AUTOADAPT: RELIABLE FEW-SHOT ADAPTATION UNDER CLINICAL DISTRIBUTION SHIFTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large pretrained clinical models provide a practical way to reuse learned prior knowledge across hospitals by adapting models to them. In practice, a target hospital may only have a small labeled patient cohort, a setting commonly referred to as few-shot adaptation. This requires making multiple decisions, such as which pretrained model to adapt, how much of the model to update, and which patients to use. Nevertheless, this process faces two primary challenges. First, the best adaptation strategy varies across clinical tasks. Second, evaluating and comparing candidate strategies becomes unreliable due to the small patient cohort. In this work, we introduce AUTOADAPT with two core designs to deal with these challenges. The *Adapter* defines an extensible space of adaptation recipes, and the *Automator* forms a weighted recipe combination from evidence within the adaptation patients. We propose a reliability rule to ensure that only the most effective strategy on most available patients will be selected. These selected strategies then form a combination for effective few-shot adaptation. We conduct extensive experiments across critical care, emergency care, and diagnostic datasets, and the results show that AUTOADAPT consistently achieves state-of-the-art performance using only a few patients for adaptation.

1 INTRODUCTION

Clinical models often face distribution shifts when deployed to new hospitals, where differences in patient data can degrade model performance. Although adapting models trained on other datasets to the target hospital can mitigate such shifts, labeled data from the new target distribution are often scarce. Particularly, clinical outcomes can be costly to annotate or unavailable for many patients, leaving only a small labeled cohort for adaptation. This limitation persists even in large clinical datasets, as many outcomes (e.g., acute kidney injury or vasopressor initiation) provide only one independent label per patient (Johnson et al., 2016; 2023; Moody et al., 2022). As an example, in the MIMIC-IV cohorts in Figure 1, only 8–30 positive patients are available per task.

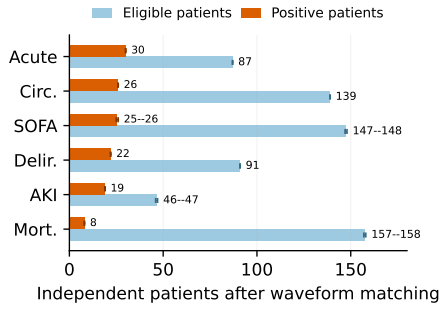


Figure 1: Independent patient support in the MIMIC-IV cohorts.

Pretrained physiological models provide a possible way for adaptation by transferring knowledge learned from large source datasets (Kiyasseh et al., 2021; Gopal et al., 2021). However, how to adapt a pretrained model to a new target distribution remains an important and complex decision. For example, it requires choosing which source checkpoint to transfer and how much of the model to update, which ranges from using a fixed encoder to partial or full parameter fine-tuning. The effectiveness of these choices can vary substantially across target tasks and distributions. Particularly, a specialized checkpoint may transfer well to a related outcome but poorly to another, while aggressive updating may help when sufficient labeled data is available but overfit when labels are scarce. Thus, there is no universally effective adaptation strategy for a new clinical target.

This creates two challenges. The first is *adaptation heterogeneity*: the appropriate checkpoint and update rule can vary with the target clinical task and distribution. As a result, an adaptation strategy

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.

that works well for one target may perform poorly on another, making a fixed adaptation choice unreliable across different distributions. The second is *selection uncertainty*: determining which strategy to use requires comparing candidates using a small cohort for adaptation. With only a few labeled patients, these comparisons can be highly unstable, and the apparent advantage of one strategy may be driven by only a small number of patients. Together, these challenges make reliable adaptation difficult when target data is limited (Ben-David et al., 2010; Zhang et al., 2021; Guo et al., 2022).

In this work, we introduce AUTOADAPT to address these challenges. It consists of two components: (1) The *Adapter* addresses adaptation heterogeneity by jointly composing choices of checkpoint, update, data, and input into complete recipes, capturing their interactions. (2) The *Automator* addresses selection uncertainty by comparing recipes based on the limited target data, while removing unreliable choices when the evidence is insufficient. Together, AUTOADAPT enables task-specific adaptation under scarce labels. Our main contributions are summarized as follows:

- We introduce an extensible *Adapter* that represents adaptation strategies as complete recipes, jointly specifying the source checkpoint, update rule, target data policy, and input route. This allows different adaptation choices to be considered together rather than optimized independently.
- We develop a reliability-aware *Automator* that uses held-out patient predictions, task information, and label-free shift measures to construct a task-specific weighted recipe combination.
- We evaluated AUTOADAPT across the ICU transfer, emergency care, and diagnostic ECG datasets. The results demonstrate improvements over fixed and existing adaptation strategies, including in settings with only one to six positive adaptation patients.

2 RELATED WORK

Clinical time-series benchmarks made ICU prediction reproducible by fixing cohorts, observation windows, and outcomes (Harutyunyan et al., 2019). Physiological pretraining asks a complementary question: whether structure learned without target labels can support new tasks. CLOCS and 3KG exploit cross-view or temporal agreement in biosignals (Kiyasseh et al., 2021; Gopal et al., 2021), while general time-series models seek one representation that transfers broadly (Goswami et al., 2024). These methods make source knowledge available. They do not determine which source checkpoint should be trusted for a particular target outcome.

Domain adaptation studies distinguish a change in observed inputs from a change in the relation between inputs and labels (Ben-David et al., 2010). This distinction is especially important in medicine because performance can vary with site, time, and care process (Zhang et al., 2021; Guo et al., 2022). Frozen transfer, partial fine-tuning, LoRA, and full fine-tuning make different assumptions about how much of that relation remains stable (Hu et al., 2022). They are candidate responses to shift rather than a rule for deciding which response is supported by a small target cohort.

Several lines of work address parts of this decision. PRAM retrieves related local patients (Jeong et al., 2026), and OTTEHR transports structured EHR features across domains (Li et al., 2026). LEEP, LogME, and TransRate score the transferability of a representation (Nguyen et al., 2020; You et al., 2021; Huang et al., 2022). AutoPEFT searches over parameter-efficient updates (Zhou et al., 2024), while TEMPLATE ranks pretrained time-series models (Zhang et al., 2025). These approaches typically optimize one component of transfer or presume that the observed ranking is reliable. AUTOADAPT instead compares complete recipes and treats the reliability of that comparison as part of the adaptation problem. It can use a specialist, update the model, or add a modality when the patient-level evidence supports that choice, and otherwise retain a predefined default recipe.

3 PROBLEM SETTING

We consider a target clinical prediction task t at a new site. Let V denote physiological waveforms, Z denote optional structured EHR variables, and Y_t denote the target outcome. The available source assets are a finite library $\mathcal{C} = \{c_1, \ldots, c_K\}$ of pretrained checkpoints. A checkpoint may be a general waveform encoder or an encoder later specialized for a source outcome. Checkpoint c_k was learned under source distribution $P_{s_k}(V, Z, Y_k)$, whereas the target follows $P_t(V, Z, Y_t)$. Transfer

may therefore involve both input shift and a change in the clinical prediction target:

$$P_{s_k}(V, Z) \neq P_t(V, Z), \qquad P_{s_k}(Y_k \mid V, Z) \neq P_t(Y_t \mid V, Z). \tag{1}$$

The first inequality captures changes in population, sensors, and acquisition. The second captures the fact that a source outcome may only be partially related to Y_t. AutoAdapt measures these two forms of compatibility separately.

Each target task specifies an index time τ, observation length O_t, gap G_t, and prediction horizon H_t. For patient i, the input contains waveform windows $X_i = (x_{ij})_{j=1}^{m_i}$ and optional structured context z_i observed in $[\tau - O_t, \tau]$. The binary label y_i records whether the target event occurs in $(\tau + G_t, \tau + G_t + H_t]$. The model outputs one patient risk $\hat{p}_i = \hat{f}_t(X_i, z_i)$ after aggregating repeated-window scores. Task-specific values of O_t, G_t, and H_t are reported in Appendix Table 4.

The target patients are divided into a small labeled adaptation cohort $\mathcal{A}_t$ and a patient-disjoint evaluation cohort $\mathcal{E}_t$. All patient-specific model fitting and numerical combination-weight estimation use $\mathcal{A}_t$. Higher-level structural choices are treated separately, and their selection scope is stated in the experimental protocol. Because windows from one patient are correlated, the patient is the evidence unit for fitting, internal validation, and uncertainty. Evaluation labels do not fit candidate predictors or numerical combination weights.

4 METHOD

AUTOADAPT separates two responsibilities that ordinary fine-tuning treats as one decision. The *Adapter* defines and executes a finite library of adaptation recipes. The *Automator* uses the small adaptation cohort to weight complete recipes for the current task. Figure 2 shows this interface. The separation allows the recipe library to grow without assuming that a larger search space can be ranked reliably from a few patients.

4.1 ADAPTER

Four candidate spaces. The Adapter starts from four finite, explicitly enumerated spaces:

- $\mathcal{C}_t \subseteq \mathcal{C}$ is the *checkpoint space*. It contains source representations compatible with the target signal geometry, including the general pretrained encoder, task-specialized encoders, and declared fixed checkpoint combinations.
- $\mathcal{U}_t$ is the *update space*. Its choices are source-head reuse, a new head on a frozen encoder, partial fine-tuning, LoRA, and full fine-tuning. Each choice specifies exactly which parameters may change.
- $\mathcal{D}_t$ is the *target-data-policy space*. A policy fixes class or patient weighting, the number and type of waveform windows sampled per patient, and the rule used to aggregate window predictions.
- $\mathcal{M}_t$ is the *input-route space*. Depending on data availability, it contains waveform only, EHR only, early fusion of waveform and EHR representations, and late fusion of their predictions.

These are spaces of concrete executable choices, not learned embedding spaces. Unavailable or incompatible choices are removed before adaptation. The valid recipe library for task t is therefore

$$\mathcal{R}_t = \{(c, u, d, m) \in \mathcal{C}_t \times \mathcal{U}_t \times \mathcal{D}_t \times \mathcal{M}_t : \mathrm{valid}_t(c, u, d, m) = 1\}. \tag{2}$$

Here $r = (c, u, d, m)$ is one complete recipe. The validity function checks, for example, that a waveform checkpoint receives the required channels and that a fusion route is proposed only when both modalities are present. The four choices are kept together because their effects interact. A specialist may work when frozen but overfit after full updating, and the useful waveform checkpoint may change when EHR context is added.

Executing a recipe. For recipe r, let $S_d(i)$ be the waveform windows selected for patient i, $w_d(y_i)$ its class weight, and A_d its patient-level aggregation rule. The route m determines whether the predictor consumes x_{ij}, z_i, or both. For an EHR-only route, waveform windows are ignored and A_d is the identity. The Adapter fits only the parameter set $\Omega_u(c, m)$ permitted by update choice u:

$$\hat{\omega}_t^r = \arg \min_{\omega \in \Omega_u(c,m)} \frac{1}{|\mathcal{A}_t|} \sum_{i \in \mathcal{A}_t} w_d(y_i)\, \ell(y_i, A_d(\{f_{c,m}(x_{ij}, z_i; \omega) : j \in S_d(i)\})). \tag{3}$$

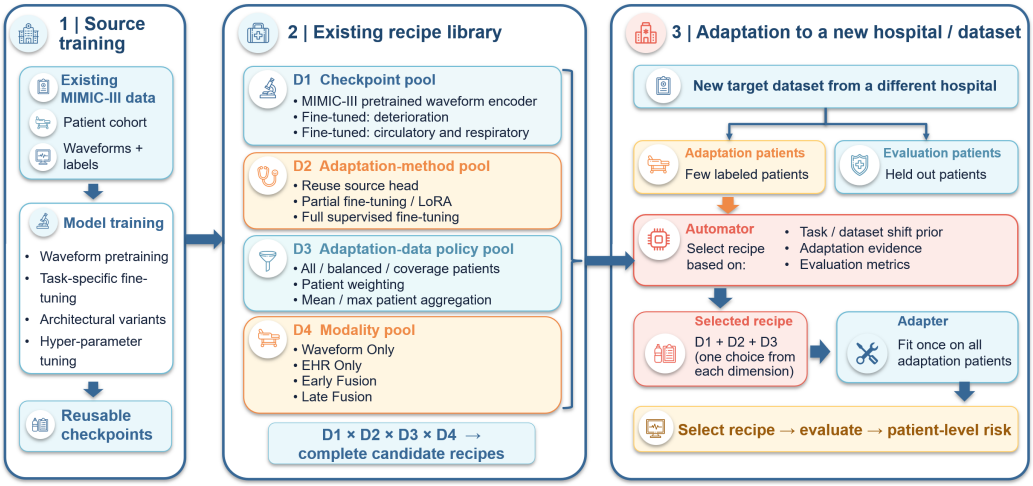


Figure 2: Overview of AUTOADAPT. The Adapter defines a recipe through its source representation, update capacity, target data policy, and input route. The Automator scores complete recipes using held-out adaptation patients. A shared evidence map forms a compact weighted combination, while the reliability check protects a predefined default recipe when the comparison is uncertain. Participating recipes are refit on the adaptation cohort before evaluation begins.

Here $f_{c,m}$ is the predictor assembled from checkpoint c and input route m, and ℓ is the supervised target loss (binary cross-entropy for the binary tasks studied here). The loss is averaged over patients rather than windows, so a long recording does not contribute more independent evidence. Equation 3 produces one fitted patient-risk predictor $\hat{f}_t^r$ for every valid recipe.

4.2 AUTOMATOR

The Automator runs separately for every target task and returns weights over $\mathcal{R}_t$. It uses three inputs: a task and shift prior, internal evidence from the adaptation cohort, and an estimate of how reliable that evidence is. The same Automator parameters are shared across tasks; only its inputs change.

Task and shift prior. For each valid recipe $r = (c, u, d, m)$, the prior combines the clinical relation between source checkpoint c and target task t with label-free compatibility between the source and target representations produced by route m:

$$\pi_t(r) = \alpha\ \mathrm{sim}_{\mathrm{clinical}}(t, c) + (1 - \alpha)\,[1 - \Delta_{\mathrm{input}}(t, c, m)]\,. \tag{4}$$

Here $\alpha \in [0, 1]$ is a global mixing coefficient and $\mathrm{sim}_{\mathrm{clinical}}(t, c) \in [0, 1]$ is the Jaccard overlap between fixed source- and target-task concept sets. Thus it is one when the two sets match and zero when they share no concept. The quantity $\Delta_{\mathrm{input}}(t, c, m) \in [0, 1]$ measures how easily unlabeled source patients can be distinguished from target adaptation patients under checkpoint c and route m; zero means indistinguishable and one means completely separated. Fixed checkpoint combinations average their component priors. Neither term uses clinical outcome labels.

Evidence from adaptation patients. No additional patient cohort is reserved for this step. Instead, $\mathcal{A}_t$ is partitioned into patient-level internal folds. For each fold, recipe r is fitted on the other folds and predicts the patients in the held-out fold. Let $\hat{p}'_{i,r}$ denote the resulting prediction for patient i; the prime indicates that patient i was excluded from the corresponding fit. Repeating this process yields exactly one internally held-out prediction for every patient in $\mathcal{A}_t$, without using $\mathcal{E}_t$. Let $\mathcal{S}_t$ denote the task-appropriate selection score, oriented so that larger is better. The adaptation utility is

$$Q_t(r) = \mathcal{S}_t\big(\{(y_i, \hat{p}'_{i,r}) : i \in \mathcal{A}_t\}\big)\,. \tag{5}$$

The score $\mathcal{S}_t$ may be any prespecified validation utility appropriate to the task. Crucially, Q_t scores the complete tuple (c, u, d, m); the four dimensions are not ranked independently.

Combining recipes from adaptation evidence. A high $Q_t(r)$ from a few patients does not by itself justify replacing the predefined default recipe. The recipe feature vector $\phi_t(r)$ contains its internally held-out utility, the fraction of leave-one-positive-patient recomputations that preserve its advantage, its prior $\pi_t(r)$, its average prediction disagreement with other recipes, its number of updated parameters, and one-hot indicators of (c, u, d, m). The task evidence vector h_t contains the numbers of positive and negative adaptation patients, adaptation prevalence, and the fraction of recipes with identifiable internal scores. A linear scoring map with coefficients θ produces

$$a_t(r) = \theta^\top [\phi_t(r);\ \phi_t(r) \otimes h_t], \qquad \widetilde{w}_t = \mathcal{N}_\lambda((a_t(r))_{r\in\mathcal{R}_t}), \tag{6}$$

where $\mathcal{N}_\lambda$ is the globally fixed regularized map from recipe scores to recipe-combination weights. The coefficients θ, the choice between nonnegative weights and a regularized signed correction, and the regularization λ are system-level structural choices. Their selection scope is reported in Section 5; once a configuration is chosen, these values remain fixed while its numerical task-specific weights are estimated from $\mathcal{A}_t$. A one-hot vector is the special case of selecting one recipe.

Finally, a fixed reliability rule maps positive-patient support, the bootstrap probability that the weighted combination improves over the default, and the leave-one-positive-patient stability defined above to $\rho_t \in [0, 1]$. The reference recipe $r_{0,t}$ is a simple valid recipe declared as the default before target evaluation. Let $e_{r_{0,t}}$ be its one-hot weight vector. The final weights and prediction are

$$w_t = \rho_t \widetilde{w}_t + (1 - \rho_t) e_{r_{0,t}}, \qquad \hat{f}_t(X_i, z_i) = \sum_{r\in\mathcal{R}_t} w_t(r) \hat{f}_t^r(X_i, z_i). \tag{7}$$

If the adaptation comparison is not identifiable, $\rho_t = 0$ and the default recipe is preserved. Otherwise, the Adapter refits every recipe with nonzero weight on all of $\mathcal{A}_t$. The weights, fitted recipes, and patient aggregation rules are then locked before $\mathcal{E}_t$ is scored. Adding a checkpoint, update rule, data policy, or modality route expands $\mathcal{R}_t$ without changing the Automator interface.

5 Experimental Setup

5.1 Datasets And Baselines

The evaluation separates hospital transfer from more distant changes in care setting and acquisition. MIMIC-III supplies the source waveform checkpoints (Johnson et al., 2016; Goldberger et al., 2000). ALOTT is the primary method-development target because it changes the hospital while preserving the ICU setting. It tests whether AutoAdapt can reuse source physiology for new outcomes without conflating the result with a completely different acquisition protocol (Lawrence et al., 2025). MIMIC-IV is a second development benchmark and provides the sharper scarcity test through its much smaller EHR and waveform cohort (Johnson et al., 2023; Moody et al., 2022). The remaining datasets test where the same decision rule remains useful. MC-MED changes the setting from intensive care to emergency care (Kansal et al., 2025). CPSC 2018 and Ningbo replace continuous bedside monitoring with diagnostic ECGs acquired under different protocols (Liu et al., 2018; Zheng et al., 2020). We report results by dataset because these outcomes and evidence regimes are not interchangeable. Every included task has a reproducible prediction time and a patient-level outcome. We consider the following baselines: (1) standard few-shot baselines, including a logistic, prototype, or 5-NN head on a frozen encoder, partial fine-tuning, LoRA, or full fine-tuning (Hu et al., 2022), (2) existing few-shot adaptation baselines: PRAM, PRAM-MI, and OTTEHR (Jeong et al., 2026; Li et al., 2026), and (3) recent automatic composition rules. LoraHub learns convex candidate weights from adaptation predictions (Huang et al., 2024); AdaMerging estimates weights by entropy minimization on unlabeled target predictions (Yang et al., 2024); and HOSO selects the blend between the source route and a target-updated route using held-out adaptation evidence (Vorster et al., 2026).

5.2 Adaptation Protocol

We convert each waveform into a five-minute, 125-Hz input on the masked channel grid used by the source model. EHR features use only measurements available before the prediction time. Patients are disjoint across splits whenever patient identifiers are available. For the binary experiments, the generic score $\mathcal{S}_t$ is instantiated as patient-level AUROC. We instantiate Δ_{input} as $2|A_{\text{domain}} - 1/2|$, where A_{domain} is the cross-validated discrimination score of a balanced linear classifier that separates

Table 1: MIMIC-III→ALOTT few-shot comparison on six clinically central tasks selected for clinical relevance, patient support, waveform coverage, and task difficulty. Values are patient-level AUROC. Every method uses the same 32 positive and 32 negative adaptation patients per task.

Method	Ventr.	Crit. rhythm	Oxy. crisis	Short VT	Pause/ asystole	ST shift	Mean
Fixed recipe baselines							
Frozen source encoder + target head	.512	.556	.515	.516	.567	.622	.548
Full fine-tuning	.642	.663	.560	.677	.654	.731	.655
LoRA	.640	.662	.557	.672	.656	.716	.651
Existing adaptation methods							
PRAM (Jeong et al., 2026)	.517	.515	.516	.545	.487	.518	.516
PRAM-MI (Jeong et al., 2026)	.527	.514	.524	.545	.487	.526	.520
OTTEHR (Li et al., 2026)	.552	.478	.629	.530	.581	.547	.553
LoraHub (Huang et al., 2024)	.634	.661	.514	.700	.657	.770	.656
AdaMerging (Yang et al., 2024)	.549	.627	.514	.600	.642	.716	.608
HOSO (Vorster et al., 2026)	.636	.662	.515	.668	.654	.753	.648
AUTOADAPT	**.677**	**.684**	**.612**	**.729**	**.692**	**.810**	**.701**

source from adaptation-patient embeddings. The clinical concept sets used in the Jaccard term are fixed before task evaluation.

Each ALOTT task uses the same 64 adaptation patients for every method, with 32 positive and 32 negative patients. Training is capped at eight waveform windows per patient so that recording duration cannot substitute for patient support. Recipe heads and task-specific evidence are learned only from these adaptation patients. The EHR baseline follows the same label budget. For the reported ALOTT analysis, combination weights and calibration are fitted on adaptation patients.

The available input-route space depends on the dataset. ALOTT exposes EHR-only, waveform-only, early-fusion, and late-fusion recipes; the selected combinations contain a fusion route in four of the six tasks in Table 1. MC-MED likewise allows EHR, waveform, and fusion, with EHR as the declared default recipe. The primary MIMIC-IV comparison fixes the route to waveform-only, while its EHR model is reported as a separate target-site baseline. CPSC and Ningbo contain no structured EHR route, so their recipe libraries are waveform-only. Thus $\mathcal{M}_t$ is selected where multiple modalities are available and becomes a singleton where they are not.

6 EXPERIMENTAL RESULTS

6.1 MAIN RESULTS ON ALOTT ICU TRANSFER

The ALOTT registry supplies twenty duration-qualified monitor-event tasks at a new ICU. Table 1 presents six clinically central tasks, and Appendix Table 3 reports the remaining fourteen. Every method receives the same 64 adaptation patients for each task. The comparison therefore measures how the same small cohort is used.

The strongest displayed baseline is LoraHub at .656 mean AUROC. In contrast, AUTOADAPT uses adaptation evidence to construct a task-specific weighted combination.

AUTOADAPT uses one shared evidence map across the six reported tasks and achieves a mean AUROC of .701. It leads the displayed baselines on five of six tasks, while OTTEHR is stronger on oxygenation crisis. Across the full twenty-task roster, AUTOADAPT reaches .686 mean AUROC versus .652 for the previous same-cohort reference. The consistent gains support the main mechanism: adaptation evidence is more useful when it controls a regularized combination than when it makes a winner-take-all choice. ALOTT is used for method development; the later datasets test the same selection principle under larger acquisition changes.

Table 2: MIMIC-IV scarce-label stress test using patient-level AUROC averaged over five evaluation runs. Every method uses the same adaptation and evaluation patients in each run. Underlining marks the strongest non-AutoAdapt result and boldface marks the highest result.

Method	Acute	Circ.	SOFA	Delir.	AKI	Mort.	Mean
Fixed recipe baselines							
Frozen source encoder + target head	.545	.596	.640	.547	.434	.637	.567
Full fine-tuning	.538	.585	.591	.519	.465	.665	.560
LoRA	.564	.586	.606	.513	.477	<u>.685</u>	.572
Existing-work baselines							
PRAM (Jeong et al., 2026)	<u>.574</u>	.605	<u>.669</u>	<u>.548</u>	.468	.649	<u>.585</u>
PRAM-MI (Jeong et al., 2026)	.565	.601	.657	.544	.464	.650	.580
OTTEHR (Li et al., 2026)	.486	.465	.583	.495	<u>.508</u>	.576	.519
LoraHub (Huang et al., 2024)	.542	.649	.650	.540	.405	.582	.561
AdaMerging (Yang et al., 2024)	.553	.614	.586	.539	.414	.613	.553
HOSO (Vorster et al., 2026)	.534	<u>.660</u>	.659	.536	.415	.582	.564
Ours							
AUTOADAPT	**.847**	**.934**	**.887**	**.842**	**.904**	**.847**	**.877**

6.2 EXTREME FEW-SHOT SELECTION ON MIMIC-IV

MIMIC-IV tests the selector where adaptation evidence is most fragile: across five fixed patient partitions and six tasks, each adaptation split contains only one to six positive patients, and every recipe is selected using only those adaptation patients. Table 2 shows that AUTOADAPT leads the displayed baselines on every task. Its shared evidence map combines patient-held-out adaptation performance with task similarity, distribution shift, target prediction geometry, and positive-patient support, without learning task- or fold-specific coefficients. This structure keeps weak evidence from dominating merely because many correlated waveform windows are available. The main finding is that reliable task-specific combinations can still be selected when independent positive patients, rather than raw window count, are the limiting resource.

6.3 CROSS-DATASET TRANSFER ON MC-MED AND DIAGNOSTIC ECG

The recipe library is useful only if it avoids forcing source physiology onto an incompatible target. MC-MED tests transfer from ICU monitoring to emergency care, while CPSC and Ningbo replace continuous bedside signals with diagnostic ECGs. As Figure 4 shows, AUTOADAPT retains the local EHR route on MC-MED unless adaptation evidence supports waveform fusion, but makes greater use of source waveform features when ECG structure remains compatible on CPSC and Ningbo. The cross-dataset finding is therefore not that source representations should always be reused, but that adaptation should depend on whether their signal structure survives the target shift. CPSC does not provide patient identifiers, so patient-disjoint splitting cannot be guaranteed for that dataset.

6.4 CHECKPOINT AND PREDICTION-HEAD ANALYSIS

A simpler system could choose a checkpoint and prediction head independently. Figure 3A tests this shortcut by pairing four source encoders with four frozen heads on the same target patients. Mean AUROC ranges from .505 to .626 across these pairs. The deterioration checkpoint with a cosine prototype is strongest on average, but the preferred pair changes across tasks. More importantly, the relative order of heads changes with the representation. This interaction rules out independent component ranking as an explanation for AUTOADAPT and motivates selection over complete recipes.

6.5 RELIABILITY-AWARE SELECTION ANALYSIS

Selecting the largest internal score is the most direct alternative to the Automator. We compare that rule with one strong fixed recipe and the complete reliability-aware selector while holding the candidate library and target patients constant in Figure 3B. Naive selection often overreacts to an unstable few-shot ranking and can be worse than keeping a fixed recipe. Task and shift priors narrow the plausible choices, while the reliability check rejects changes that lack patient-level support.

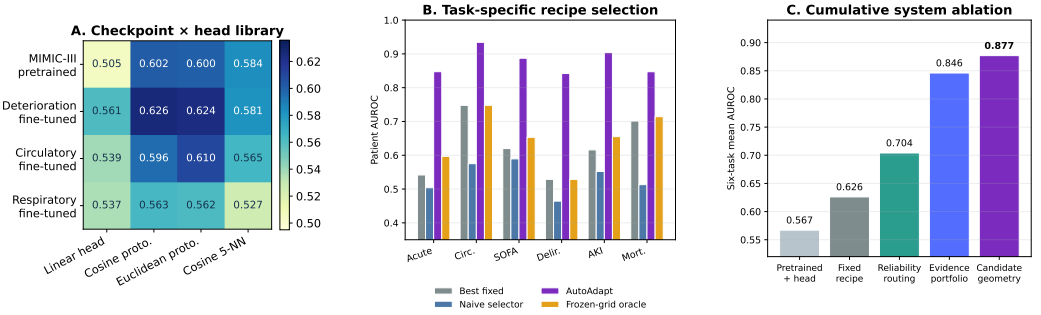


Figure 3: Checkpoint and recipe analysis on MIMIC-IV. Panel A tests source checkpoint and head pairings. Panel B compares a fixed recipe, naive adaptation-only selection, and the reliability-aware Automator. Panel C adds the Adapter and Automator components cumulatively.

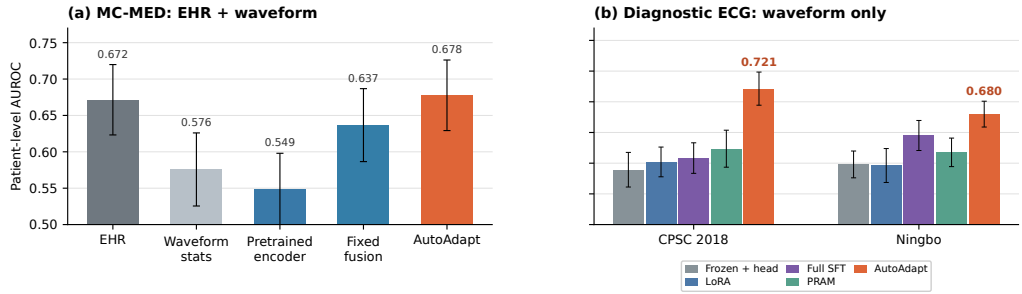


Figure 4: Cross-dataset comparisons. Panel A reports MC-MED macro AUROC over four tasks with sufficient patient support, comparing local EHR, waveform transfer, fixed fusion, and AUTOADAPT. Panel B reports macro AUROC over five shared diagnostic ECG tasks in CPSC and Ningbo. Methods use the same target folds within each dataset. The reference line marks chance AUROC.

Weighted combinations and target prediction geometry then preserve complementary candidates without trusting a single noisy winner. The central finding is therefore not that a larger search space is inherently better, but that reliability-aware evidence makes a diverse search space usable.

6.6 ADAPTER AND AUTOMATOR ABLATION STUDY

Figure 3C presents the cumulative ablation and tests whether a larger library alone explains the result. The stages start with the general pretrained encoder, add the strongest fixed recipe, introduce task and shift routing, and finally apply complete recipe selection with the reliability gate. Each component produces a cumulative improvement, but expanding the Adapter alone explains only the first gain. The larger improvement appears when the Automator routes among those alternatives using reliability and shift evidence. Keeping checkpoint, update-method, and data-policy evidence separate prevents one strong signal from being diluted, while candidate-level target geometry distinguishes redundant predictions from complementary ones. The ablation therefore supports the intended division of labor: the Adapter supplies useful alternatives, and the Automator decides when and how to use them.

6.7 CLINICAL-TASK AND WAVEFORM-SHIFT ANALYSIS

The prior assumes that clinical relation and signal compatibility describe different parts of transfer. We test this assumption with a fixed target head and data policy so that Figure 5 isolates source checkpoint routing. Clinical task distance correlates at $-.43$ with specialist transfer gain, while waveform distance correlates at $-.24$. Each association is informative but incomplete. Combining both signals with adaptation evidence raises mean AUROC from .562 to .646 and reduces harmful transfer by 67% relative to adaptation-only routing. The task relation identifies source knowledge that could matter, while the target waveform reveals whether that knowledge remains compatible with the new acquisition distribution.

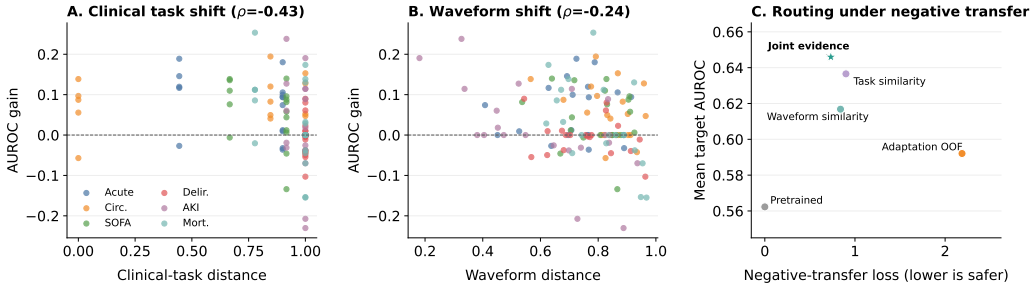


Figure 5: Checkpoint routing under distribution shift. Panels A and B show how transfer gain changes with task distance and waveform distance. Panel C compares routing rules. Every point uses the same target patients. Evaluation labels measure transfer gain but do not select a checkpoint.

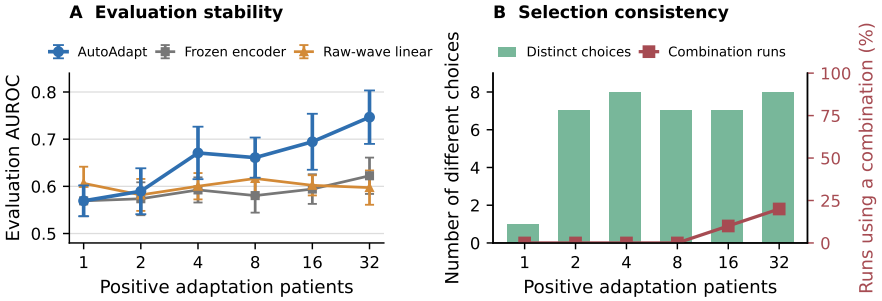


Figure 6: Stability on ALOTT new-pressor prediction as the adaptation panel grows from one to 32 positive patients, with an equal number of negative patients. Panel A reports evaluation AUROC over ten repeated panels against two fixed baselines. Panel B reports the number of distinct selections and the fraction using a recipe combination.

6.8 Checkpoint-Pool Composition Analysis

We grow the source pool from the general pretrained encoder to test whether checkpoint count alone helps. The deterioration specialist raises mean routing AUROC from .562 to .641, and the circulatory specialist raises it to .646. The respiratory specialist is never selected and adds no gain. The library therefore benefits from complementary physiological knowledge, not from accumulating redundant checkpoints.

6.9 Adaptation-Patient Budget And Stability Analysis

To test the MIMIC-IV failure mode directly, we vary positive-patient support rather than total window count. Ten balanced ALOTT adaptation panels are drawn at each budget from one to 32 positives while evaluation remains fixed (Figure 6). With one positive patient, the support rule always retains the pretrained default. From four positives onward, AutoAdapt exceeds both fixed baselines and reaches .747 mean AUROC at 32. Different panels still select different recipes, showing that more independent outcomes strengthen the gate without making the choice deterministic.

7 Conclusion

Adapting a clinical model with a few labeled patients is a decision problem as much as an optimization problem. The useful source checkpoint, update rule, data policy, and modality route change across tasks, while the cohort used to compare them may contain only a handful of positive outcomes. AutoAdapt addresses these two constraints with an extensible Adapter and a patient-reliability-aware Automator. The experiments show that this separation turns heterogeneous source models into useful task-specific predictors, while the predefined default limits unsupported transfer under larger shifts. Reliable recipe selection provides a practical path for reusing clinical models where target labels are scarce and no single adaptation rule is consistently appropriate.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, code drafting, and consistency checks. The authors reviewed the manuscript, source code, experimental artifacts, and references, and remain responsible for all claims and reported results.

ETHICS STATEMENT

This study uses retrospective, de-identified clinical datasets under their respective data-use agreements and access requirements. It involves no patient contact or intervention. We report aggregate results, make no attempt to re-identify individuals, and do not claim prospective clinical utility.

REPRODUCIBILITY STATEMENT

Appendix Section B records the data contract, patient counts, source-model architecture, trainable parameter counts, fixed seed, recipe definitions, and development provenance. The anonymous supplement contains machine-readable configurations, plotting data, and evaluation code. Access-controlled clinical data cannot be redistributed, but public dataset identifiers and label definitions are provided.

REFERENCES

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010. doi: 10.1007/s10994-009-5152-4.

E. Wesley Ely, Richard Margolin, Joseph Francis, Lisa May, Brenda Truman, Robert Dittus, Theodore Speroff, Shiva Gautam, Gordon R. Bernard, and Sharon K. Inouye. Evaluation of delirium in critically ill patients: validation of the confusion assessment method for the intensive care unit (CAM-ICU). *Critical Care Medicine*, 29(7):1370–1379, 2001. doi: 10.1097/00003246-200107000-00012.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Bryan Gopal, Ryan Han, Gautham Raghupathi, Andrew Ng, Geoff Tison, and Pranav Rajpurkar. 3KG: Contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations. In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pp. 156–167. PMLR, 2021. URL https://proceedings.mlr.press/v158/gopal21a.html.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16115–16152. PMLR, 2024. URL https://proceedings.mlr.press/v235/goswami24a.html.

Lin Lawrence Guo, Stephen R. Pfohl, Jason Fries, Alistair E. W. Johnson, Jose Posada, Catherine Aftandilian, Nigam Shah, and Lillian Sung. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Scientific Reports*, 12:2726, 2022. doi: 10.1038/s41598-022-06484-1.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6:96, 2019. doi: 10.1038/s41597-019-0103-9. URL https://doi.org/10.1038/s41597-019-0103-9.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International*

Conference on Learning Representations, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. LoraHub: Efficient cross-task generalization via dynamic LoRA composition. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=lyRpY2bpBh.

Long-Kai Huang, Junzhou Huang, Yu Rong, Qiang Yang, and Ying Wei. Frustratingly easy transferability estimation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9201–9225. PMLR, 2022. URL https://proceedings.mlr.press/v162/huang22d.html.

Inyong Jeong, Taeyeong Lee, Byeongsu Kim, Jin-Hyun Park, Yeongmin Kim, and Hwamin Lee. Pram: Post-hoc retrieval augmentation for parameter-free domain adaptation of icu clinical prediction models. *medRxiv*, pp. 2026–04, 2026.

Alistair E. W. 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*, 3:160035, 2016. doi: 10.1038/sdata.2016.35.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo Anthony Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023. doi: 10.1038/s41597-022-01899-x.

Aman Kansal, Emma Chen, Tom Jin, Pranav Rajpurkar, David Kim, et al. Multimodal clinical monitoring in the emergency department (mc-med). *Published online March*, 3, 2025.

Arif Khwaja. KDIGO clinical practice guidelines for acute kidney injury. *Nephron Clinical Practice*, 120(4):c179–c184, 2012. doi: 10.1159/000339789.

Dani Kiyasseh, Tingting Zhu, and David A. Clifton. CLOCS: Contrastive learning of cardiac signals across space, time, and patients. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5606–5615. PMLR, 2021. URL https://proceedings.mlr.press/v139/kiyasseh21a.html.

John Lawrence, Mike Rayo, and Timothy Huerta. ALarms, Outcomes Telemetry with Timing (ALOTT): a Bedside-EMR Database. *PhysioNet*, March 2025. doi: 10.13026/sbq5-dy17. URL https://doi.org/10.13026/sbq5-dy17. Version 1.0.0.

Wanxin Li, Saad Ahmed, Yongjin P. Park, and Khanh Dao Duc. Transport-based transfer learning on electronic health records: application to detection of treatment disparities. *Journal of the American Medical Informatics Association*, 33(1):15–25, 2026. doi: 10.1093/jamia/ocaf134.

Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, Jianqing Li, and Eddie Ng Yin Kwee. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018. doi: 10.1166/jmihi.2018.2442.

Benjamin Moody, Sicheng Hao, Brian Gow, Tom Pollard, Wei Zong, and Roger Mark. MIMIC-IV Waveform Database. *PhysioNet*, July 2022. doi: 10.13026/a2mw-f949. URL https://doi.org/10.13026/a2mw-f949. Version 0.1.0.

Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7294–7305. PMLR, 2020. URL https://proceedings.mlr.press/v119/nguyen20b.html.

Jean-Louis Vincent, Rui Moreno, Jukka Takala, Sheila Willatts, Arnaldo De Mendonca, Henk Bruining, Konrad Reinhart, Peter M. Suter, and L. G. Thijs. The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22:707–710, 1996. doi: 10.1007/BF01709751.

Chris Vorster, Mayug Maniparambil, Noel O'Connor, Noel Murphy, and Derek Molloy. Hold-one-shot-out (HOSO) for validation-free few-shot CLIP adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 7820–7829, 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Vorster_Hold-One-Shot-Out_HOSO_for_Validation-Free_Few-Shot_CLIP_Adapters_CVPRF_2026_paper.html.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. AdaMerging: Adaptive model merging for multi-task learning. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=nZP6NgD3QY.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12133–12143. PMLR, 2021. URL https://proceedings.mlr.press/v139/you21b.html.

Haoran Zhang, Natalie Dullerud, Laleh Seyyed-Kalantari, Quaid Morris, Shalmali Joshi, and Marzyeh Ghassemi. An empirical framework for domain generalization in clinical settings. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 279–290. ACM, 2021. doi: 10.1145/3450439.3451878.

Weiyang Zhang, Xinyang Chen, Xiucheng Li, Kehai Chen, Weili Guan, and Liqiang Nie. Unified transferability metrics for time series foundation models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 41155–41178. Curran Associates, Inc., 2025. doi: 10.52202/085713-1374. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3adfe6e3d8da4a75d174f466e1efc039-Paper-Conference.pdf.

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*, 7:48, 2020. doi: 10.1038/s41597-020-0386-x.

Han Zhou, Xingchen Wan, Ivan Vulić, and Anna Korhonen. AutoPEFT: Automatic configuration search for parameter-efficient fine-tuning. *Transactions of the Association for Computational Linguistics*, 12:525–542, 2024. doi: 10.1162/tacl_a_00662. URL https://aclanthology.org/2024.tacl-1.29/.

A Detailed Task And Data Contracts

Task labels are constructed from clinical events after the waveform cutoff. In Table 4, O/G/H gives the observation period, blank period, and prediction horizon in hours. Acute deterioration combines new organ support with death. Circulatory deterioration uses new pressor initiation. SOFA requires a future component increase of at least two (Vincent et al., 1996). Delirium requires a new positive CAM record (Ely et al., 2001). AKI uses incident urine output KDIGO stage 2 or 3 (Khwaja, 2012). Dynamic mortality uses future in hospital death. Windows are masked when the future state cannot be observed.

The primary MIMIC-IV input is waveform only. Every interval is resampled to 125 Hz and mapped to the fixed 62 channel vocabulary. Missing channels are zero with an explicit presence mask. We corrected inconsistent aliases for respiration and aVR before running the reported comparison. Labels and patient partitions remain unchanged.

B Reproducibility Details

Evaluation contract. Patient level AUROC is the primary metric. The MIMIC-IV table uses five executions recorded in a fixed manifest. Each execution assigns about 20% of patients to adaptation. The remaining patients are used for evaluation. Every MIMIC-IV analysis reuses these assignments and seeds. CPSC uses five native folds. Duplicate records remain in one fold, but patient disjointness cannot be established because identifiers are unavailable. Ningbo uses three participant disjoint folds with one ECG per participant. Their results are reported separately. The primary MIMIC-IV comparison uses waveform input. Structured EHR defines the outcome but does not enter the model. The multimodal analysis adds only features observed before the prediction cutoff.

Model and optimization. The source input is the final five minutes of the observation window on 62 canonical physical channels at 125 Hz, giving $x \in \mathbb{R}^{62\times 37,500}$. Missing channels are zero-filled with a presence mask. A magnitude FFT companion uses 250 samples and expands the input to 124 internal channels. Raw physical units are retained in the time branch. Only FFT segments are standardized. Concatenating 25 samples from each internal channel yields 1,500 tokens. TF-dual contains eight Transformer blocks with width 512. It uses eight attention heads, rotary attention, and an MLP width of 2,048. Mean pooling produces a 512 dimensional embedding. The checkpoint contains 37.7M parameters. The active classification path uses 26.81M of them.

The source manifest contains 398,204 one hour MIMIC-III windows. The loader exposes 4,778,448 five minute segments without overlap. Pretraining uses 3,500 optimizer steps with global batch 512. The deterioration, respiratory, and circulatory source tasks use 484,667, 551,232, and 578,274 windows. Their patient counts are 5,693, 4,421, and 4,955 after validation holdout. Partial fine tuning updates the final two blocks, the final LayerNorm, and the target head. This exposes 6,306,305 parameters. Rank 8 LoRA updates the attention projections and target head, which exposes 197,121 parameters. Full tuning updates the 26.81M parameter classification path. Optimizer and schedule settings belong to the selected recipe.

All runs use the same four TF-dual source states. The library contains the general pretrained encoder and three specialized encoders. Those specialists were trained for deterioration, circulatory decline, and respiratory decline. A machine readable manifest records each object name, size, digest, and source commit. Result generation verifies the manifest before loading predictions.

The task distance prior uses the fixed concept sets in Table 9. Evaluation labels do not alter these tokens.

Patient aggregation. Every metric is computed after combining a patient's windows into one risk. Each recipe fixes its aggregation rule before recipe combination. Candidate rules include the mean, maximum, upper quantile, and a bounded temporal summary. Bootstrap sampling operates on patients.

Compute and access. PhysioNet waveforms are read in WFDB form and resampled once into task manifests. Frozen methods reuse cached representations. Partial tuning, LoRA, and full tuning use GPU execution. Machine readable artifacts generate both the reported values and the figures.

Table 3: The fourteen ALOTT tasks not included in the six-task main panel. Values are patient-level AUROC on the same evaluation cohorts used for the main comparison.

Task	Prior reference	AUTOADAPT	Gain
AF vs sustained VT/VF alarm proxy	.727	.793	+.066
Critical rhythm vs short VT severity proxy	.612	.655	+.043
Critical rhythm vs ST-shift severity proxy	.663	.697	+.033
Critical rhythm vs VPC severity proxy	.597	.647	+.050
Critical vs noncritical electrical proxy	.569	.617	+.047
Pause vs sustained VT/VF proxy	.595	.653	+.057
Short VT vs nonrun ectopy proxy	.593	.629	+.035
Short VT vs pause proxy	.639	.660	+.021
VPC vs pause proxy	.649	.704	+.056
VPC vs short VT proxy	.645	.659	+.014
VPC vs ST-shift proxy	.682	.733	+.051
VT family vs pause proxy	.641	.689	+.048
Future VPC-burden escalation proxy	.659	.674	+.015
Respiratory crisis proxy	.560	.612	+.052

Table 4: Independent patient support in the five MIMIC-IV runs. Ranges show variation across patient partitions. Evaluation counts include patients with a usable prediction. Waveform row counts are omitted because repeated measurements are not independent events.

		All eligible		Adaptation/run		Evaluation/run	
Task	O/G/H (h)	Patients	Positive	Patients	Positive	Patients	Positive
Acute	6/0/6	87	30	17–18	6	69–70	24
Circ.	6/0/6	139	26	27–29	5–6	110–112	20–21
SOFA	6/6/24	147–148	25–26	29–30	5	117–119	20–21
Delir.	6/3/9	91	22	17–19	4–5	72–74	17–18
AKI	6/1/23	46–47	19	9–10	4	36–37	15
Mort.	6/6/24	157–158	8	31–32	1–2	125–127	6–7

Table 5: Source checkpoint library. Specialized checkpoints start from the general pretrained model and receive supervised MIMIC-III training. A source classifier is reused only when its label matches the target contract.

Checkpoint	MIMIC-III training state	Saved model	Target use
Pretrained	Masked time and frequency pretraining at selected step 1,300	37,699,612	Encoder with target fitted head
Deterioration	Pretrained start with escalation or death supervision at epoch 8	37,700,125	Encoder with aligned acute source head
Circulatory	Pretrained start with five hemodynamic labels at epoch 5	37,702,177	Encoder with aligned pressor source head
Respiratory	Pretrained start with five respiratory labels at epoch 15	37,702,177	Encoder only for main targets

Table 6: Target datasets and evaluation scope. Support gives positive evaluation patients after waveform quality control for ALOTT and MC-MED. The MIMIC-IV entry gives the full eligible range.

Target	Setting and input	Positive support	Purpose
ALOTT	ICU with EHR and bedside waveforms	16–779	Primary 64-patient transfer
MIMIC-IV	ICU waveform with EHR extension	8–30	Scarce label stress test
MC-MED	ED with EHR and bedside waveforms	23–162	Reliability of the default recipe
CPSC and Ningbo	Diagnostic 12 lead ECG	Native folds	Acquisition shift

Table 7: Target adaptation performed by each comparison. Candidate models and combination weights use labels from the declared adaptation patients.

Method	Target labels	Encoder update	Target operation
Traditional baselines			
Source head	No	None	Reuse the original source classifier unchanged
Frozen + head	Yes	None	Fit a new logistic head on frozen embeddings
Frozen + cosine prototype	Yes	None	Compare with adaptation class centroids using cosine similarity
Frozen + Euclidean prototype	Yes	None	Compare with adaptation class centroids using Euclidean distance
Frozen + cosine 5-NN	Yes	None	Vote among five nearby adaptation examples
Partial fine tuning	Yes	Last two blocks	Fit the final blocks and a new head
LoRA	Yes	Rank 8 adapters	Fit low rank attention updates and a new head
Full fine tuning	Yes	Active path	Update the classification path and new head
Existing work baselines			
PRAM	Yes	None	Mix predictions retrieved from adaptation patients
PRAM-MI	Yes	None	Use mutual-information-weighted Euclidean retrieval
OTTEHR	Yes	None	Transport frozen validation features toward the adaptation reference
LoraHub	Yes	None	Learn convex candidate weights from adaptation predictions
AdaMerging	No	None	Minimize prediction entropy on unlabeled target examples
HOSO	Yes	None	Select a source and target-updated prediction blend
Ours			
AUTOADAPT	Yes	Recipe dependent	Weight complete checkpoint, update, data-policy, and input-route recipes

Table 8: Fixed Automator thresholds. Minimum support controls eligibility. A challenger below the preferred count also needs a favorable paired confidence bound. Complexity resolves task-utility differences within δ_Q. A source head is reused only for an aligned label.

D2 method	Min. pos.	Preferred pos.	Rank	Updated params.
Aligned source head	0	0	0	0
Frozen + target head	4	4	1	head only
LoRA	4	8	2	197,121
Partial fine tuning	4	8	3	6,306,305
Full fine tuning	4	20	4	26.81M

For the binary tasks in this study, $\mathcal{S}_t$ is AUROC, $\delta_Q = .01$, and the paired safety margin is $\epsilon_Q = .02$ AUROC. A .01 Brier margin provides the secondary calibration check. Recipe selection uses 5,000 patient bootstrap draws. Main shift analyses use 2,000 draws.

Table 9: Clinical concepts used for task distance. Tokens are fixed before analysis. The general pretrained encoder has no outcome specific concept prior, so its set is empty. It remains eligible through distribution and adaptation evidence.

Source or target	Concept set
Pretrained source	$\varnothing$
Deterioration source	death, mortality, organ support, respiratory, circulatory, renal, ventilation, pressor, RRT
Circulatory source	circulatory, pressor, hypotension, MAP, lactate, shock
Respiratory source	respiratory, ventilation, intubation, FiO_2, PEEP, gas exchange
Acute target	death, organ support, ventilation, pressor, RRT
Circulatory target	circulatory, pressor, hypotension, MAP, lactate, shock
SOFA target	organ support, respiratory, circulatory, renal, neurologic, coagulation, liver
Delirium target	neurologic, delirium
AKI target	renal, AKI, creatinine, urine output
Mortality target	death, mortality