0.870}_{\pm 0.007}$
modality-level missing (20%)								
FuseMoE	$0.502_{\pm 0.005}$	$0.890_{\pm 0.001}$	$0.314_{\pm 0.019}$	$0.662_{\pm 0.004}$	$0.562_{\pm 0.005}$	$0.901_{\pm 0.004}$	$0.327_{\pm 0.003}$	$0.644_{\pm 0.002}$
Flex-MoE	$\underline{0.651}_{\pm 0.005}$	$\underline{0.961}_{\pm 0.001}$	$0.447_{\pm 0.029}$	$0.728_{\pm 0.009}$	$0.570_{\pm 0.007}$	$0.909_{\pm 0.000}$	$0.357_{\pm 0.010}$	$0.699_{\pm 0.006}$
Maestro	$0.646_{\pm 0.012}$	$0.953_{\pm 0.003}$	$\underline{0.604}_{\pm 0.010}$	$\underline{0.872}_{\pm 0.002}$	$\underline{0.675}_{\pm 0.013}$	$0.931_{\pm 0.003}$	$0.413_{\pm 0.015}$	$0.758_{\pm 0.003}$
MIRA	$0.561_{\pm 0.003}$	$0.925_{\pm 0.005}$	$0.504_{\pm 0.007}$	$0.826_{\pm 0.001}$	$0.663_{\pm 0.003}$	$\underline{0.932}_{\pm 0.003}$	$\underline{0.415}_{\pm 0.024}$	$\underline{0.827}_{\pm 0.010}$
PAMF	$\mathbf{0.653}_{\pm 0.009}$	$\mathbf{0.963}_{\pm 0.005}$	$\mathbf{0.648}_{\pm 0.002}$	$\mathbf{0.878}_{\pm 0.002}$	$\mathbf{0.691}_{\pm 0.015}$	$\mathbf{0.944}_{\pm 0.005}$	$\mathbf{0.539}_{\pm 0.009}$	$\mathbf{0.840}_{\pm 0.016}$

Table 3: PTB-XL classification under mixed and high missingness. Mixed denotes simultaneous within-modality missing+modality-level missing at 20% each. High-rate settings evaluate either within-modality missing or modality-level missing at 50%. Means and standard deviations over three runs are rounded to three decimals. **Bold**: best; underlined: second-best.

	Mixed		within-modality missing 50%		modality-level missing 50%	
Method	F1	AUROC	F1	AUROC	F1	AUROC
FuseMoE	$0.211_{\pm 0.040}$	$0.601_{\pm 0.003}$	$0.176_{\pm 0.041}$	$0.612_{\pm 0.002}$	$0.237_{\pm 0.014}$	$0.617_{\pm 0.021}$
Flex-MoE	$0.428_{\pm 0.024}$	$0.757_{\pm 0.003}$	$0.411_{\pm 0.008}$	$0.708_{\pm 0.006}$	$0.460_{\pm 0.005}$	$0.709_{\pm 0.006}$
Maestro	$\underline{0.608}_{\pm 0.009}$	$\mathbf{0.867}_{\pm 0.002}$	$\underline{0.514}_{\pm 0.012}$	$\underline{0.856}_{\pm 0.003}$	$\underline{0.566}_{\pm 0.030}$	$\underline{0.848}_{\pm 0.004}$
MIRA	$0.503_{\pm 0.005}$	$0.828_{\pm 0.003}$	$0.455_{\pm 0.004}$	$0.812_{\pm 0.001}$	$0.505_{\pm 0.001}$	$0.823_{\pm 0.005}$
PAMF (ours)	$\mathbf{0.641}_{\pm 0.011}$	$\mathbf{0.867}_{\pm 0.003}$	$\mathbf{0.626}_{\pm 0.011}$	$\mathbf{0.865}_{\pm 0.003}$	$\mathbf{0.600}_{\pm 0.012}$	$\mathbf{0.852}_{\pm 0.005}$

missing rate. For within-modality missing, instead of adding Gaussian noise, we apply contiguous block masks with lengths drawn uniformly from $[5\%, 10\%]$ of T, per channel, to mimic transient sensor dropout or signal failure. For modality-level missing, we remove whole modalities to mimic absent recordings, such as subjects not wearing or recording a given sensor. Beyond these settings, we evaluate a mixed protocol where within-modality missing and modality-level missing occur simultaneously, real clinical cohorts with native observation gaps, high-missingness protocols, and two signal-dependent missing protocols detailed in Section 4.4 and Appendix I.3.

Baselines and Metrics. We compare against four recent multimodal methods designed for incomplete inputs. FuseMoE (Han et al., 2024) routes inputs through modality-specific experts, tolerating absent streams via sparse gating. Flex-MoE (Yun et al., 2024) substitutes missing modalities with learned embeddings in a mixture-of-experts architecture. Maestro (Mohapatra et al., 2025) applies adaptive sparse attention over partially observed modality streams. MIRA (Li et al., 2025) is a channel-independent sparse MoE Transformer trained for multimodal classification under missingness. We report AUROC and F1, where F1 denotes macro-averaged F1 for multiclass and multilabel classification and positive-class F1 for the binary clinical cohorts, and for reconstruction comparisons, we report mean squared error (MSE). Reported comparison results are means and standard deviations over three independent runs.

4.2 Comparison with Baselines

Basic Missingness. Table 2 reports the original four-benchmark comparison at 20% missingness. PAMF achieves the highest F1 and AUROC on all eight dataset–setting pairs, demonstrating consistent downstream improvements across both within-modality missing and modality-level missing settings. Here, F1 gives equal weight to each label and complements AUROC on these imbalanced tasks (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015). The strongest baseline varies: Flex-MoE on Sleep-EDF, Maestro on PTB-XL and PPG-DaLiA, and MIRA on Chapman.

Table 4: PTB-XL reconstruction comparison with diffusion-based methods. MSE, F1 and latency are reported as mean ± SD over three runs. **Bold**: best; underlined: second-best.

Method	MSE ↓	F1 ↑	NFE	Latency (ms/batch)
CSDI	1.148 ± 0.004	$\underline{0.626} \pm 0.016$	$\underline{50}$	$\underline{137.774 \pm 6.279}$
SSSD	$\underline{0.782 \pm 0.014}$	0.616 ± 0.009	$\underline{50}$	526.982 ± 23.749
PAMF	$\mathbf{0.477 \pm 0.004}$	$\mathbf{0.641} \pm 0.011$	**20**	$\mathbf{75.070 \pm 7.718}$

Table 5: Results on real-world clinical datasets. Challenge2012 reports mortality prediction; MIMIC-III reports early Sepsis-3 onset prediction. Both datasets have native missingness. Results are reported as mean ± SD over three seeds. **Bold**: best; underlined: second-best.

	Challenge2012		MIMIC-III	
Method	F1 ↑	AUROC ↑	F1 ↑	AUROC ↑
MIRA	0.383 ± 0.011	0.742 ± 0.015	0.085 ± 0.005	0.665 ± 0.004
Flex-MoE	0.393 ± 0.023	0.767 ± 0.010	0.117 ± 0.002	$\underline{0.763} \pm 0.001$
FuseMoE	0.423 ± 0.003	0.806 ± 0.004	0.116 ± 0.006	0.753 ± 0.004
Maestro	$\underline{0.438} \pm 0.013$	$\underline{0.812} \pm 0.009$	$\underline{0.122} \pm 0.001$	0.760 ± 0.006
PAMF	$\mathbf{0.451} \pm 0.012$	$\mathbf{0.821} \pm 0.002$	$\mathbf{0.172} \pm 0.008$	$\mathbf{0.791} \pm 0.003$

Mixed and High Missingness. Table 3 compares simultaneous 20% within-modality and 20% modality-level masking, and separate 50% conditions on PTB-XL. PAMF has the highest mean F1 in all three, matching the highest AUROC in the mixed condition, indicating that the same design remains effective when the two missingness types co-occur or become more severe.

4.3 Reconstruction, Downstream Utility, and Inference Cost

We compare PAMF with the diffusion-based imputers CSDI (Tashiro et al., 2021) and SSSD (Lopez Alcaraz & Strodthoff, 2023) on PTB-XL under mixed 20% missingness, pairing each imputer with the same downstream task component. Table 4 shows that PAMF obtains the lowest MSE (0.477) under the matched evaluation protocol. PAMF also achieves the highest F1 (0.641) and the lowest latency (75.070 ms/batch), as flow matching uses 20 function evaluations instead of the 50 iterative denoising steps required by the diffusion baselines. Lower MSE is therefore accompanied by stronger downstream utility, suggesting that improved signal recovery preserves information relevant to classification.

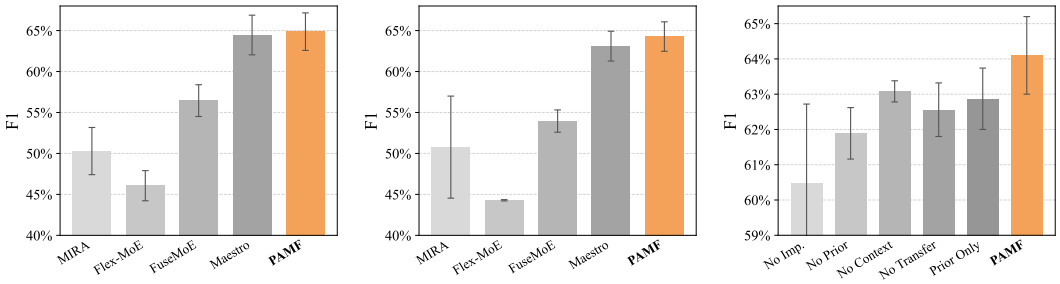
4.4 Real-World Clinical Datasets and Signal-Dependent Missingness

To evaluate PAMF under native missingness in real-world applications, we conduct experiments on PhysioNet Challenge 2012 and MIMIC-III Sepsis-3. During training, additional masks are applied only to observed entries, creating known imputation targets while preserving the original missingness patterns. Table 5 shows that PAMF achieves the strongest overall performance among the compared methods on both datasets. Appendix I provides the detailed dataset construction, training protocols, and additional results.

We also evaluate two signal-dependent mechanisms on PTB-XL. Under missing at random (MAR), the mask depends on observed leads, whereas under missing not at random (MNAR), it depends on values that are subsequently hidden. Both mechanisms are widely considered when analyzing incomplete medical data (Little & Rubin, 2019; Molenberghs & Kenward, 2007). Figure 2 shows that PAMF achieves the highest mean F1 under both mechanisms; Appendix I.3 provides the detailed results. Together, the native and signal-dependent evaluations extend the evidence beyond independent random masking and show consistent gains under non-independent observation processes.

4.5 Ablation and Hyperparameter Analysis

We first assess the contribution of each component by comparing full PAMF with variants that bypass imputation, remove prior-aware initialization, remove cross-modal conditioning, omit task-informed representation transfer, or use the source prior without flow-matching refinement. As shown in Figure 3, full PAMF achieves the highest performance, while all ablated variants perform worse,



(a) Synthetic MAR (b) Synthetic MNAR

Figure 2: Signal-dependent missingness analyses on PTB-XL. (a–b) F1 under synthetic MAR and MNAR masking. Bars show mean ± SD over three runs; orange denotes full PAMF.

Figure 3: PTB-XL component ablation under mixed 20% missingness (F1 mean ± SD).

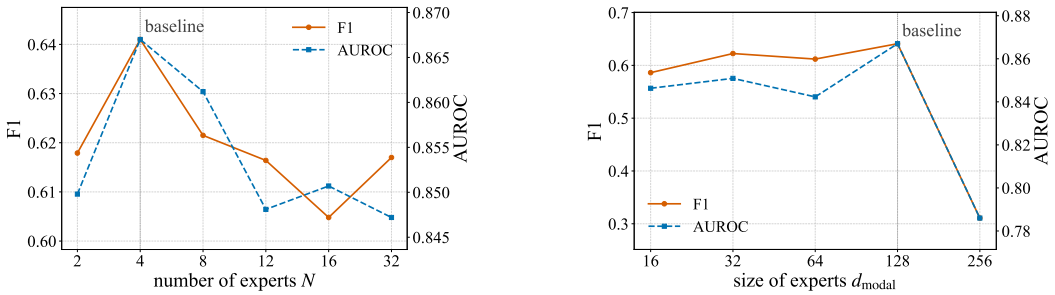


Figure 4: Hyperparameter sensitivity on PTB-XL under mixed 20% missingness (three-run means). Left: number of experts N; right: expert size d_{modal}.

supporting the contribution of each component to the overall framework. The ablations therefore support complementary roles for the source prior, cross-modal conditioning, and task-informed transfer rather than attributing the gain to a single component.

We examine two encoder capacity controls: the number of experts N, which determines the granularity of modality-specific specialization, and expert size d_{modal}, which controls per-expert capacity. Figure 4 reports three-run means and supports N=4 and d_{modal}=128. Larger expert counts become less stable, possibly because specialization becomes overly fragmented (Shazeer et al., 2017), while d_{modal}=256 degrades sharply. We therefore use N=4 and d_{modal}=128.

We also compare direct joint optimization of imputation and prediction with the loss function $\mathcal{L}_{\text{ds}} + \gamma\mathcal{L}_{\text{fm}}$. Table 6 shows that all joint-loss variants underperform our design, supporting task-specific objectives that may avoid conflict between reconstruction and task losses. Additional details are provided in Appendix J.

Table 6: PTB-XL joint-loss controls.

Method	F1	AUROC
PAMF	**0.641** ± 0.011	**0.867** ± 0.003
Joint, $\gamma = 0.1$	0.619 ± 0.012	0.859 ± 0.004
Joint, $\gamma = 1$	0.604 ± 0.005	0.856 ± 0.001
Joint, $\gamma = 10$	0.523 ± 0.020	0.834 ± 0.006

5 Conclusion

Missing data remains a major obstacle in multimodal time-series analysis because both within-modality missing and modality-level missing can degrade downstream performance. PAMF addresses both types through prior-aware flow matching and task-informed representation transfer. Across four synthetic-missingness benchmarks and two native-missingness datasets, it delivers the strongest overall downstream performance against missing-aware baselines and direct imputation pipelines, supporting missingness-aware recovery coupled with downstream learning. These results demonstrate PAMF's effectiveness across diverse missingness patterns while jointly improving reconstruction quality and inference efficiency.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing, support retrieval and discovery of related work, and draft sections of the manuscript. In these uses, the tools helped revise prose, organize literature and existing experimental evidence, and adapt the LaTeX source to the conference format. The authors reviewed AI-assisted text and evidence summaries before inclusion, verified the reported claims against the manuscript sources and experiment records, and take responsibility for the final content of the paper.

ETHICS STATEMENT

This study uses existing healthcare and physiological benchmarks. Potential benefits and risks of model-generated signal completion, including erroneous estimates, distribution shift, and privacy, are discussed in Appendix B. The reported experiments evaluate research models and do not establish clinical deployment readiness.

REPRODUCIBILITY STATEMENT

Dataset processing and missingness protocols are described in Appendix D, with additional native-missingness and signal-dependent protocols in Appendix I. Appendix G specifies the model, training schedule, checkpoint selection, and hardware. Supplementary experiments report the evaluated settings and repeated-run statistics.

REFERENCES

Parikshit Bansal, Prathamesh Deshpande, and Sunita Sarawagi. Missing value imputation on multidimensional time series. *arXiv preprint arXiv:2103.01600*, 2021.

Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems*, 31, 2018.

Yu Chen, Wei Deng, Shikai Fang, Fengpei Li, Nicole Tianjiao Yang, Yikai Zhang, Kashif Rasul, Shandian Zhe, Anderson Schneider, and Yuriy Nevmyvaka. Provably convergent schrödinger bridge with applications to probabilistic time series imputation. In *International conference on machine learning*, pp. 4485–4513. PMLR, 2023.

Ruiting Dai, Chenxi Li, Yandong Yan, Lisi Mo, Ke Qin, and Tao He. Unbiased missing-modality multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24507–24517, 2025.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, 2006.

Wenjie Du, David Côté, and Yan Liu. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023.

Elizabeth Fons, Alejandro Sztrajman, Yousef El-Laham, Luciana Ferrer, Svitlana Vyetrenko, and Manuela Veloso. Lscd: Lomb-scargle conditioned diffusion for time series imputation. *arXiv preprint arXiv:2506.17039*, 2025.

Vincent Fortuin, Dmitry Baranchuk, Gunnar Rätsch, and Stephan Mandt. Gp-vae: Deep probabilistic time series imputation. In *International conference on artificial intelligence and statistics*, pp. 1651–1661. PMLR, 2020.

Xing Han, Huy Nguyen, Carl Harris, Nhat Ho, and Suchi Saria. Fusemoe: Mixture-of-experts transformers for fleximodal fusion. *Advances in Neural Information Processing Systems*, 37: 67850–67900, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Maximilian Hoffmann, Leander Kotzur, Detlef Stolten, and Martin Robinius. A review on time series aggregation methods for energy system models. *Energies*, 13(3):641, 2020.

Wei-Bang Jiang, Xi Fu, Yi Ding, and Cuntai Guan. Towards robust multimodal physiological foundation models: Handling arbitrary missing modalities. *arXiv preprint arXiv:2504.19596*, 2025.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):160035, 2016.

Bob Kemp, Aeilko Zwinderman, Bert Tuk, Hilbert Kamphuisen, and Josefien Oberyé. Sleep-edf database expanded. *Physionet org*, 2018.

Hao Li, Bowen Deng, Chang Xu, Zhiyuan Feng, Viktor Schlegel, Yu-Hao Huang, Yizheng Sun, Jingyuan Sun, Kailai Yang, Yiyao Yu, et al. Mira: Medical time series foundation model for real-world health data. *arXiv preprint arXiv:2506.07584*, 2025.

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 6555–6565, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2019.

Haoxin Liu, Harshavardhan Kamarthi, Zhiyuan Zhao, Shangqing Xu, Shiyu Wang, Qingsong Wen, Tom Hartvigsen, Fei Wang, and B Aditya Prakash. How can time series analysis benefit from multiple modalities? a survey and outlook. *arXiv preprint arXiv:2503.11835*, 2025.

Mingzhe Liu, Han Huang, Hao Feng, Leilei Sun, Bowen Du, and Yanjie Fu. Pristi: A conditional diffusion framework for spatiotemporal imputation. In *2023 IEEE 39th international conference on data engineering (ICDE)*, pp. 1927–1939. IEEE, 2023a.

Ran Liu, Ellen L Zippi, Hadi Pouransari, Chris Sandino, Jingping Nie, Hanlin Goh, Erdrin Azemi, and Ali Moin. Frequency-aware masked autoencoders for multimodal pretraining on biosignals. *arXiv preprint arXiv:2309.05927*, 2023b.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Juan Miguel Lopez Alcaraz and Nils Strodthoff. Diffusion-based time series imputation and forecasting with structured atate apace models. *Transactions on machine learning research*, pp. 1–36, 2023.

Yonghong Luo, Xiangrui Cai, Ying Zhang, Jun Xu, et al. Multivariate time series imputation with generative adversarial networks. *Advances in neural information processing systems*, 31, 2018.

Payal Mohapatra, Yueyuan Sui, Akash Pandey, Stephen Xia, and Qi Zhu. Maestro: Adaptive sparse attention and robust learning for multimodal dynamic time series. *arXiv preprint arXiv:2509.25278*, 2025.

Geert Molenberghs and Michael Kenward. *Missing data in clinical studies*. John Wiley & Sons, 2007.

Weizhu Qian, Dalin Zhang, Yan Zhao, and Yunyao Cheng. Conditional lagrangian wasserstein flow for time series imputation. *arXiv preprint arXiv:2410.07550*, 2024.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 2359–2369, 2020.

Attila Reiss, Ina Indlekofer, Philip Schmidt, and Kristof Van Laerhoven. Deep ppg: Large-scale heart rate estimation with convolutional neural networks. *Sensors*, 19(14):3079, 2019.

Md Kaykobad Reza, Ashley Prater-Bennette, and M Salman Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 47(2):742–754, 2024.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Satya Narayan Shukla and Benjamin M Marlin. Multi-time attention networks for irregularly sampled time series. *arXiv preprint arXiv:2101.10318*, 2021.

Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *2012 computing in cardiology*, pp. 245–248. IEEE, 2012.

Vaidotas Simkus and Michael U Gutmann. Cfmi: Flow matching for missing data imputation. *arXiv preprint arXiv:2506.09258*, 2025.

Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, 5(1):149, 2022.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in neural information processing systems*, 34:24804–24816, 2021.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 6558–6569, 2019.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):154, 2020.

Hao Wang, Haoxuan Li, Xu Chen, Mingming Gong, Zhichao Chen, et al. Optimal transport for time series imputation. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15878–15887, 2023.

Jun Wang, Wenjie Du, Yiyuan Yang, Linglong Qian, Wei Cao, Keli Zhang, Wenjia Wang, Yuxuan Liang, and Qingsong Wen. Deep learning for multivariate time series imputation: a survey. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 10696–10704, 2025b.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024a.

Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri. Multimodal patient representation learning with missing modalities and labels. In *The Twelfth International Conference on Learning Representations*, 2024b.

Jinsung Yoon, William R Zame, and Mihaela Van Der Schaar. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Transactions on Biomedical Engineering*, 66(5):1477–1490, 2018.

Jiahao Yu, Qizhen Ying, Leyang Wang, Ziyue Jiang, and Song Liu. Missing data imputation by reducing mutual information with rectified flows. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. *Advances in Neural Information Processing Systems*, 37:98782–98805, 2024.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 1103–1114, 2017.

Xinlu Zhang, Shiyang Li, Zhiyu Chen, Xifeng Yan, and Linda Ruth Petzold. Improving medical predictions by irregular multimodal electronic health records modeling. In *International conference on machine learning*, pp. 41300–41313. PMLR, 2023.

Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific data*, 7(1):48, 2020.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

Jianping Zhou, Junhao Li, Guanjie Zheng, Xinbing Wang, and Chenghu Zhou. Mtsci: A conditional diffusion model for multivariate time series consistent imputation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 3474–3483, 2024.

A LIMITATIONS

Although PAMF improves prediction under diverse missingness settings, it still requires observed multimodal training examples from which reliable temporal and cross-modal structure can be learned.

B BROADER IMPACTS

This work studies missing-aware multimodal time-series modeling for healthcare, so its main potential positive impact is improved robustness when some sensors or clinical measurements are unavailable. In settings where full multimodal acquisition is difficult, a model that better handles both within-modality missing and modality-level missing may help maintain downstream predictive performance and make multimodal decision support more usable in practice.

At the same time, the method also carries several risks. First, explicit imputation can produce plausible but incorrect physiological signals, which may encourage over-trust if the completed inputs are treated as ground truth rather than model estimates. Second, because the approach depends on learned temporal and cross-modal correlations, performance may degrade under distribution shift, different missingness mechanisms, or populations that are underrepresented in the training data, which could amplify unfair error patterns in deployment. Third, the target application domain involves sensitive health data, so any real-world use would require privacy-preserving data governance, secure access control, and clinical oversight. For these reasons, we view PAMF as a research-stage method rather than a deployment-ready clinical system, and any practical use should include site-specific validation and monitoring under missingness.

C PRIOR-AWARE SOURCE INITIALIZATION

This appendix provides the complete initialization formulas for the prior-aware source initialization x_0 introduced in Section 3.3, along with the design-choice ablation that supports the default prior choices used in our main experiments.

C.1 WITHIN-MODALITY PRIOR

For each missing position $(t, c) \in \Omega_w$, we initialize $x_0^{\mathrm{wm}}[t, c]$ as a convex combination of the nearest observed temporal neighbor and the per-channel batch mean:

$$x_0^{\mathrm{wm}}[t, c] = \alpha_{t,c}\, x_{t,c}^{\mathrm{nbr}} + (1 - \alpha_{t,c})\, \bar{x}_c^{\mathrm{batch}}, \tag{9}$$

where $x_{t,c}^{\mathrm{nbr}}$ is the value of the nearest observed entry in the same channel:

$$x_{t,c}^{\mathrm{nbr}} = x_{t^*,c}, \quad t^* = \arg \min_{t' :\, R_{t',c}^{(m)}=1} |t' - t|, \tag{10}$$

Here $R^{(m)}$ is the observation mask from Section 3.1, and $\bar{x}_c^{\mathrm{batch}}$ is the fallback batch mean when no neighbor exists in channel c.

The blending weight $\alpha_{t,c}$ encodes confidence in the neighbor: it should be high when the gap is short and decay as the gap widens. We consider two instantiations:

Rule-Based (Deterministic).

$$\alpha_{t,c} = \exp(-\lambda \cdot d_{t,c}), \quad d_{t,c} = |t^* - t|, \tag{11}$$

where $\lambda > 0$ is a fixed decay coefficient (default $\lambda = 0.1$; ablated in Table 10). When no observed neighbor exists, $\alpha_{t,c} = 0$ (pure batch mean).

Learned (Adaptive).

$$\alpha_{t,c} = \sigma(\mathrm{MLP}(d_{t,c},\, \mathbf{c}_{\mathrm{imp}})), \tag{12}$$

where σ is the sigmoid function and $\mathbf{c}_{\mathrm{imp}}$ provides cross-modal context, allowing the network to adapt the confidence based on the broader signal structure.

The rule-based variant adds no learned parameters. Its exponential distance weight is a heuristic that gives closer observations more influence; the experiments assess this initialization rather than claiming an exact Gaussian-process posterior identity.

C.2 Modality-Level Prior

For modality-level missing positions $(t, c) \in \Omega_m$, no within-modality signal exists. We consider three source strategies. The default uses observed channel statistics; the retrieval variants additionally use the encoder context $\mathbf{c}_{\text{imp}}$. All are refined by the conditional flow.

Batch Mean (B1).

$$x_0^{\text{ml}} = \bar{x}^{\text{batch}}, \tag{13}$$

The per-channel mean is computed over observed values across time and samples with modality m present, then broadcast over time. Missing entries are excluded. If no usable observation exists for a channel, the implementation uses zero as the fallback source value. This strategy is simple and parameter-free but ignores sample-level heterogeneity.

Batch Matching (B2). For a query sample b with missing modality m, we retrieve the most context-similar sample in the batch that has modality m observed:

$$b^* = \arg \max_{b' \in \mathcal{B},\, m \in \text{obs}(b')} \cos\Big(\mathbf{c}_{\text{imp}}^{(b)},\, \mathbf{c}_{\text{imp}}^{(b')}\Big), \quad x_0^{\text{ml}} = x_m^{(b^*)}, \tag{14}$$

Here $\mathcal{B}$ is the current batch, $\text{obs}(b')$ denotes the observed modalities of sample b', and $x_m^{(b^*)}$ is modality m from the matched sample b^*. The most context-similar sample that has modality m present provides the initialization; if no such sample exists in the batch, the method falls back to B1.

Memory Bank (B3). A set of P learnable prototype vectors $\mathcal{C}_m = \{c_p\}_{p=1}^P \subset \mathbb{R}^{T \times C_m}$ is maintained per modality. The initialization is a soft retrieval:

$$x_0^{\text{ml}} = \sum_{p=1}^{P} \pi_p\, c_p, \quad \pi_p = \frac{\exp(\cos(\mathbf{c}_{\text{imp}},\, \tilde{c}_p))}{\sum_{p'} \exp(\cos(\mathbf{c}_{\text{imp}},\, \tilde{c}_{p'}))}, \tag{15}$$

where $\tilde{c}_p$ is the key associated with prototype c_p, and π_p is its softmax retrieval weight. This variant provides a source without requiring a matching sample in the current batch, but adds learnable prototype vectors and projection parameters.

These three modality-level priors reflect a clear trade-off between simplicity, sample adaptivity, and parameterization. The batch-mean baseline is maximally stable and easy to compute, but it ignores instance-specific variation. Batch matching injects sample-specific information through cross-modal similarity, yet its quality depends on whether the current mini-batch contains a suitable reference example. The memory-bank design tries to remove that batch dependence by storing reusable prototypes, but doing so also compresses the space of possible initializations into a small learned dictionary. To make these trade-offs explicit, we report a quantitative comparison of the prior variants later in Appendix H.

D Dataset Details

D.1 Dataset Overview

The four benchmarks used in Section 4.1 cover sleep staging, ECG diagnosis, wearable activity recognition, and rhythm classification, for a total of 108,098 processed samples. Although all four datasets are physiological time series, they stress the benchmark in different ways: modality counts range from 3 grouped lead streams to 12 single-lead streams, channel counts range from 5 to 12, and sequence lengths range from 250 to 1000 time steps after preprocessing. This diversity is useful for our missingness study because it tests whether the same prior-aware and missingness-aware design can remain effective across both compact and fine-grained multimodal layouts.

To keep the appendix readable, each dataset summary below follows the same structure: source and task, modality construction, and preprocessing with split policy. This standardization makes it easier to compare benchmark assumptions across datasets without overloading the main paper's experiment section. At the same time, the four datasets differ not only in scale but also in how the multimodal structure is defined. Sleep-EDF and PPG-DaLiA combine signals from different physiological

sources, whereas PTB-XL and Chapman-Shaoxing use ECG-based modality decompositions that preserve clinically meaningful lead organization. As a result, the benchmark covers both settings where modalities are semantically distinct by nature and settings where the multimodal view is constructed from a larger sensing system to expose structured missingness.

Another important source of variation is temporal resolution. Sleep-EDF starts from long 30-second physiological epochs, PTB-XL and Chapman-Shaoxing come from fixed-length diagnostic ECG recordings, and PPG-DaLiA is built from shorter sliding windows over free-living wearable streams. These differences matter for our study because within-modality missingness and modality-level missingness can interact differently with long sequences, compact diagnostic snapshots, and motion-heavy activity windows.

In addition to these four controlled benchmarks, we evaluate two clinical cohorts with native observation patterns. PhysioNet Challenge 2012 provides an ICU mortality prediction setting with substantial pre-existing missingness, and MIMIC-III provides an early Sepsis-3 onset prediction setting from numerical clinical observations. These cohorts complement the controlled benchmarks by testing whether the same missing-aware pipeline remains useful when the input mask is inherited from real clinical measurement processes rather than fully specified by synthetic corruption.

D.2 Per-Dataset Details

Sleep-EDF Dataset. The Sleep-EDF dataset provides overnight polysomnographic recordings with manual hypnograms and standard sleep-related physiological signals (Kemp et al., 2018). In our benchmark, it serves as a heterogeneous-but-aligned multimodal sleep-staging task, where all modalities share the same 30-second epoch but contribute different physiological evidence.

- **Task.** Five-class sleep-stage classification.
- **Labels.** ‘W‘ = wakefulness; ‘N1‘ = light sleep onset; ‘N2‘ = stable light sleep with spindles/K-complexes; ‘N3‘ = deep slow-wave sleep (with S3/S4 merged); ‘REM‘ = rapid-eye-movement sleep.
- **Modalities.** We retain five channels and group them into four modalities: EEG (2 channels), EOG (1 channel), Resp (1 channel), and EMG (1 channel). This preserves clinically meaningful separation between neural, ocular, respiratory, and muscle-tone signals. Body-temperature channels available in some releases are not used.
- **Preprocessing and splits.** The loader uses fixed 30-second epochs from the benchmark-provided `split_v2.json` partition, yielding 10,918 processed samples. Each epoch is downsampled to 250 time steps and normalized with per-sample z-score scaling at load time, producing inputs of shape $\mathbb{R}^{5\times 250}$.

PTB-XL Dataset. PTB-XL is a large public 12-lead ECG benchmark with coarse diagnostic annotations over standard clinical superclasses (Wagner et al., 2020). In our benchmark, it represents a structured ECG setting where the multimodal view is constructed from clinically related lead groups rather than distinct sensor families.

- **Task.** Multi-label ECG diagnosis prediction with five diagnostic superclasses.
- **Labels.** ‘NORM‘ = normal ECG; ‘MI‘ = myocardial infarction; ‘STTC‘ = ST/T-wave change; ‘CD‘ = conduction disturbance; ‘HYP‘ = hypertrophy. A record may activate more than one label.
- **Modalities.** We split the twelve leads into three lead-group modalities: Limb (I/II/III), Augmented (aVR/aVL/aVF), and Precordial (V1–V6). This grouping gives modality-level missingness a concrete interpretation as the loss of an anatomically related lead group rather than a single isolated channel.
- **Preprocessing and splits.** The repository loader uses split files keyed by `ecg_id`. After preprocessing, each sample is represented with 12 channels and 250 time steps, i.e., $\mathbb{R}^{12\times 250}$. Per-lead normalization is applied before classification and missingness evaluation.

PPG-DaLiA Dataset. PPG-DaLiA is a wearable sensing dataset collected from 15 subjects in free-living activity settings, with synchronized chest and wrist recordings and ECG-based heart-rate reference signals (Reiss et al., 2019). In our benchmark, we use it as a multimodal wearable activity-recognition task, which makes it a natural testbed for missingness under motion-heavy, sensor-diverse conditions.

- **Task.** Nine-class activity classification.
- **Labels.** 'transient' = transition windows between stable activities; 'sitting' = seated rest; 'ascending_stairs' = walking upward on stairs; 'descending_stairs' = walking downward on stairs; 'table_soccer' = playing table soccer; 'cycling' = cycling activity; 'driving_car' = driving; 'lunch_break' = eating or quiet break period; 'walking' = ordinary walking.
- **Modalities.** We use five sensor modalities defined by the current code path: chest ACC (3 axes), wrist ACC (3 axes), wrist BVP, wrist EDA, and wrist TEMP. This gives 9 total channels and preserves the distinction between motion-dominant and physiology-dominant wearable measurements rather than merging them into a single generic stream.
- **Preprocessing and splits.** The benchmark loader operates on segmented 8-second windows mapped to a common 256-step grid, uses benchmark split JSONs, and produces 64,697 processed segments in total. Per-channel z-score normalization is applied within each segment, giving final inputs of shape $\mathbb{R}^{9\times256}$.

Chapman-Shaoxing Dataset. Chapman-Shaoxing is a 12-lead ECG corpus with diagnostic annotations covering common and additional cardiovascular conditions (Zheng et al., 2020). In our benchmark-aligned configuration, it complements PTB-XL by exposing a more fine-grained lead-wise multimodal formulation, where each ECG lead is treated as its own modality.

- **Task.** Multi-label diagnosis prediction with seven superclasses.
- **Labels.** 'normal' = normal rhythm or morphology; 'cd' = conduction disturbance; 'mi' = myocardial infarction; 'sttc' = ST/T-wave change; 'other' = other cardiac abnormality not covered by the previous groups; 'afib' = atrial fibrillation; 'hyp' = hypertrophy. As in PTB-XL, a record may activate multiple labels.
- **Modalities.** We treat each lead as its own modality: I, II, III, aVR, aVL, aVF, and V1–V6. This yields 12 single-lead modalities, so modality-level missingness corresponds to the loss of an individual lead rather than a coarse group.
- **Preprocessing and splits.** The loader resamples each 10-second record from 500 Hz to 100 Hz, keeps 1000 time steps per lead, applies per-channel z-score normalization, and reads the benchmark `climb_splits` for train/val/test evaluation. If those split files are regenerated from the CSV release, the code maps the CSV `val` partition to test and repartitions the CSV `train` subset into train/val with a 90/10 split and seed 42. The resulting tensor shape is $\mathbb{R}^{12\times1000}$.

D.3 Native Clinical Cohorts

PhysioNet Challenge 2012. PhysioNet Challenge 2012 is an intensive-care corpus for in-hospital mortality prediction (Silva et al., 2012). We use it as a native clinical missingness benchmark because the original ICU measurements are sparse and irregularly observed during the first 48 hours.

- **Task.** Binary in-hospital mortality prediction from the first 48 ICU hours.
- **Labels.** The outcome label is read from the official `Outcomes-a.txt` file and matched to each `RecordID`; died cases are treated as the positive class.
- **Modalities.** The processed input has 42 channels: 34 hourly time-series variables, four static numeric variables, and a four-way ICU-type one-hot encoding. For the encoder interface, the 34 time-series channels are grouped into two 17-channel branches, with the eight static/ICU channels as a third branch.
- **Preprocessing and splits.** Records from Challenge 2012 `set-a` are mapped to a 48-bin hourly grid, with multiple measurements of the same variable within an hour averaged. Time-series channels and static numeric variables are standardized with train-split statistics,

Table 7: Summary of the four benchmark datasets used in this work. Sample counts and tensor shapes correspond to the processed benchmark inputs used by our loaders rather than the raw source files.

Dataset	Domain	Task	Samples	Modalities	Channels	Processed shape
Sleep-EDF	Sleep staging	5-class classification	10,918	4	5	$\mathbb{R}^{5\times 250}$
PTB-XL	ECG diagnosis	5-label prediction	21,837	3	12	$\mathbb{R}^{12\times 250}$
PPG-DaLiA	Wearable activity recognition	9-class classification	64,697	5	9	$\mathbb{R}^{9\times 256}$
Chapman-Shaoxing	ECG rhythm/diagnosis	7-label prediction	10,646	12	12	$\mathbb{R}^{12\times 1000}$
Total / Range	**4 tasks**	—	**108,098**	3–12	5–12	$T = 250$–1000

Table 8: Summary of the additional native clinical cohorts. These cohorts are reported separately from the four controlled benchmarks in Table 7.

Dataset	Domain	Task	Samples	Modalities	Channels	Processed shape
Challenge2012	ICU prognosis	Mortality prediction	4,000 stays	3 groups	42	$\mathbb{R}^{42\times 48}$
MIMIC-III	ICU prognosis	Early Sepsis-3 prediction	318,726 windows	5 clinical groups	30	$\mathbb{R}^{30\times 6}$
Total / Range	**2 tasks**	—	**322,726**	3–5 groups	30–42	$T = 6$–48

while Gender remains binary and ICU type is one-hot encoded. The benchmark-aligned split is stratified 60/20/20 with split seed 42, producing tensors of shape $\mathbb{R}^{42\times 48}$ before collation.

MIMIC-III Sepsis-3. MIMIC-III v1.4 is a de-identified intensive-care database with irregularly sampled vital signs, laboratory measurements and input/output observations (Johnson et al., 2016). We use an audited Sepsis-3 onset cohort for early clinical prediction under native observation patterns.

- **Task.** Binary prediction of first infection-associated acute organ-dysfunction onset after a one-hour warning gap. For an anchor time T, the input window is $[T - 6\text{h}, T)$ and the label window is $[T + 1\text{h}, T + 6\text{h})$.
- **Labels.** The label `y_sepsis3_onset`=1 indicates that the first audited Sepsis-3 proxy onset falls in the future five-hour label window. Windows with a prevalent onset at or before T are excluded by the cohort construction.
- **Modalities.** The processed input is a 30-channel numerical ICU stream organized into five clinical groups: vital signs, temperature and weight, chemistry labs, hematology labs, and additional laboratory or input/output variables.
- **Preprocessing and splits.** Events are mapped to 30 variables, floored to six one-hour bins, and averaged within each ICU-stay, hour, and variable cell. The frozen cohort contains 318,726 windows from 7,413 subjects, split at the patient level into 255,835 training, 31,253 validation, and 31,638 test windows with split seed 42. The processed tensor shape is $\mathbb{R}^{30\times 6}$ before collation.

D.4 Overall Summary Table

Table 7 consolidates the benchmark properties used throughout our experiments. It is intended as a quick reference for dataset scale, modality design, and processed input size, complementing the longer descriptions above. In particular, it highlights that the benchmark spans both coarse and fine-grained modality decompositions, moderate and large sample counts, and short to relatively long processed sequences, which together define the scope of the missingness scenarios studied in the paper.

D.5 Missingness Protocol Details

The main paper reports within-modality missing, modality-level missing, mixed missingness, and high-missingness stress tests. Here we summarize how those conditions are implemented in the benchmark pipeline. Missingness is injected at data-loading time rather than written out as a separate copy of every dataset, so different missingness settings can be evaluated without duplicating the

underlying processed signals on disk. To keep the stochastic patterns reproducible, the injector constructs a local random generator from the pair (`seed`, `sample_id`), ensuring that the same sample receives the same missingness pattern whenever the same configuration is rerun.

Modality-Level Missingness. For modality-level missingness, each currently present modality group is dropped independently with Bernoulli probability p_{mod}. In the main experiments, we set $p_{\text{mod}} = 0.2$ for the basic modality-level missing setting. A safeguard enforces that at least one modality remains: if all groups would be dropped, one dropped group is restored uniformly at random. Once a group is removed, its entry in `mod_mask` is set to zero and the corresponding row block in the temporal mask is zeroed as well. The meaning of one dropped modality therefore depends on the benchmark-specific grouping: one of four physiological groups in Sleep-EDF, one of three lead groups in PTB-XL, one of five wearable sensor streams in PPG-DaLiA, or one of twelve single leads in Chapman-Shaoxing.

Within-Modality Missingness. For within-modality missingness, we use contiguous time-block masking rather than scattered point deletion. For a sequence of length T, each block length is sampled from the interval $[0.05T,\ 0.10T]$, matching the 5–10% range described in Section 4.1. The number of blocks is chosen to reach approximately 20% total temporal corruption, and non-overlap is enforced by rejection sampling with a bounded retry budget. These blocks are applied independently across time-series rows, so different channels or modality groups typically lose different temporal windows. This design better mimics sensor dropouts and short recording interruptions than iid masking would.

Mixed and High Missingness Settings. The mixed setting combines the same two mechanisms: 20% block-based within-modality missing together with 20% Bernoulli modality-level missing. In implementation, the temporal block mask is constructed first and then any fully dropped modalities overwrite their corresponding temporal rows, so modality removal takes precedence whenever both types affect the same group. For the high-missingness stress test, we increase the corruption rate to 50% while keeping the same missingness semantics, allowing us to measure whether performance degrades gracefully when observation loss becomes severe.

Synthetic MAR and MNAR Settings. For the PTB-XL signal-dependent evaluations, we also construct synthetic MAR and MNAR masks. In the MAR setting, leads I, II and III remain observed and their mean absolute amplitude determines the temporal windows masked in the augmented and precordial leads; the mask therefore depends only on information that is still observed. In the MNAR setting, each target lead's own absolute amplitude determines which windows are masked, so the masking process depends on values that become unobserved after masking. These settings are intended as controlled evaluations of signal-dependent missingness rather than as claims about the true PTB-XL acquisition mechanism.

E Evaluation Metric Details

For multiclass and multilabel experiments, F1 denotes macro-averaged F1 and AUROC denotes macro-averaged AUROC. For binary clinical cohorts, F1 denotes positive-class F1. Let N be the number of test samples and K the number of classes. For class k, let TP_k, FP_k, and FN_k denote the numbers of true positives, false positives, and false negatives, respectively. The class-wise precision and recall are

$$P_k = \frac{TP_k}{TP_k + FP_k}, \qquad R_k = \frac{TP_k}{TP_k + FN_k}, \tag{16}$$

and the class-wise F1 score is

$$F1_k = \frac{2P_k R_k}{P_k + R_k}. \tag{17}$$

The reported F1 for multiclass and multilabel tasks is the unweighted average over classes,

$$F1 = \frac{1}{K}\sum_{k=1}^{K} F1_k. \tag{18}$$

For AUROC, we use a one-vs.-rest formulation and then average across classes. Let $y_{i,k} \in \{0, 1\}$ denote the ground-truth indicator of whether sample i belongs to class k, and let $s_{i,k}$ be the predicted confidence score for that class. The class-wise AUROC is

$$\mathrm{AUROC}_k = \frac{1}{|\mathcal{P}_k|\,|\mathcal{N}_k|} \sum_{i \in \mathcal{P}_k} \sum_{j \in \mathcal{N}_k} \left[\mathbb{I}(s_{i,k} > s_{j,k}) + \tfrac{1}{2}\mathbb{I}(s_{i,k} = s_{j,k}) \right], \tag{19}$$

where $\mathcal{P}_k = \{i \mid y_{i,k} = 1\}$ and $\mathcal{N}_k = \{i \mid y_{i,k} = 0\}$ are the positive and negative sample sets for class k. The reported macro-AUROC is

$$\text{Macro-AUROC} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{AUROC}_k. \tag{20}$$

This matches our implementation, which computes F1 and AUROC with a one-vs.-rest reduction across classes.

F Baseline Methods

This section discusses how the representative baselines used in our experiments handle missing inputs and, more importantly, what they still leave unresolved. The core distinction in our paper is between methods that tolerate missingness at the representation level and methods that explicitly reconstruct missing signal content before downstream prediction.

Flex-MoE (Yun et al., 2024). Flex-MoE is one of the clearest examples of a missing-aware fusion model. It explicitly organizes computation around modality subsets through its routers and missing-modality bank, so it is built for the combinatorial setting in which different samples arrive with different observed-modality patterns. Among the classification baselines, this makes it particularly relevant for modality-level missing, because the model does not assume that all modalities are present at test time. Even so, the substitute for a missing modality remains a learned subset-aware representation rather than a sample-level reconstruction of the missing signal itself.

- **What it solves.** It handles arbitrary observed-modality subsets explicitly and learns routing behavior that remains effective when whole modalities are absent.
- **What it does not solve.** It does not reconstruct the missing signal trajectory; the missing-modality bank provides a learned placeholder for fusion, not an imputed waveform or sensor trace tied to the current sample.
- **Why this matters here.** This makes Flex-MoE a strong baseline for missing-aware classification, but not a direct answer to the question of whether explicit signal recovery can provide more downstream value.

MIRA (Li et al., 2025). MIRA is stronger than ordinary fusion baselines in the sense that it natively handles irregular sampling, sparse observations, and masked values within a large time-series foundation-model framework. Its continuous-time positional design and expert specialization make it well matched to temporally sparse clinical signals, and in our benchmark it is the baseline that comes closest to treating missingness as part of the time-series learning problem rather than as an afterthought. Still, MIRA remains prediction-oriented: it uses masks to tell the encoder which observations are reliable, but it does not build an explicit sample-specific recovery process for unobserved signals.

- **What it solves.** It can process partially observed time series without heavy dataset-specific redesign and is naturally robust to irregular temporal sparsity.
- **What it does not solve.** It does not explicitly reconstruct missing signal values, and it does not separate within-modality missing from modality-level missing through different source priors or recovery paths.
- **Why this matters here.** Our setting distinguishes between local gaps inside an observed modality and the complete absence of a modality, whereas MIRA mainly treats both through masked observation encoding.

FuseMoE (Han et al., 2024). FuseMoE is a cross-modal Transformer with sparse expert routing, so it is designed to keep multimodal fusion effective even when different streams have different reliability or sampling structure. In our setting, its main advantage is that missing modalities can be represented by masked or placeholder tokens while the observed modalities continue to interact through cross-attention. This makes it a strong benchmark for missing-aware fusion under modality-level missing, especially when the remaining modalities still contain enough evidence for classification.

- **What it solves.** Its mTAN-based temporal module encodes irregularly sampled and partially observed signals, while expert fusion supports prediction from incomplete modality subsets.
- **What it does not solve.** The missing modality is bypassed rather than reconstructed, so the model never produces a sample-specific estimate of the absent stream.
- **Why this matters here.** When the missing modality contains complementary rather than redundant information, missing-aware routing alone may preserve stability but still leave task-relevant evidence unrecovered.

Maestro (Mohapatra et al., 2025). Maestro is the most explicitly robustness-oriented classifier among the fusion baselines because it combines sparse attention, modality-aware tokenization, and curriculum modality dropout. That training strategy exposes the model to progressively harder missingness patterns and helps it learn to classify from partial evidence instead of collapsing when an input stream disappears. In our benchmark, this makes Maestro a particularly strong representative of the "tolerate missingness" family.

- **What it solves.** It improves downstream robustness to partial observation through training-time modality dropout and missing-aware token handling.
- **What it does not solve.** Its robustness comes from better classification under absent inputs, not from reconstructing the missing waveform or sensor content itself.
- **Why this matters here.** Maestro tests how far carefully trained missing-aware fusion can go, but it still leaves open whether explicit recovery can add information beyond robust masking-based classification.

CSDI (Tashiro et al., 2021). CSDI addresses the opposite side of the problem from the fusion baselines above: it is an explicit imputer. Its role in our benchmark is to provide a principled diffusion-based reconstruction baseline that fills in missing values before the downstream classifier is applied. This makes it highly relevant for within-modality missing, where the central question is not just whether the classifier can ignore corrupted segments, but whether the underlying signal can be recovered well enough to improve the final prediction.

- **What it solves.** It explicitly reconstructs missing values and therefore targets signal recovery directly rather than only masking or bypassing corrupted inputs.
- **What it does not solve.** It is primarily designed for within-series imputation and does not directly address multimodal modality-level absence with different priors for different missingness mechanisms.
- **Why this matters here.** CSDI is a strong test of whether generic diffusion-based imputation helps, but not of whether missingness-type-aware initialization and downstream-coupled training help further.

SSSD (Lopez Alcaraz & Strodthoff, 2023). SSSD is also an explicit diffusion-style imputer, but with a structured state-space design intended to model temporal dependencies more efficiently. Within our benchmark, it serves as a second reconstruction baseline that tests whether stronger temporal generative modeling alone is sufficient to close the gap under missingness. Like CSDI, however, its focus is still generic signal completion rather than multimodal missingness reasoning.

- **What it solves.** It reconstructs missing signal values with a dedicated generative model and therefore provides a direct signal-recovery comparison to our method.
- **What it does not solve.** It does not naturally distinguish within-modality missing from modality-level missing, and its imputation stage remains separate from downstream task supervision.

Table 9: Model and training configurations used for PAMF. The encoder output dimension $\mathrm{fused_dim}$ equals $N_{\mathrm{mod}} \times 128$.

Configuration Type	Parameter Name / Value
Encoder	Per-modal projection dimension: 128
	Per-modal Informer layers: 2
	Per-modal attention heads: 2
	Cross-modal MoE blocks with Informer attention: 1
	Cross-modal MoE experts: 4
	Cross-modal expert hidden dimension: 128
	Encoder output dimension ($\mathrm{fused_dim}$): $N_{\mathrm{mod}} \times 128$
	Classifier dropout: 0.1
Flow Matching	Residual channels: 64
	Number of residual blocks: 4
	FM time embedding dimension: 128
	Feature embedding dimension: 64
	Time-attention heads: 8
	Side-information dimension: $128 + 64 + 1 + \mathrm{fused_dim}$
	Euler ODE steps: 20
Training and Prior	Task-specific pretraining epochs: 5
	Imputation training epochs: 20
	Full-pipeline adaptation epochs: 20
	Batch size: 64
	Classification learning rate: 1e-3
	FM learning rate: 5e-5
	Prior initialization: neighbor_rule + batch_mean
	Imputation checkpoint selection: lowest validation RMSE
	Adaptation checkpoint selection: highest validation F1
	Random seeds: 42 / 52 / 62

- **Why this matters here.** SSSD helps test whether better generic imputation is enough, but it does not answer whether recovery should be conditioned on missingness type and tied more tightly to the end task.

Summary of the Gap. Taken together, the baselines in our study split into two incomplete families. Missing-aware fusion models can classify with absent inputs but usually do not reconstruct the missing signal, whereas explicit imputers can reconstruct signals but are not built to handle multimodal modality-level absence with type-specific priors or downstream-coupled training. This is precisely the gap that motivates PAMF in the main paper. The benchmark is therefore not just comparing stronger versus weaker classifiers; it is comparing three different responses to missing data: tolerating it, generically imputing it, or explicitly reconstructing it with priors matched to the missingness type while still optimizing for the downstream task.

G Implementation Details

Module Roles. The encoder, f_{fm}, and f_{ds} serve different purposes in PAMF. The shared encoder backbone is instantiated twice, as $\mathrm{Enc}_{\mathrm{imp}}$ for cross-modal conditioning and as $\mathrm{Enc}_{\mathrm{fus}}$ for downstream prediction. In both cases, each modality is linearly projected, enriched with temporal positional encoding, processed by two per-modality Informer layers, and then fused across modalities by a cross-modal MoE block with Informer attention (Zhou et al., 2021). The resulting fused vector is used either as the imputation context $\mathbf{c}_{\mathrm{imp}}$ or as the downstream-task context $\mathbf{c}_{\mathrm{fus}}$.

f_{fm} is implemented as a conditional velocity network. Given the interpolated state x_τ, it combines a continuous FM time embedding with side information built from time embeddings, feature embeddings, the observation mask, and $\mathbf{c}_{\mathrm{imp}}$, then predicts the velocity through four residual blocks with alternating time- and feature-wise attention. During training, we sample one continuous FM time

Table 10: Ablation of prior initialization strategies on PTB-XL. within-modality missing variants compare neighbor-rule initialization and learned-neighbor initialization under within-modality missing 20%. modality-level missing variants compare batch-matched, batch-mean (default), and memory-bank initialization under modality-level missing 20%. Means and standard deviations over three runs are rounded to three decimals. Best per group in **bold**.

Missing type	Prior variant	F1
within-modality missing	Neighbor rule	$0.661_{\pm 0.003}$
	Neighbor learned	$\mathbf{0.668}_{\pm 0.007}$
modality-level missing	Batch-matched	$\mathbf{0.650}_{\pm 0.002}$
	Batch-mean (default)	$0.648_{\pm 0.002}$
	Memory Bank	$0.475_{\pm 0.002}$

$\tau \sim \mathcal{U}[0,1]$ independently for each sample in the batch, construct $x_\tau = (1-\tau)x_0 + \tau x_1$, and inject τ into the network through a sinusoidal continuous-time embedding followed by a two-layer MLP. At inference time, we discretize $[0,1]$ into 20 uniform Euler steps and evaluate the velocity field on that fixed grid. Finally, f_{ds} is a lightweight residual MLP head consisting of two linear projections with ReLU and dropout, followed by a linear output layer that maps $\mathbf{c}_{\mathrm{fus}}$ to the task label.

Hyperparameters. Table 9 summarizes the default model and training configuration used for the controlled benchmark experiments, including the standard 5/20/20 epoch allocation for task-specific pretraining, imputation training, and full-pipeline adaptation.

Checkpoint Selection. The imputation checkpoint is selected by the lowest validation RMSE. With the selected imputation branch fixed, the final prediction checkpoint is selected by the highest validation F1. This aligns each checkpoint criterion with its corresponding objective.

Training Hardware. All experiments were trained on NVIDIA DGX B200 systems, using one NVIDIA B200 GPU per run. For the standard 5/20/20 epoch allocation, the representative wall-clock training time was approximately 55.2 minutes on PPG-DaLiA (median over three runs), 9.8 hours on Sleep-EDF, 25.3 minutes on PTB-XL, and 1.64 hours on Chapman-Shaoxing.

H Additional Results

This section collects appendix-only analyses that elaborate on design choices supporting the main paper while remaining secondary to the core narrative. We focus on secondary quantitative results that help interpret design choices and complete the experimental picture, but that are not essential for understanding the main method or its headline performance. We begin with a direct ablation of the prior initialization strategies introduced in Section C.

H.1 Design Choice Ablation

Table 10 reports the effect of prior initialization choices on downstream classification under the corresponding missingness type. The goal of this ablation is not to maximize every appendix number with a more complicated prior, but to test whether additional adaptivity consistently translates into better downstream behavior. Because the proposed framework uses the prior only as a source initialization for flow matching, a stronger prior should ideally shorten the transport path without introducing instability or overfitting. The table therefore helps distinguish improvements that are structurally meaningful from those that merely add parameters.

For the refreshed PTB-XL modality-level missing reruns at $d_{\mathrm{modal}} = 128$, the corresponding AUROC values are $0.880_{\pm 0.001}$ for batch-matched and $0.878_{\pm 0.002}$ for the default batch-mean prior. In contrast, the learned memory-bank prior performs much worse. One possible explanation is that a small set of learned prototypes may produce overly averaged modality-level templates that preserve insufficient sample-specific cross-modal structure.

Table 11: Clean-input classification performance on Sleep-EDF, PTB-XL, and PPG-DaLiA. To match the main-paper comparison set, the table reports the same primary baselines and excludes ShaSpec and CLIMB. Means and standard deviations over three runs are rounded to three decimals. **Bold**: best; underlined: second-best within the reported comparison set.

	Sleep-EDF		PTB-XL		PPG-DaLiA	
Method	F1	AUROC	F1	AUROC	F1	AUROC
FuseMoE	$0.582_{\pm 0.004}$	$0.930_{\pm 0.001}$	$0.224_{\pm 0.034}$	$0.686_{\pm 0.023}$	$0.623_{\pm 0.019}$	$0.925_{\pm 0.002}$
Flex-MoE	$\underline{0.704}_{\pm 0.011}$	$\underline{0.973}_{\pm 0.002}$	$0.440_{\pm 0.016}$	$0.767_{\pm 0.004}$	$0.624_{\pm 0.004}$	$0.928_{\pm 0.001}$
Maestro	$0.699_{\pm 0.004}$	$0.971_{\pm 0.002}$	$\underline{0.622}_{\pm 0.009}$	$\underline{0.882}_{\pm 0.002}$	$\mathbf{0.769}_{\pm 0.004}$	$\mathbf{0.958}_{\pm 0.002}$
MIRA	$0.629_{\pm 0.000}$	$0.958_{\pm 0.001}$	$0.482_{\pm 0.022}$	$0.826_{\pm 0.006}$	$0.676_{\pm 0.004}$	$0.939_{\pm 0.002}$
PAMF	$\mathbf{0.744}_{\pm 0.007}$	$\mathbf{0.981}_{\pm 0.002}$	$\mathbf{0.657}_{\pm 0.005}$	$\mathbf{0.884}_{\pm 0.002}$	$\underline{0.748}_{\pm 0.011}$	$\underline{0.951}_{\pm 0.003}$

The within-modality missing results suggest that adding limited adaptivity can be helpful when the relevant uncertainty is still strongly local: the learned-neighbor variant slightly outperforms the rule-based decay, which is consistent with the idea that the model can refine how much trust to place in nearby observations without abandoning the temporal-neighbor inductive bias. For modality-level missing, the picture is different. Batch-matched retrieval performs only marginally better than the parameter-free batch mean, showing limited gains from coarse cross-sample context in this setting. The large drop of the memory-bank variant further suggests that introducing learnable global prototypes is not automatically beneficial under the evaluated protocol.

Default configuration. In the main experiments, we use the rule-based within-modality prior and the batch-mean modality-level prior. Appendix H.1 shows that batch-matched retrieval is only slightly stronger than the default batch mean, whereas the learned memory-bank variant performs substantially worse. We therefore keep the parameter-free batch-mean prior in the primary setting because it remains competitive without introducing additional learned prototypes. More broadly, these results support our main design principle that prior-aware initialization should encode the right missingness structure with as little unnecessary complexity as possible.

H.2 Encoder Capacity Sensitivity

Our encoder introduces two key capacity controls: the number of experts N and the expert size d_{modal}. The former determines how finely the model partitions modality-specific patterns, while the latter controls the representation capacity of each expert. Both are important for performance and therefore require empirical tuning. We conduct the hyperparameter sweep in Figure 4 under mixed 20% missingness, with each point reporting the mean over three independent runs. Performance is strongest around $N{=}4$; the subsequent oscillation and decline may reflect increasingly fragmented expert specialization and less effective routing supervision. Performance also reaches its best point at $d_{\text{modal}}{=}128$ before dropping sharply at 256. Based on these trends, we use $N{=}4$ and $d_{\text{modal}}{=}128$ in the final configuration.

H.3 Clean-Input Performance

Table 11 reports clean-input performance on Sleep-EDF, PTB-XL, and PPG-DaLiA. We include this table to show that the missing-aware design of PAMF does not depend on evaluating only corrupted inputs. To keep this appendix table aligned with the main-paper comparison set, we report the same primary baselines as in the main results and do not include ShaSpec or CLIMB here. Even without synthetic missingness, the model remains competitive and achieves the best F1 on Sleep-EDF and PTB-XL within this primary comparison set, while remaining close to the strongest baseline on PPG-DaLiA.

These clean results provide a useful complement to the missing-input tables in the main paper. Within the primary baseline set used throughout the main text, PAMF remains the strongest method on Sleep-EDF and PTB-XL even without injected missingness, suggesting that the matched encoder architecture and imputed-data fine-tuning strategy do not trade away clean-case classification quality.

On PPG-DaLiA, Maestro remains the strongest baseline, but PAMF still ranks second on both F1 and AUROC within the same comparison set.

I Additional Missingness Protocols

I.1 PhysioNet Challenge 2012: Native Missingness

PhysioNet Challenge 2012 (Silva et al., 2012) contains intensive-care records for in-hospital mortality prediction. Our processed input has 42 channels: 34 time-series channels and eight static/ICU channels. The loader uses the first 48 ICU hours on an hourly grid, giving 42×48 inputs and a stratified 60/20/20 split with split seed 42. Its encoder interface groups the 34 time-series channels into two 17-channel branches plus an eight-channel static branch. The recorded observation fractions are summarized in Table 12. The original four-benchmark statistics in Table 7 exclude this additional cohort.

Table 12: Observed and natively missing entries in the processed Challenge2012 cohort.

Channels	Observed (%)	Native missing (%)
All 42	41.43	58.57
34 time-series	29.29	70.71
8 static/ICU	93.05	6.95

The native-target formulation is as follows. Let $\mathbf{A}$ indicate native observed entries and $\mathbf{H} \leq \mathbf{A}$ indicate an additional random holdout of 10% of those observations. Conditioning uses $\mathbf{R} = \mathbf{A} - \mathbf{H}$, and the imputation loss is restricted to $\mathcal{H} = \{i : H_i = 1\}$. This formulation excludes unknown native missing values from supervision. At inference it completes native gaps without introducing additional holdouts. The experimental input lies on the processed time grid, rather than arbitrary irregular observation times.

Table 13 reports the native-missingness comparison used in the main paper. For the binary mortality task, F1 denotes positive-class F1.

Table 13: Challenge2012 native-missingness results. Mean $\pm$ SD across three seeds.

Model	F1	AUROC
MIRA	0.383 ± 0.011	0.742 ± 0.015
Flex-MoE	0.393 ± 0.023	0.767 ± 0.010
FuseMoE	0.423 ± 0.003	0.806 ± 0.004
Maestro	0.438 ± 0.013	0.812 ± 0.009
PAMF	$\mathbf{0.451} \pm 0.012$	$\mathbf{0.821} \pm 0.002$

I.2 MIMIC-III Sepsis-3: Native Clinical Observations

MIMIC-III (Johnson et al., 2016) provides the clinical cohort for early Sepsis-3 onset prediction. Each input contains six one-hour bins over the observation window $[T - 6\text{h}, T)$, a one-hour gap, and 30 numerical channels. The Sepsis-3 label is defined over $[T + 1\text{h}, T + 6\text{h})$. The patient-level split contains 255,835 training windows, 31,253 validation windows, and 31,638 test windows; the test prevalence is 2.41%. Native observations are retained as the input mask, and the three reported runs use seeds 42, 52 and 62. The same native-target formulation used for Challenge2012 is applied here: 10% of observed entries are additionally held out as known imputation targets during training, while native missing entries are excluded from the imputation loss.

PAMF obtains the highest F1 and AUROC among the compared methods on this clinical cohort. The MIMIC-III result complements the native-missingness evaluation on Challenge2012 while using a distinct early-onset prediction task and cohort construction.

Table 14: MIMIC-III Sepsis-3 native-observation results. Mean ± SD across three seeds.

Model	F1	AUROC
MIRA	0.085 ± 0.005	0.665 ± 0.004
Flex-MoE	0.117 ± 0.002	0.763 ± 0.001
FuseMoE	0.116 ± 0.006	0.753 ± 0.004
Maestro	0.122 ± 0.001	0.760 ± 0.006
PAMF	$\mathbf{0.172} \pm 0.008$	$\mathbf{0.791} \pm 0.003$

I.3 PTB-XL Synthetic MAR and MNAR

The MAR mechanism preserves leads I, II and III as an observed driver. Their mean absolute amplitude ranks temporal locations; contiguous high-score windows are masked in the augmented and precordial leads until 20% of each target lead is hidden. The mask therefore depends only on information that remains observed. For MNAR, each lead's own absolute amplitude ranks candidate windows, and its highest-energy 20% is masked. That mechanism depends on values subsequently hidden. Both use the same train/validation/test split; neither is claimed to recover the true PTB-XL acquisition mechanism.

The mean ± SD across seeds 42, 52 and 62 is reported in Table 15. PAMF obtains F1/AUROC of 0.649 ± 0.023/0.888 ± 0.001 for MAR and 0.643 ± 0.018/0.883 ± 0.001 for MNAR. Maestro obtains 0.645 ± 0.024/0.886 ± 0.002 and 0.631 ± 0.018/0.879 ± 0.002, respectively. The margins are interpreted as differences of means, without a significance claim.

Table 15: PTB-XL synthetic MAR/MNAR evaluations, reported to three decimals as mean ± SD over three seeds.

	MAR		MNAR	
Method	F1	AUROC	F1	AUROC
MIRA	0.503 ± 0.029	0.828 ± 0.020	0.508 ± 0.062	0.830 ± 0.023
Flex-MoE	0.461 ± 0.018	0.768 ± 0.008	0.443 ± 0.001	0.744 ± 0.007
FuseMoE	0.565 ± 0.019	0.853 ± 0.002	0.540 ± 0.014	0.849 ± 0.003
Maestro	0.645 ± 0.024	0.886 ± 0.002	0.631 ± 0.018	0.879 ± 0.002
PAMF	$\mathbf{0.649} \pm 0.023$	$\mathbf{0.888} \pm 0.001$	$\mathbf{0.643} \pm 0.018$	$\mathbf{0.883} \pm 0.001$

J Training Controls

J.1 Joint-Loss Alternatives

Table 16 compares the proposed training strategy with direct joint objectives of the form $\mathcal{L}_{\mathrm{ds}} + \gamma \mathcal{L}_{\mathrm{fm}}$ under three loss weights. All three joint alternatives have lower mean F1 than the proposed strategy, suggesting that the reconstruction and task losses can interfere when optimized simultaneously in the evaluated implementation.

Table 16: PTB-XL mixed-missingness joint-loss controls, three seeds.

Training	F1	AUROC
Proposed	$\mathbf{0.641} \pm 0.011$	$\mathbf{0.867} \pm 0.003$
Joint, $\gamma = 0.1$	0.619 ± 0.012	0.859 ± 0.004
Joint, $\gamma = 1$	0.604 ± 0.005	0.856 ± 0.001
Joint, $\gamma = 10$	0.523 ± 0.020	0.834 ± 0.006

J.2 Training-Duration Sensitivity

Each row in Table 17 changes one training duration relative to the 5/20/20 allocation for task-specific pretraining, imputation training, and full-pipeline adaptation. The 5/20/20 allocation has the highest F1 among the tested settings, while 3/20/20 is the closest alternative.

Table 17: Training-duration sensitivity across task-specific pretraining, imputation training, and full-pipeline adaptation, evaluated over seeds 42, 52 and 62.

Pretrain / Impute / Adapt	F1	AUROC
1/20/20	0.619 ± 0.009	0.856 ± 0.011
3/20/20	0.635 ± 0.012	0.865 ± 0.003
5/20/20	$\mathbf{0.641} \pm 0.011$	$\mathbf{0.867} \pm 0.003$
7/20/20	0.626 ± 0.017	0.855 ± 0.016
9/20/20	0.622 ± 0.018	0.860 ± 0.012
5/15/20	0.627 ± 0.011	0.861 ± 0.004
5/20/20	$\mathbf{0.641} \pm 0.011$	$\mathbf{0.867} \pm 0.003$
5/25/20	0.627 ± 0.016	0.860 ± 0.010
5/20/10	0.620 ± 0.010	0.853 ± 0.005
5/20/15	0.622 ± 0.010	0.856 ± 0.005
5/20/20	$\mathbf{0.641} \pm 0.011$	$\mathbf{0.867} \pm 0.003$
5/20/25	0.628 ± 0.001	0.857 ± 0.006
5/20/30	0.623 ± 0.008	0.855 ± 0.005

K Additional Reconstruction Evaluation

For a fair reconstruction comparison, PAMF, CSDI, and SSSD use the same PTB-XL split, normalization, mixed 20% missingness protocol, and held-out reconstruction targets. Checkpoints are selected on the validation set and frozen before evaluation on 2,163 test records. Table 4 reports test MSE over three runs under this matched protocol.