Cross-Sample Augmented Test-Time Adaptation for Personalized Intraoperative Hypotension Prediction

Kanxue Li[1], Yibing Zhan[1*], Hua Jin[2*], Chongchong Qi[3], Xu Lin[3], Baosheng Yu[4]

[1]School of Computer Science, Wuhan University
[2]First People's Hospital of Yunnan Province
[3]Yunnan United Vision Technology Company Limited
[4]Nanyang Technological University
likanxue@whu.edu.cn

Abstract

Intraoperative hypotension (IOH) poses significant surgical risks, but accurate prediction remains challenging due to patient-specific variability. While test-time adaptation (TTA) offers a promising approach for personalized prediction, the rarity of IOH events often leads to unreliable test-time training. To address this, we propose CSA-TTA, a novel **C**ross-**S**ample **A**ugmented **T**est-**T**ime **A**daptation framework that enhances training by incorporating hypotension events from other individuals. Specifically, we first construct a cross-sample bank by segmenting historical data into hypotensive and non-hypotensive samples. Then, we introduce a coarse-to-fine retrieval strategy for building test-time training data: we initially apply K-Shape clustering to identify representative cluster centers and subsequently retrieve the top-K semantically similar samples based on the current patient signal. Additionally, we integrate both self-supervised masked reconstruction and retrospective sequence forecasting signals during training to enhance model adaptability to rapid and subtle intraoperative dynamics. We evaluate the proposed CSA-TTA on both the VitalDB dataset and a real-world in-hospital dataset by integrating it with state-of-the-art time series forecasting models, including TimesFM and UniTS. CSA-TTA consistently enhances performance across settings—for instance, on VitalDB, it improves Recall and F1 scores by +1.33% and +1.13%, respectively, under fine-tuning, and by +7.46% and +5.07% in zero-shot scenarios—demonstrating strong robustness and generalization.

Code — https://github.com/kanxueli/CSA-TTA

Introduction

Intraoperative hypotension (IOH) — typically defined as blood pressure falling below a critical threshold for a sustained period (Dong et al. 2024; Wesselink et al. 2018)—is a common but serious complication during surgery. It is strongly associated with adverse outcomes such as acute kidney injury, myocardial infarction, stroke, and even mortality (Jeong et al. 2024; Lee et al. 2021). Accurate and timely prediction of IOH is critical for enabling early interventions before blood pressure drops to dangerous levels (Hwang et al. 2023; Yoon et al. 2020), thereby reducing both the incidence and severity of these adverse events (Mukkamala et al. 2025). However, due to the complex, dynamic, and highly patient-specific nature of physiological responses during surgery, reliable IOH prediction remains a significant challenge despite advances in intraoperative monitoring and machine learning.

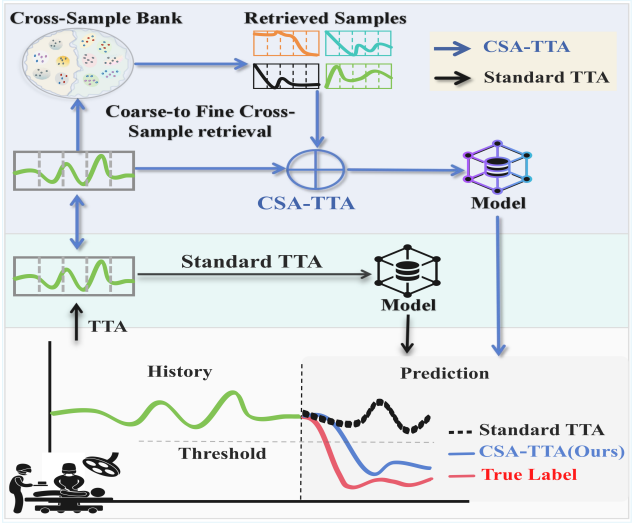


Figure 1: An illustrative comparison. Standard TTA, relying on recent stable history, often produces overly smooth predictions and misses sudden changes. CSA-TTA leverages a cross-sample augmented dataset to capture diverse temporal patterns, enabling personalized IOH prediction.

Recent studies have explored various methods to improve IOH prediction (Shi et al. 2023a; Sidiropoulou et al. 2022; Lee et al. 2018). For instance, CMA (Lu et al. 2023) employed attention mechanisms to capture temporal and feature-level dependencies, while HMF (Cheng et al. 2024) integrated contextual, physiological, and temporal features. However, the ability of these models to generalize remains limited by individual differences in patients' physiology and the influence of clinical interventions (e.g., anesthesia or drug administration) (Cai et al. 2025; Mohammadi et al.

*Corresponding authors: Yibing Zhan and Hua Jin.

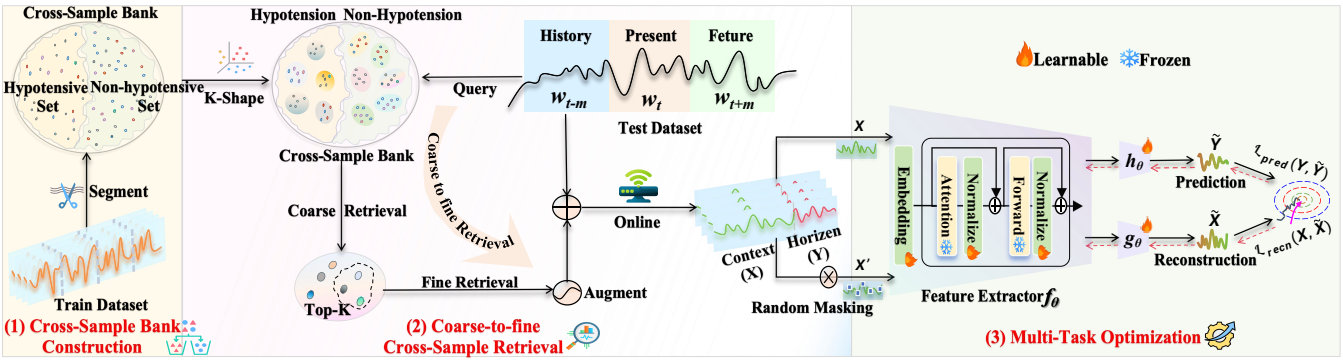


Figure 2: The main proposed CSA-TTA framework. It comprises three key steps: (1) Cross-sample bank construction, (2) Coarse-to-fine retrieval, and (3) Multi-task optimization.

2024; Li et al. 2022). These factors introduce implicit distribution shifts in real-time signals that are typically poorly captured by population-level models.

TTA offers a promising paradigm to tackle such distribution shifts by adapting models using test data during inference (Liang, He, and Tan 2025). Techniques such as TTT (Sun et al. 2020) and TTT++ (Liu et al. 2021) leverage self-supervised auxiliary tasks to refine models at inference time and have shown effectiveness in fields like computer vision (Karmanov et al. 2024; Lim et al. 2023) and NLP (Shi et al. 2024). In the context of IOH, TTA holds potential for personalizing predictions by utilizing recent patient history to dynamically adjust to individual physiological patterns. Despite its promise, standard TTA struggles due to the rarity of hypotensive events. For example, in the VitalDB dataset (Lee et al. 2018), hypotension accounts for just 12.6% of all samples, with the majority of patients experiencing it in less than 10% of the surgical timeline. Standard TTA, which typically relies on single-sample adaptation, often fails to detect sudden blood pressure drops, especially when such transitions are not present in the current patient's history. As shown in Figure 1, these methods yield overly smooth predictions and underperform during abrupt physiological changes caused by clinical interventions.

To overcome these limitations, we propose cross-sample augmented test-time adaptation (CSA-TTA) for personalized IOH prediction. As depicted in Figure 2, CSA-TTA first builds a cross-sample bank by partitioning historical data into hypotensive and non-hypotensive segments, increasing the availability and diversity of critical training signals. To retrieve relevant samples efficiently, we design a coarse-to-fine strategy: K-Shape clustering (Paparrizos and Gravano 2016) is used to identify representative cluster centers in the coarse stage, followed by fine-grained retrieval of the top-K semantically similar samples based on the patient's current data. These retrieved samples—optionally augmented with perturbations—are combined with the patient's signal to form a balanced and representative adaptation set. Finally, CSA-TTA employs a multi-task optimization strategy, incorporating both self-supervised masked reconstruction and retrospective forecasting to improve model adaptability to subtle and rapid intraoperative dynamics.

Our main contributions are twofold: **1)** To the best of our knowledge, this is the first attempt to apply test-time adaptation for personalized IOH prediction. **2)** We propose CSA-TTA, which introduces a cross-sample bank, a coarse-to-fine retrieval strategy, and multi-task optimization. Together, these components enable robust test-time adaptation for personalized IOH prediction, despite the extreme scarcity of individual IOH events during surgery.

Related Work

Intraoperative Hypotension Prediction IOH prediction is critical for timely intervention and reducing postoperative risks (Dong et al. 2024; de Keijzer et al. 2024; Meng et al. 2018). With the growing availability of continuous physiological data, many deep learning models have been developed, following (Mukkamala et al. 2025; Mohammadi et al. 2024). The first treats IOH prediction as a classification task based on time series inputs, but suffers from limited interpretability and sensitivity to varying clinical definitions (Jeong et al. 2024; Yoon et al. 2020). The second forecasts continuous blood pressure trajectories, followed by post hoc IOH detection, offering improved flexibility but still struggling with patient-specific variability and intraoperative distribution shifts—factors that undermine robustness in real-world deployment (Cheng et al. 2024).

Test-time Adaptation TTA is a paradigm for mitigating distribution shifts between training (source) and test (target) data, enabling models to adjust to new data distributions during inference (Liang, He, and Tan 2025). The foundational approach to TTA is TTT (Sun et al. 2020), which incorporates self-supervised auxiliary tasks during inference to update model parameters and enhance generalization under distribution shift. Building on this, TTT++ (Liu et al. 2021) systematically analyzes the scenarios where self-supervised test-time training is effective or limited, and proposes strategies to improve its stability and robustness. These methods have shown success in various fields, such as vision (Karmanov et al. 2024) and language (Tan et al. 2025; Shi et al. 2024). Despite its potential, standard test-time adaptation struggles in IOH prediction due to rapidly shifting patient-specific physiological signals and rare IOH events.

Method

In this section, we detail CSA-TTA for personalized IOH prediction. As illustrated in Figure 2, CSA-TTA comprises three key components: cross-sample bank construction, coarse-to-fine retrieval, and multi-task optimization.

Problem Formulation

Let $X \in \mathbb{R}^{L \times C}$ represent a multivariate time series of a patient's vital signs, where L is the lookback window length and C is the number of variables (e.g., mean arterial pressure, body temperature, heart rate, ECG). The objective of IOH prediction is to forecast the blood pressure sequence $Y \in \mathbb{R}^{1 \times H}$ over the next H time steps using a model F_θ: $\hat{Y} = F_\theta(X)$. Traditional models aim to learn the conditional distribution $P(Y|X)$, assuming that the observed features X fully capture the relevant physiological dynamics. However, unmeasured patient-specific factors and subtle signal shifts—denoted as X_{latent}—can significantly affect blood pressure. The true target distribution is thus better represented as $P(Y|X, X_{\text{latent}})$.

Since X_{latent} is not observable at inference, models trained only on X often suffer from distributional mismatch, leading to poor generalization—especially in the presence of individualized IOH patterns. Test-time adaptation (TTA) addresses this by leveraging a patient's recent history D_{hist} to infer latent physiological context, implicitly aligning predictions with:

$$P(Y|X) \cdot P(X, X_{\text{latent}}|D_{\text{hist}}) \approx P(Y|X, X_{\text{latent}}). \quad (1)$$

Here, $P(X, X_{\text{latent}} \mid D_{\text{hist}})$ serves as an empirical posterior over latent dynamics, estimated from the patient's trajectory. However, due to the episodic and sparse nature of hypotension, short-term patient data often lacks sufficient hypotensive information. To address this, we posit that patients with similar physiological profiles exhibit similar intraoperative responses. Thus, we extend the adaptation context using cross-sample retrieval:

$$P(X, X_{\text{latent}}|D_{\text{hist}}) \approx \hat{P}(X, X_{\text{latent}}|R(D_{\text{hist}})), \quad (2)$$

where $\mathcal{R}(D_{\text{hist}})$ is a set of retrieved cross-sample segments that are temporally and semantically aligned with the current patient trajectory. This strategy enriches the adaptation signal with personalized priors from similar patients, enabling more robust estimation of $P(Y|X, X_{\text{latent}})$, even in the absence of local hypotensive events.

Cross-Sample Bank Construction

CSA-TTA introduces a cross-sample bank to enrich the diversity of temporal patterns. Specifically, it first segments physiological time series data from all patients in the historical dataset into fixed-length fragments, forming a cross-sample bank B. The original intraoperative hypotension dataset is denoted as:

$$\mathcal{D} = \{(X_n, Y_n)\}_{n=1}^{N}, \quad (3)$$

where N is the total number of samples, X_n represents the n-th time series sample, and Y_n denotes the corresponding label. Each sample derived from D is represented as $s = (X_n, Y_n)$, where x is a contiguous input subsequence and y is its label/future sequence. The cross-sample collection $\mathcal{B}$ is then partitioned into two subsets based on the occurrence of hypotensive events, i.e.,

$$\mathcal{B} = \mathcal{B}_{\text{hypo}} \cup \mathcal{B}_{\text{non-hypo}}. \quad (4)$$

A hypotensive event is defined as a mean arterial pressure below 65 mmHg lasting at least one minute (Shi et al. 2023b; Wesselink et al. 2018). During adaptation, CSA uses the target patient's historical data to retrieve semantically relevant segments from a cross-sample bank, effectively combining personalized patterns with population-level insights. This strategy improves the model's ability to handle sudden state changes and enhances generalization.

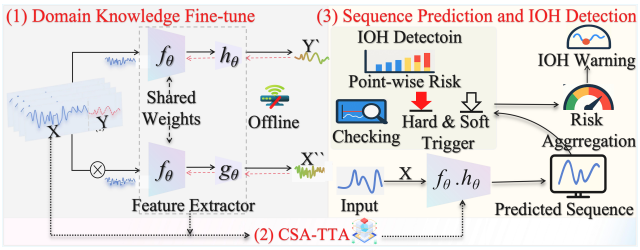


Figure 3: Illustration of the key steps in applying CSA-TTA for personalized IOH prediction. (1) Domain knowledge adaptation, (2) CSA-TTA, and (3) Sequence prediction and IOH detection.

Coarse-to-Fine Retrieval

Direct retrieval of relevant samples from the cross-sample bank can be computationally expensive for large datasets. To alleviate this, we propose a coarse-to-fine retrieval strategy that enhances effectiveness while preserving relevance. It employs an adaptive context window to process the patient's streaming vital data. At each time step t, a present window of length m is defined as the test window, while the previous window of the same length serves as the history window, denoted as:

$$\mathcal{W}^{\text{hist}}_{t-m:t} = [x_{t-m}, \ldots, x_{t-1}], \quad (5)$$

where $x_t \in \mathbb{R}^C$ represents the C-dimensional vital signals collected at time t, and each window contains multiple time series samples.

Coarse-grained Retrieval. To efficiently narrow the search space, CSA-TTA conducts clustering separately on the hypotensive subset $\mathcal{B}_{\text{hypo}}$ and the non-hypotensive subset $\mathcal{B}_{\text{non-hypo}}$. Specifically, the K-Shape clustering(Paparrizos and Gravano 2016) is employed independently on each subset to identify representative cluster centroids. K-Shape is particularly well-suited for physiological time series data due to its shape-based clustering approach, which aligns temporal dynamics of sequences without requiring explicit time warping or amplitude normalization(Moe et al. 2023; Yang et al. 2017). When a query sample s from the adaptive context window $W^{\text{hist}}_{t-m:t}$ is available, CSA-TTA first determines its category and then computes its similarity to the

cluster centroids of the corresponding subset. The query is then assigned to the cluster with the most semantically relevant centroid, thereby localizing its nearest temporal neighborhood within the large cross-sample bank.

Fine-grained Retrieval. CSA-TTA refines the candidate selection by computing the semantic similarity between the query sample s and all samples within the cluster identified during the coarse retrieval phase from the cross-sample bank $\mathcal{B}$. We employ Dynamic Time Warping (DTW) (Berndt and Clifford 1994; Senin 2008) as similarity metric for cross-sample retrieval. DTW measures the optimal alignment between two time series, allowing it to capture temporal shifts and variations, making it effective for identifying similarities in physiological signals (Bringmann et al. 2024). The top-K samples with the highest similarity scores are then retrieved to form a refined candidate set. Formally, the refined candidate set $\mathcal{D}_{\text{retrieval}}$ is given by:

$$\mathcal{D}_{\text{retrieval}} = \left\{ r \in \mathcal{B}, s \in W^{\text{hist}}_{t-m:t} \mid r \in \text{TopK}\left(\mathcal{B}, \text{DTW}\left(r, s\right)\right) \right\}, \tag{6}$$

where $\text{TopK}(\mathcal{B}, \text{DTW}(\cdot,\cdot))$ denotes the subset of K samples from $\mathcal{B}$ having the highest semantic similarity scores with the query s, and $\text{DTW}(\cdot,\cdot)$ is a semantic similarity function between time series samples. The retrieved samples are further augmented with perturbations (e.g., Gaussian noise, temporal scaling) to increase variability and better simulate potential patient-specific variations. At each adaptation step, we construct a cross-sample augmented dataset by combining the patient's own history window with these augmented reference samples:

$$\mathcal{D}_t^{\text{CSA-TTA}} = W^{\text{hist}}_{t-m:t} \cup \text{Aug}(\mathcal{D}_{\text{retrieval}}). \tag{7}$$

Multi-task Optimization

We adopt a multi-task optimization framework, i.e., the model is refined through two learning objectives: self-supervised masked reconstruction and retrospective sequence forecasting. As shown in Figure 2-(3), the architecture includes a shared feature encoder f_θ, and two task-specific branches: h_θ for the primary prediction task and g_θ for the auxiliary self-supervised task.

The complete model is denoted as $F_\theta = (f_\theta, h_\theta, g_\theta)$. For self-supervision, we employ a masked reconstruction objective, a lightweight yet effective strategy for enhancing time-series representation learning (Nie et al. 2023; Woo et al. 2024). The model is trained to minimize a combined loss:

$$\min_{f_\theta, h_\theta, g_\theta} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{\text{Pred}}(X_n, Y_n; f_\theta, h_\theta) + \mathcal{L}_{\text{Recon}}(X_n; f_\theta, g_\theta). \tag{8}$$

The total loss $\mathcal{L}_{\text{CSA-TTA}}$ is computed over the cross-sample adaptation dataset, and model parameters are updated via gradient descent:

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{CSA-TTA}}, \tag{9}$$

where η is the learning rate. To preserve generalization while enabling personalization, only a subset of parameters—typically in the input, output, and normalization layers—are updated during adaptation (Kang et al. 2024).

CSA-TTA supports two modes: fine-tuning and zero-shot. In the fine-tuning setting (Figure 3-(1)), both the main and auxiliary tasks are jointly fine-tuned offline before test-time adaptation, following (Sun et al. 2020). In contrast, the zero-shot mode eliminates the need for offline fine-tuning or architectural changes; test-time adaptation is performed directly via retrospective regression. This allows CSA-TTA to remain effective even when auxiliary task training is impractical or historical data is unavailable.

To handle uncertainty and signal fluctuations during prediction, we implement a hybrid mechanism (Figure 3-(3)). Predicted blood pressure sequences are converted into pointwise risk scores representing the probability of MAP falling below a clinical threshold. We then apply two triggers: a hard trigger to detect sustained hypotensive periods and a soft trigger that evaluates average risk in sliding windows. These are combined to produce a final probabilistic estimate, balancing clinical reliability with model flexibility.

Experiments

Experimental Setups

Dataset. We conduct experiments using the VitalDB dataset, a real-world clinical database collected at Seoul National University Hospital. It contains vital sign data from 6,388 patients who underwent noncardiac surgeries between June 2016 and August 2017 (Lee et al. 2018; Shi et al. 2023b). From the raw records, we extract temporal features such as mean arterial pressure, body temperature, and heart rate, consistent with prior studies (Lu et al. 2023). Rigorous quality control is applied to exclude cases with over 20% missing or abnormal values, as well as surgeries lasting less than one hour. This results in a curated dataset of 2,150 patient cases. To assess the impact of temporal resolution, the data are sampled at both 2-second and 30-second intervals. The dataset is split at the patient level into training (70%), validation (20%), and test (10%) subsets. The training and validation sets are primarily used for offline fine-tuning. The training set also serves to construct the cross-sample bank required for test-time adaptation. The test set is reserved exclusively for evaluating model performance under test-time adaptation using CSA-TTA. In addition, we utilize a real-world clinical dataset from a collaborating hospital. The in-hospital test data are fully de-identified, and all experiments are conducted in compliance with the institution's ethical review and data governance protocols. These data were collected from standard clinical devices—including blood pressure monitors, ventilators, and infusion pumps—and processed using the same quality control procedures as VitalDB. This dataset is sampled at one-minute intervals and comprises 130 cases for evaluation, with a additional 910 cases used exclusively to construct the cross-sample retrieval bank for test-time adaptation.

Baselines. We use two state-of-the-art pre-trained time series models as backbones: TimesFM (Das et al. 2024), a decoder-only foundation model trained on a large corpus of real and synthetic time series, and UniTS (Gao et al. 2024), a unified multi-task model designed for diverse time series applications. We compare against two test-time training base-

(a) Zero-shot Setting

Model		VitalDB (2S)						VitalDB (30S)					
		F1↑	Rec↑	Prec↑	Acc↑	MAE↓	MSE↓	F1↑	Rec↑	Prec↑	Acc↑	MAE↓	MSE↓
TimesFM	Test(Das et al. 2024)	62.40	64.97	60.53	88.07	6.19	83.61	64.17	58.87	70.87	87.60	6.49	92.77
	CSA-TTA (Ours)	**64.10**	**66.09**	**62.88**	**88.40**	**6.07**	**80.55**	**64.90**	**59.27**	**72.20**	**87.93**	**6.28**	**85.28**
	Δ(%)	▲1.70	▲1.13	▲2.35	▲0.33	▲1.94	▲3.66	▲0.73	▲0.40	▲1.33	▲0.33	▲3.24	▲8.07
Units	Test(Gao et al. 2024)	49.67	44.83	58.44	87.53	7.36	98.88	52.23	43.24	71.14	85.47	7.32	99.96
	CSA-TTA (Ours)	**54.53**	**52.50**	**59.32**	**88.03**	**7.22**	**96.62**	**57.30**	**50.70**	66.56	**85.83**	**7.19**	**95.84**
	Δ(%)	▲4.87	▲7.67	▲0.88	▲0.50	▲1.90	▲2.29	▲5.07	▲7.46	▼4.58	▲0.36	▲1.78	▲4.12

(b) Fine-tuned Setting

Model		VitalDB (2S)						VitalDB (30S)					
		F1↑	Rec↑	Prec↑	Acc↑	MAE↓	MSE↓	F1↑	Rec↑	Prec↑	Acc↑	MAE↓	MSE↓
TimesFM	Test(Das et al. 2024)	64.20	64.93	65.13	89.13	6.03	77.87	65.80	**62.67**	69.97	87.77	5.94	76.27
	TTT(Sun et al. 2020)	64.00	64.77	64.79	89.07	6.02	77.70	65.77	62.63	69.93	87.77	5.94	76.28
	TTT++(Liu et al. 2021)	64.10	64.80	65.13	89.13	6.02	77.68	65.80	62.60	70.00	87.80	5.93	76.19
	CSA-TTA (Ours)	**64.83**	**65.99**	65.42	**89.40**	**5.94**	**76.19**	**66.07**	62.33	**70.97**	**88.03**	**5.82**	**72.93**
	Δ(%)	▲0.63	▲1.06	▲0.28	▲0.27	▲1.49	▲2.16	▲0.27	▼0.34	▲1.00	▲0.26	▲2.02	▲4.38
Units	Test(Gao et al. 2024)	64.60	63.93	67.13	89.50	5.80	71.65	63.83	59.07	70.00	87.97	6.37	81.77
	TTT(Sun et al. 2020)	64.64	63.67	67.48	89.57	5.82	71.58	63.47	55.23	76.09	88.07	6.33	81.20
	TTT++(Liu et al. 2021)	64.71	63.90	67.39	89.53	5.81	71.45	63.17	56.66	73.23	88.00	6.30	81.20
	CSA-TTA (Ours)	**65.73**	**65.27**	**68.13**	**89.70**	**5.72**	**70.07**	**65.57**	**59.93**	72.88	**88.47**	**6.23**	**79.31**
	Δ(%)	▲1.13	▲1.33	▲0.99	▲0.20	▲1.38	▲2.21	▲1.74	▲0.86	▲2.88	▲0.50	▲2.20	▲3.01

Table 1: Performance comparison under different adaptation settings: (a) zero-shot setting; (b) fine-tuning setting. All results are averaged over three predict horizons (5, 10, and 15 minutes), using a fixed lookback length of 15 minutes. The best performance is highlighted in **bold**, and the second-best is underlined. Metrics with the ↑ symbol (e.g., F1, Recall) indicate higher is better, while those with ↓ (e.g., MAE, MSE) indicate lower is better. Performance changes are marked with ▲ for improvement and ▼ for degradation.

lines: TTT (Sun et al. 2020), which adapts the model during inference via a simple self-supervised task (e.g., rotation prediction) to mitigate distribution shifts, and TTT++ (Liu et al. 2021), which enhances this by incorporating feature alignment to improve robustness and reduce catastrophic forgetting during adaptation.

Implementation Details. We evaluate model performance using both regression and classification metrics. For regression, we use Mean Absolute Error (MAE) and Mean Squared Error (MSE). For classification, we report Accuracy, Recall, Precision, and F1-score. For all experiments, we set the lookback window length L to 15 minutes and vary the forecast horizon H among 5, 10, and 15 minutes. During the fine-tuning stage, models are trained on training data for 10 epochs with a learning rate of 1×10^{-4}, batch size of 64, and dropout rate of 0.01. During test-time adaptation, models are updated for one epoch in the fine-tuning setting and three epochs in the zero-shot setting. We apply a partial fine-tuning strategy, where only the parameters in the input layer, output layer, and layer normalization layers of the backbone are updated, to balance adaptability with generalization. To ensure fairness, the same architecture and hyperparameter configurations are applied across all datasets. All experiments were accelerated by four NVIDIA A100 GPUs.

Performance Comparison

Zero-Shot Setting. In the zero-shot setting, where no fine-tuning is performed before test-time adaptation, CSA-TTA achieves significant improvements. As shown in Table 1(a), TimesFM + CSA-TTA improves Recall by 1.13% and F1 by 1.70% on VitalDB(2S). Units + CSA-TTA shows even greater performance, with Recall increasing by 7.46% and F1 improving by 5.07% on VitalDB(30S). Similar improvements are observed on the in-hospital test set, Units + CSA-TTA improves Recall by 9.56% (43.77% → 53.33%) and F1 by 7.70% (56.10% → 63.80%), as shown in Table 2. This substantial improvement in Recall and F1 is clinically significant for intraoperative hypotension prediction, allowing it to more effectively identify hypotension events and reduce missed diagnosis risks. Additionally, CSA-TTA demonstrates notable regression performance, with reductions in MAE and MSE of 3.24% (6.49 → 6.28) and 8.07% (92.77 → 85.28) on VitalDB, and 4.17% (6.00 → 5.75) and 5.28% (88.02 → 83.37) on the in-hospital test set, respectively. These improvements underscore CSA-TTA's ability to effectively address the challenges of personalized IOH dynamics, enhancing predictive accuracy and reducing regression errors by adapting to rapid signal shifts and rare hypotensive events.

Model		TimesFM(Das et al. 2024)						UniTS(Gao et al. 2024)					
		F1↑	Rec↑	Prec↑	Acc↑	MAE↓	MSE↓	F1↑	Rec↑	Prec↑	Acc↑	MAE↓	MSE↓
Zero-shot	Test	70.23	60.77	83.28	86.63	6.00	88.02	56.10	43.77	80.09	82.93	6.45	91.97
	Random samples	70.30	60.70	83.53	86.57	5.93	86.40	56.17	42.03	**85.43**	83.10	6.40	90.12
	CSA-TTA (Ours)	**71.60**	**62.44**	**84.03**	**87.20**	**5.75**	**83.37**	**63.80**	**53.33**	80.43	**84.47**	**6.30**	**88.69**
	Δ(%)	▲1.37	▲1.67	▲0.75	▲0.57	▲4.17	▲5.28	▲7.70	▲9.56	▲0.34	▲1.54	▲2.33	▲3.57

Table 2: Performance comparison on the in-hospital test set. All results are averaged over prediction horizons (5, 10, and 15 minutes), using a fixed 15-minute lookback window.

Config			VitalDB Dataset					
Horizon	Pred	Recon	F1↑	Rec↑	Prec↑	Acc↑	MAE↓	MSE↓
5min	✗	✓	70.00	**71.90**	68.20	91.50	4.82	55.81
	✓	✗	**70.60**	71.40	**69.80**	**91.80**	4.79	54.08
	✓	✓	**70.60**	71.60	69.70	**91.80**	**4.77**	**53.17**
10min	✗	✓	64.60	60.50	69.30	87.30	6.20	82.14
	✓	✗	64.40	59.50	**70.20**	87.40	6.15	80.49
	✓	✓	**64.70**	**60.80**	69.20	**87.50**	**6.05**	**77.60**
15min	✗	✓	62.80	**55.30**	72.70	84.60	6.70	90.02
	✓	✗	62.80	54.59	**74.00**	**84.80**	6.69	89.29
	✓	✓	**62.90**	54.60	**74.00**	**84.80**	**6.64**	**88.02**

Table 3: Ablation study on multi-task optimization. Performance of TimesFM in the fine-tuning setting using different optimization strategies: Pred (supervised prediction) and Recon (self-supervised masked reconstruction).

Fine-Tuning Setting. CSA-TTA shows consistent improvements across all performance metrics. As shown in Table 1(b), TimesFM + CSA-TTA improves Recall by 1.06% and F1 by 0.63% on VitalDB (2S), surpassing all other test-time adaptation methods. Units + CSA-TTA achieves even greater gains, with Recall increasing by 1.33%, F1 by 1.13%, Precision by 0.99%, and Accuracy by 0.20%, demonstrating consistent improvements across multiple metrics. Similar trends appear on VitalDB (30S), where Units + CSA-TTA raises Recall by 0.86% and F1 by 1.74%. Regarding regression metrics, CSA-TTA further improves performance, with the TimesFM-based model reducing MAE and MSE by 2.02% and 4.38%, respectively, on VitalDB (30S). These improvements, particularly in Recall, are clinically significant. In intraoperative hypotension prediction, higher Recall reduces false negatives—missed hypotensive events that can lead to myocardial injury or death. Even modest Recall gains can improve patient outcomes. CSA-TTA enhances Recall while maintaining stable Precision, helping to reduce unnecessary interventions.

Case Study. To compare CSA-TTA with the baseline, we evaluate vanilla TimesFM and TimesFM + CSA-TTA on two VitalDB (2S) cases. At the 5-minute horizon (Figure 4a), the static model smooths the blood pressure drop, while CSA-TTA—guided by cross-sample retrieval and multi-task adaptation—closely follows the ground truth.

Config		Zero-Shot Setting					
Model	Top-K	F1↑	Rec↑	Prec↑	Acc↑	MAE↓	MSE↓
TimesFM	1	64.57	58.83	71.97	87.83	6.32	86.96
	2	64.73	59.03	72.13	87.80	**6.27**	86.28
	3	**64.90**	**59.27**	**72.20**	**87.93**	6.28	**85.28**

Table 4: Ablation study on Top-K of coarse-to-fine cross-sample retrieval. Performance of the TimesFM backbone under zero-shot settings.

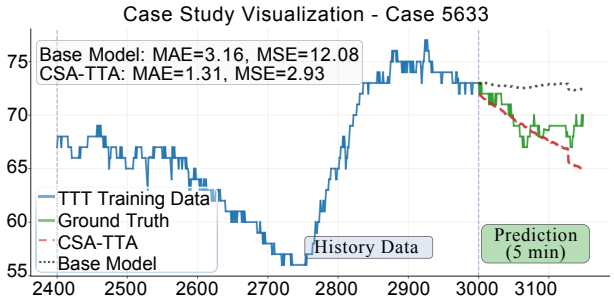
Model	#ToP	#TuP	#PuP	#Time(Avg)
TimesFM	477.36M	5.05M	1.06%	6.587s
Units	1.01M	0.05M	5.44%	1.688s

Table 5: Computational cost of CSA-TTA per epoch (30-minute history window). "#ToP" and "#TuP" indicate the total and fine-tuned parameter counts, respectively. "#PuP" is the percentage of parameters updated, and "#Time" denotes the average wall-clock time per TTA epoch.

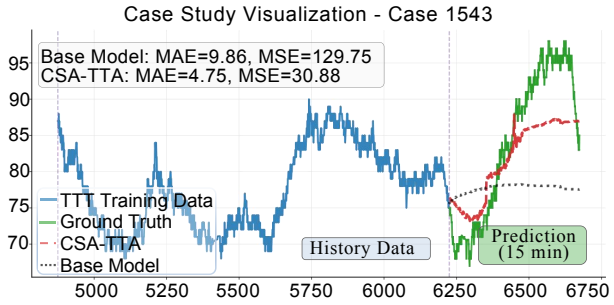
This advantage continues at the 15-minute horizon (Figure 4b), where CSA-TTA accurately captures the sharp decline and rebound, reducing errors by more than half (MAE: 9.86 → 4.75; MSE: 129.75 → 30.88). These examples show how incorporating patient-specific context at inference allows TimesFM to model abrupt, individualized IOH dynamics missed by static models.

Ablation Study

Multi-Task Optimization. Table 3 compares supervised prediction optimization ("Pred"), self-supervised reconstruction optimization ("Recon"), and their combination on TimesFM under the fine-tuning setting across 5-, 10-, and 15-minute horizons. Multi-task optimization consistently outperforms the single-task variants, achieving the best F1, MAE, and MSE across all settings. It also yields the highest Recall at the 10-minute horizons (60.80%). While single-task setups occasionally show higher Recall—for instance, "Recon"-only achieves 71.90% at 5 minutes—they perform worse on other metrics. These results highlight the effectiveness of multi-task optimization in adapting to personalized intraoperative hypotension dynamics.



(a) 5-min horizon



(b) 15-min horizon

Figure 4: Case study visualizations on VitalDB (2-second sampling). TimesFM + CSA-TTA predicts 5-minute (top) and 15-minute (bottom) horizons.

Top-K of retrieval. We ablate the Top-K hyperparameter in the fine-grained retrieval stage of CSA-TTA using a TimesFM backbone on VitalDB under zero-shot settings. As shown in Table 4, with results averaged over 5/10/15-minute horizons, increasing K yields improved or stable F1, Recall, and regression errors. For example, Recall rises from 58.83 (K=1) to 59.27 (K=3), while MSE drops from 86.96 to 85.28. We therefore adopt K=3, which balances relevance and diversity and improves robustness and accuracy.

Augmentation Strategy. Integrating the cross-sample bank with small perturbations consistently enhances model performance. As shown in Table 6, using the cross-sample bank alone improves F1 by 0.5% (64.47% → 64.97%) and Recall by 1.46% in the zero-shot setting on the VitalDB dataset. Adding small perturbations further boosts performance—under the fine-tuning setting, the combined approach increases Recall by 0.13% (62.20% → 62.33%) and reduces MAE by 1.69% (74.19 → 72.93). These results highlight that while each augmentation method offers individual benefits, their combination yields the best overall performance.

Computational Cost. To assess the computational cost, we report adaptation time in Table 5. By using partial fine-tuning (updating only 1.06%–5.44% of parameters), it enables efficient adaptation, with per-epoch latency as low as 1.7 seconds for lightweight models and 6.6 seconds for larger ones. This demonstrates its efficiency and practicality for real-world deployment.

Config			VitalDB Dataset					
Setting	Bank	Aug	F1↑	Rec↑	Prec↑	Acc↑	MAE↓	MSE↓
Zero-shot	✗	✗	64.47	58.67	71.90	87.80	6.33	86.70
	✗	✓	64.27	57.60	**72.97**	87.83	6.27	85.66
	✓	✗	**64.97**	**60.13**	71.13	87.77	**6.26**	**85.03**
	✓	✓	64.90	59.27	72.20	**87.93**	6.28	85.28
fine-tuned	✗	✗	65.90	62.17	70.82	87.90	5.87	74.79
	✗	✓	65.72	61.38	**71.43**	87.97	5.84	73.95
	✓	✗	66.03	62.20	71.10	88.00	5.85	74.19
	✓	✓	**66.07**	**62.33**	70.97	**88.03**	**5.82**	**72.93**

Table 6: Ablation study on data augmentation strategy. Performance of TimesFM under zero-shot and fine-tuned settings using different augmentation methods: Bank (cross-sample bank) and Aug (perturbation-based augmentation).

Setting / Data		VitalDB Dataset					
		F1↑	Recall↑	Prec↑	Acc↑	MAE↓	MSE↓
Zero-shot	Random	63.93	58.73	70.57	87.60	6.47	92.30
	CSA-TTA	**64.90**	**59.27**	**72.20**	**87.93**	**6.28**	**85.28**
Fine-tuned	Random	65.73	62.13	70.39	87.77	5.92	75.84
	CSA-TTA	**66.07**	**62.33**	**70.97**	**88.03**	**5.82**	**72.93**

Table 7: Ablation study on personalized adaptation. Performance of TimesFM + CSA-TTA with random data and personalized cross-sample data.

Personalized Adaptation. To validate the effectiveness of personalized adaptation, we compare against a "Random" baseline—CSA-TTA without personalized history or augmentation, using the same number of randomly selected sequences from other patients. Table 7 shows that, on VitalDB with TimesFM + CSA-TTA, F1 rises from 63.93 to 64.90 with a concurrent MSE drop from 92.30 to 85.28 in the zero-shot setting, while fine-tuning yields smaller yet consistent gains (F1: 65.73 → 66.07; MSE: 75.84 → 72.93). These results demonstrate that leveraging personalized history aligns the model more closely with the target distribution than random sampling.

Conclusions

We present CSA-TTA, the first test-time adaptation framework specifically designed for personalized intraoperative hypotension (IOH) prediction. CSA-TTA addresses the challenges of individual IOH dynamics by leveraging a cross-sample bank with a coarse-to-fine retrieval strategy to provide richer temporal context during adaptation. It further combines supervised and self-supervised objectives in a unified optimization framework to enhance adaptation stability and effectiveness. Extensive experiments on two real-world clinical datasets, using TimesFM and UniTS as backbones, demonstrate that CSA-TTA consistently improves prediction performance, offering a robust and efficient solution for real-time, personalized monitoring in surgical settings.

Acknowledgments

This work was partially supported by the Yunnan Provincial Department of Science and Technology-Major Science and Technology Special Project (202502AS080002), funding from the National Natural Science Foundation (81860218), the Yunling Scholar Talent Program of Yunnan Province under Grant No. (K264202230207), and the Deng Cheng Expert Workstation of Yunnan Province (202305AF150202).

References

Berndt, D. J.; and Clifford, J. 1994. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd international conference on knowledge discovery and data mining*, 359–370.

Bringmann, K.; Fischer, N.; van der Hoog, I.; Kipouridis, E.; Kociumaka, T.; and Rotenberg, E. 2024. Dynamic dynamic time warping. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 208–242. SIAM.

Cai, X.; Wang, X.; Zhu, Y.; Yao, Y.; and Chen, J. 2025. Advances in automated anesthesia: a comprehensive review. *Anesthesiology and Perioperative Science*, 3(1): 1–20.

Cheng, M.; Zhang, J.; Liu, Z.; Liu, C.; and Xie, Y. 2024. HMF: A Hybrid Multi-Factor Framework for Dynamic Intraoperative Hypotension Prediction. *CoRR*, abs/2409.11064.

Das, A.; Kong, W.; Sen, R.; and Zhou, Y. 2024. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning, ICML 2024*. OpenReview.net.

de Keijzer, I. N.; Vos, J. J.; Yates, D.; Reynolds, C.; Moore, S.; Lawton, R. J.; Scheeren, T. W.; and Davies, S. J. 2024. Impact of clinicians' behavior, an educational intervention with mandated blood pressure and the hypotension prediction index software on intraoperative hypotension: a mixed methods study. *Journal of Clinical Monitoring and Computing*, 38(2): 325–335.

Dong, S.; Wang, Q.; Wang, S.; Zhou, C.; and Wang, H. 2024. Hypotension prediction index for the prevention of hypotension during surgery and critical care: A narrative review. *Computers in Biology and Medicine*, 170: 107995.

Gao, S.; Koker, T.; Queen, O.; Hartvigsen, T.; Tsiligkaridis, T.; and Zitnik, M. 2024. Units: A unified multi-task time series model. *Advances in Neural Information Processing Systems*, 37: 140589–140631.

Hwang, E.; Park, Y.-S.; Kim, J.-Y.; Park, S.-H.; Kim, J.; and Kim, S.-H. 2023. Intraoperative hypotension prediction based on features automatically generated within an interpretable deep learning model. *IEEE Transactions on Neural Networks and Learning Systems*.

Jeong, H.; Kim, D.; Kim, D. W.; Baek, S.; Lee, H.-C.; Kim, Y.; and Ahn, H. J. 2024. Prediction of intraoperative hypotension using deep learning models based on non-invasive monitoring devices. *Journal of Clinical Monitoring and Computing*, 1–9.

Kang, J.; Kim, N.; Ok, J.; and Kwak, S. 2024. MemBN: Robust Test-Time Adaptation via Batch Norm with Statistics Memory. In *Computer Vision - ECCV 2024 - 18th European Conference*, volume 15086, 467–483. Springer.

Karmanov, A.; Guan, D.; Lu, S.; El Saddik, A.; and Xing, E. 2024. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14162–14171.

Lee, H.-C.; Park, Y.; Yoon, S. B.; Yang, S. M.; Park, D.; and Jung, C.-W. 2018. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9(1): 279.

Lee, S.; Lee, H.-C.; Chu, Y. S.; Song, S. W.; Ahn, G. J.; Lee, H.; Yang, S.; and Koh, S. B. 2021. Deep learning models for the prediction of intraoperative hypotension. *British journal of anaesthesia*, 126(4): 808–817.

Li, W.; Hu, Z.; Yuan, Y.; Liu, J.; and Li, K. 2022. Effect of hypotension prediction index in the prevention of intraoperative hypotension during noncardiac surgery: a systematic review. *Journal of Clinical Anesthesia*, 83: 110981.

Liang, J.; He, R.; and Tan, T. 2025. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1): 31–64.

Lim, H.; Kim, B.; Choo, J.; and Choi, S. 2023. TTN: A Domain-Shift Aware Batch Normalization in Test-Time Adaptation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Liu, Y.; Kothari, P.; Van Delft, B.; Bellot-Gurlet, B.; Mordan, T.; and Alahi, A. 2021. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 34: 21808–21820.

Lu, F.; Li, W.; Zhou, Z.; Song, C.; Sun, Y.; Zhang, Y.; Ren, Y.; Liao, X.; Jin, H.; Luo, A.; et al. 2023. A composite multi-attention framework for intraoperative hypotension early warning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 14374–14381.

Meng, L.; Yu, W.; Wang, T.; Zhang, L.; Heerdt, P. M.; and Gelb, A. W. 2018. Blood pressure targets in perioperative care: provisional considerations based on a comprehensive literature review. *Hypertension*, 72(4): 806–817.

Moe, T. E.; Pereira, T. M.; Calvo, F.; and Leenaarts, J. 2023. Shape-based clustering of synthetic Stokes profiles using k-means and k-Shape. *Astronomy & Astrophysics*, 675: A130.

Mohammadi, I.; Firouzabadi, S. R.; Hosseinpour, M.; Akhlaghpasand, M.; Hajikarimloo, B.; Tavanaei, R.; Izadi, A.; Zeraatian-Nejad, S.; and Eghbali, F. 2024. Predictive ability of hypotension prediction index and machine learning methods in intraoperative hypotension: a systematic review and meta-analysis. *Journal of Translational Medicine*, 22(1): 725.

Mukkamala, R.; Schnetz, M. P.; Khanna, A. K.; and Mahajan, A. 2025. Intraoperative hypotension prediction: current methods, controversies, and research outlook. *Anesthesia & Analgesia*, 141(1): 61–73.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Paparrizos, J.; and Gravano, L. 2016. k-Shape: Efficient and Accurate Clustering of Time Series. *SIGMOD Rec.*, 45(1): 69–76.

Senin, P. 2008. Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, 855(1-23): 40.

Shi, M.; Zheng, Y.; Wu, Y.; and Ren, Q. 2023a. Multitask attention-based neural network for intraoperative hypotension prediction. *Bioengineering*, 10(9): 1026.

Shi, M.; Zheng, Y.; Wu, Y.; and Ren, Q. 2023b. Multitask attention-based neural network for intraoperative hypotension prediction. *Bioengineering*, 10(9): 1026.

Shi, W.; Xu, R.; Zhuang, Y.; Yu, Y.; Sun, H.; Wu, H.; Yang, C.; and Wang, M. D. 2024. Medadapter: Efficient test-time adaptation of large language models towards medical reasoning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2024, 22294.

Sidiropoulou, T.; Tsoumpa, M.; Griva, P.; Galarioti, V.; and Matsota, P. 2022. Prediction and prevention of intraoperative hypotension with the hypotension prediction index: a narrative review. *Journal of Clinical Medicine*, 11(19): 5551.

Sun, Y.; Wang, X.; Liu, Z.; Miller, J.; Efros, A.; and Hardt, M. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, 9229–9248. PMLR.

Tan, M.; Chen, G.; Wu, J.; Zhang, Y.; Chen, Y.; Zhao, P.; and Niu, S. 2025. Uncertainty-calibrated test-time model adaptation without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wesselink, E.; Kappen, T.; Torn, H.; Slooter, A.; and Van Klei, W. 2018. Intraoperative hypotension and the risk of postoperative adverse outcomes: a systematic review. *British journal of anaesthesia*, 121(4): 706–721.

Woo, G.; Liu, C.; Kumar, A.; Xiong, C.; Savarese, S.; and Sahoo, D. 2024. Unified Training of Universal Time Series Forecasting Transformers. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. PMLR.

Yang, J.; Ning, C.; Deb, C.; Zhang, F.; Cheong, D.; Lee, S. E.; Sekhar, C.; and Tham, K. W. 2017. k-Shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. *Energy and Buildings*, 146: 27–37.

Yoon, J. H.; Jeanselme, V.; Dubrawski, A.; Hravnak, M.; Pinsky, M. R.; and Clermont, G. 2020. Prediction of hypotension events with physiologic vital sign signatures in the intensive care unit. *Critical Care*, 24: 1–9.