# Knowledge-Guided Temporal-Language-Action Model for Hierarchical Ventilator Control in Clinical Settings

Anonymous Author(s)

## Abstract

Recent advancements have highlighted the potential of training multimodal large language models (MLLMs) as general-purpose controllers for various robotics and industrial control tasks. These models utilize their reasoning abilities and multimodal understanding to handle complex tasks effectively. In ventilator control, clinicians follow a hierarchical decision-making process to choose the appropriate management strategy, such as maintenance or weaning, by assessing the patient's condition. This is followed by fine-tuning specific ventilator parameters. The process relies heavily on clinical expertise and the continuous monitoring of waveforms, which can be time-consuming. In this paper, we introduce the **temporal-language-action (TLA)** model for hierarchical optimization of ventilator parameters. Our approach combines strong temporal representations with the advanced reasoning capabilities of large language models. The TLA model first determines the appropriate management strategy by evaluating the patient's condition and then optimizes the ventilator parameters using a dedicated action head. Experimental results on clinical ventilator control datasets provided by the collaborative university hospital show that the TLA model accurately selects the mode and fine-tunes the parameters, offering a more efficient and reliable approach to ventilator control.

## CCS Concepts

• **Applied computing** → **Life and medical sciences**.

## Keywords

Large language models, temporal-language-action, ventilator decision-making.

**ACM Reference Format:**

Anonymous Author(s). 2025. Knowledge-Guided Temporal-Language-Action Model for Hierarchical Ventilator Control in Clinical Settings. In *Proceedings of CIKM* . ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction

Effective ventilator management is vitally important in respiratory therapy, directly impacting patient survival and recovery. This complex task requires the careful adjustment of multiple interconnected parameters, such as tidal volume, respiratory rate, and oxygen concentration [10]. This process involves continuous interpretation driven by clinical expertise and consumes significant time, highlighting the need for safe automation frameworks to improve efficiency. Even subtle modifications can significantly alter respiratory support, necessitating clinicians to constantly interpret a wealth of patient data-respiratory waveforms, blood gas analyses, and pulmonary mechanics measurements [23]. Waveform analysis is particularly crucial, offering intuitive insights into respiratory mechanics, enabling the identification of trends, early detection of complications like airway obstructions or pulmonary disease changes (e.g., through airway pressure waveform analysis), and guiding informed parameter adjustments [9].

Clinically, effective ventilator management intrinsically follows a hierarchical decision-making approach. Expert clinicians first meticulously assess the patient's overall condition to determine the most appropriate overarching ventilator management strategy—typically deciding between a maintenance strategy aimed at rapid stabilization or a weaning strategy designed to facilitate the patient's return to spontaneous breathing. This foundational strategic decision then critically guides all subsequent, precise adjustments to individual ventilator parameters, allowing for dynamic and tailored treatment responsive to subtle changes in the patient's status. However, many existing computational or AI-driven approaches to ventilator optimization often oversimplify this complex process. They may attempt direct, single-step parameter prediction without this crucial strategic stratification, or they might struggle to effectively integrate and interpret the full spectrum of relevant patient data, particularly the rich, continuous information embedded in respiratory waveforms. Such limitations, as conceptualized in the upper path of Figure 1, can lead to less robust, suboptimal, or even clinically misaligned recommendations. In contrast, our Temporal-Language-Action model, depicted in the lower path of Figure 1, is explicitly designed to emulate this vital clinical hierarchy and overcome these prevalent challenges. By first an TLA-driven determination of the management strategy and then guiding context-aware parameter optimization—all while deeply integrating multimodal data including effectively reprogrammed waveforms—TLA provides decision support that is not only data-driven but also fundamentally aligned with proven clinical practice, thereby enhancing the potential for safe and effective automated assistance.

The advent of large language models (LLMs) has sparked significant interest due to their advanced capabilities in text processing and generation, with great potential in healthcare [33]. Integrating these models with domain expertise, as demonstrated by OpenAGI [8], further enhances their practical applicability. Multimodal large language models , capable of processing diverse data types [17, 30], hold particular promise in medicine. For instance, visual language analysis (VLA) has shown utility in rapid epidemic response and clinical decision support through the integration of visual and textual data [13, 18]. Furthermore, machine learning

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

*CIKM* , *Seoul,Korea.*

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

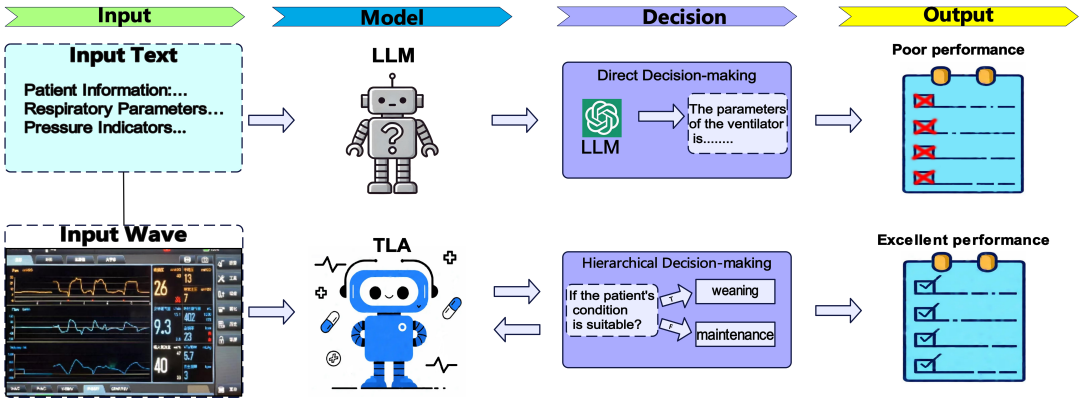


**Figure 1: TLA Model vs. Traditional Methods in Hierarchical Clinical Decision-Making**

techniques, including reinforcement learning (RL), have shown potential in optimizing ventilator settings, improving the precision of care [21, 22]. However, effectively leveraging the richness of multimodal ventilator data, especially continuous waveform signals, for truly precise and clinically-informed parameter adjustments remains a significant challenge. While methods like MedTSLLM [1] explore waveform data in clinical tasks, they often lack deep integration of clinical expertise in parameter optimization. Inspired by the paradigm shift of "Time-LLM" [16], which reprogrammed LLMs for time series forecasting, we recognize the need for a specialized architecture that can bridge the gap between the temporal complexities of respiratory waveforms and the structured clinical decision-making process. Current general-purpose LLMs struggle to effectively process continuous temporal data and inherently incorporate the hierarchical clinical strategies essential for optimal ventilator management.

To address these limitations and more effectively harness both clinical knowledge and the rich multimodal data from ventilators for refined parameter adjustments, we introduce the TLA model (Figure 2). TLA is designed to overcome the inherent limitations of standard LLMs in this domain by explicitly integrating clinical heuristic knowledge and effectively processing temporal waveform data. Inspired by the clinical workflow, the TLA model mimics the hierarchical decision-making process, first evaluating the patient's condition to choose between maintenance or weaning, and then adjusting the ventilator parameters. By effectively "reprogramming" temporal waveform data for LLM comprehension, TLA achieves more accurate and clinically relevant parameter optimization. This comprehensive approach aims to make ventilator parameter optimization not only more intelligent and data-driven but also deeply aligned with established clinical practice. Our experimental results demonstrate TLA's effectiveness through significant performance improvements on clinical datasets. Our key contributions are summarized as follows:

- We formalize ventilator parameter adjustment as a clinically-informed hierarchical decision-making process, establishing a novel and structured framework for this complex critical care task.
- We introduce the Temporal-Language-Action TLA model, a novel architecture specifically designed for clinically-aligned, hierarchical ventilator management. TLA achieves this by uniquely integrating discrete physiological data, reprogrammed continuous ventilator waveforms, and clinical heuristic knowledge within a large language model framework to guide both ventilation strategy selection and precise parameter optimization.
- We demonstrate the successful application of the TLA model in addressing key challenges within automated ventilator management, such as effective multimodal data fusion and the integration of clinical reasoning. Experimental results validate the TLA model's efficacy, showcasing its superior performance on real-world clinical datasets.

## 2 Related Work

### 2.1 Time Series Analysis with Large Language Models

The success of Large Language Models in natural language processing has spurred significant interest in adapting these powerful architectures for time series analysis. A fundamental challenge is transforming continuous time series data into discrete sequences amenable to LLM processing. Inspiration for such tokenization often stems from computer vision, where the "patching" strategy for Vision Transformers (ViTs) [6, 32] effectively tokenized continuous spatial data. This aligns with the broader concept of "model reprogramming" [3], a resource-efficient strategy for leveraging pre-trained models across different domains and modalities, often by repurposing input and output interfaces without extensive retraining of the core model.

These paradigms of input tokenization and model reprogramming are increasingly explored to adapt pre-trained LLMs for various time series tasks, predominantly forecasting. Notable examples include TimeGPT-1 [7], which employs a pre-trained LLM for zero-shot forecasting, LLM4TS [2] with its specialized two-stage fine-tuning, and prompt-based approaches like GPT4MTS [15] for multimodal time series forecasting. Among efforts to adapt LLM internals more deeply for temporal data, Time-LLM [16] stands out by "reprogramming" LLMs through redesigned input embedding and

attention mechanisms, serving as a key inspiration for our work. Building upon such concepts of deep reprogramming, our TLA model's waveform component is designed to make complex physiological waveforms LLM-interpretable for clinical decision-making. As detailed in our Methods section, TLA's clinically-tailored reprogramming uniquely employs learnable text prototypes and cross-attention mechanisms to link numerical waveform patterns with interpretable clinical semantics. This specialized approach, integral to TLA's hierarchical structure, aims to create richer, clinically-informed representations and address challenges in domain-specific interpretation, long-sequence management, and bridging the modality gap in physiological time series analysis.

## 2.2 Large Language Models in Healthcare and Medicine

The application of LLMs in the medical domain has experienced rapid growth, with models being developed and fine-tuned to understand and generate human-like text for various clinical applications. General-purpose medical LLMs, such as Med-PaLM [25], GatorTron [29], and the Llama3-Med42-8B model [5], are typically trained or fine-tuned on vast corpora of biomedical literature, electronic health records (EHRs), and clinical notes. These models have shown promise in tasks such as medical question answering, diagnostic assistance, clinical text summarization, and patient-facing applications.

Beyond these general text-based applications, the field is rapidly advancing towards specialized foundation models and sophisticated data integration techniques to tackle complex medical challenges. In oncology, for instance, efforts are focused on creating general-purpose foundation models for computational pathology [4] and leveraging automated integration of diverse real-world data to significantly improve cancer outcome prediction [14]. Similarly, the principles underlying LLMs, particularly transformer architectures, are being adapted to interpret complex physiological time-series data. In electroencephalography (EEG) analysis, models like EEGPT aim to learn universal representations of EEG signals [27], while broader initiatives such as BioSignal Transformers (BioT) are designed for robust cross-data learning from various real-world clinical biosignals [28]. These advancements underscore the trend towards using deep learning to extract nuanced insights from diverse and complex medical data modalities.

Despite this remarkable progress in developing specialized models for various medical data types, effectively translating these advanced data interpretations and representations into seamlessly integrated clinical decision support remains a significant challenge. Fully harnessing these capabilities, particularly for dynamic, continuous physiological signals within complex, hierarchical clinical workflows, continues to be an active area of research and development, paving the way for more refined and context-aware clinical tools.

# 3 Method

As shown in Figure 2, the **TLA** model consists of three components: reprogrammed waveform data processing, Ventilation Strategy Selection, and an attention-aware hidden state mapper.

## 3.1 Reprogram Wave Data

This section describes how the **TLA** model processes raw waveform data into a representation suitable for a Large Language Model (LLM). This process is essential for the **TLA** model's handling of time series data.

First, the input raw time series waveform sequence $X \in \mathbb{R}^T$, where $T$ is the number of time steps, is segmented using **overlapping window segmentation** to generate $P$ contextual "patches" $\{X_p\}_{p=1}^{P}$. Each patch has a window size $L_p$ and a stride $S$. The number of patches $P$ is calculated using the following formula, as referenced in [20]:

$$P = \left\lfloor \frac{T - L_p}{S} \right\rfloor + 2 \tag{1}$$

The motivation for this segmentation is twofold: 1) better preserving local semantic information by aggregating local information into each patch, and 2) serving as a form of "tokenization" to create a compact sequence of input tokens, thereby reducing computational burden. After segmenting the time series into patches, each patch $X_p \in \mathbb{R}^{L_p}$ is linearly projected through learnable weights $W_e \in \mathbb{R}^{L_p \times d_{\text{llm}}}$ to obtain embedded representations $X'_p = X_p W_e^\top$. This step aims to align the dimension of the temporal features with the hidden dimension of the LLM, $d_{\text{llm}}$.

To establish a latent bridge between numerical patterns and linguistic semantics, we introduce **learnable text prototypes** $E' \in \mathbb{R}^{k \times d_{\text{llm}}}$, where $k \ll |V|$ (vocabulary size). These text prototypes are learned from the pre-trained word embeddings $E \in \mathbb{R}^{V \times D'}$ within the LLM backbone (assuming $D' = d_{\text{llm}}$). They are designed to connect linguistic cues, such as "short up" and "steady down", and combine them to represent local patch information. Compared to directly editing time series or describing them losslessly in natural language, text prototypes offer an efficient and adaptive way to select relevant source information.

The reprogramming process for waveform data employs a multi-head cross-attention (MHCA) mechanism [26]. Specifically, for each attention head $k$, we define query matrices $Q = X^p W_q$, key matrices $K = E' W_k$, and value matrices $V = E' W_v$. Here, $\{W_q, W_k, W_v\} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{llm}}}$ are projection matrices. The cross-attention mechanism computes the output $O$ as follows:

$$\begin{aligned} &Q = X^p W_q, \quad K = E' W_k, \quad V = E' W_v \\ &O = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_{\text{llm}}}}\right) V \end{aligned} \tag{2}$$

This attention mechanism enables adaptive pattern matching between temporal features and linguistic prototypes. By aggregating the outputs of each attention head and performing a linear projection, we obtain the output $O^{(i)} \in \mathbb{R}^{P \times d_{\text{llm}}}$ with dimensions aligned to the LLM's hidden layer dimension.

To further ground the LLM's reasoning in domain-specific statistics, we prepend a **prompt tensor** $P_{\text{stat}} \in \mathbb{R}^{3x \times d_{\text{llm}}}$ to the sequence input to the LLM. This tensor contains learnable embeddings of three key statistics: minimum ($x_{\text{min}}$), maximum ($x_{\text{max}}$), and median ($x_{\text{median}}$), where each statistic is embedded into a vector of dimension $x$. These statistics can describe the overall trend (upward or downward) by calculating the sum of differences between consecutive time steps and determine the top five lags by computing the

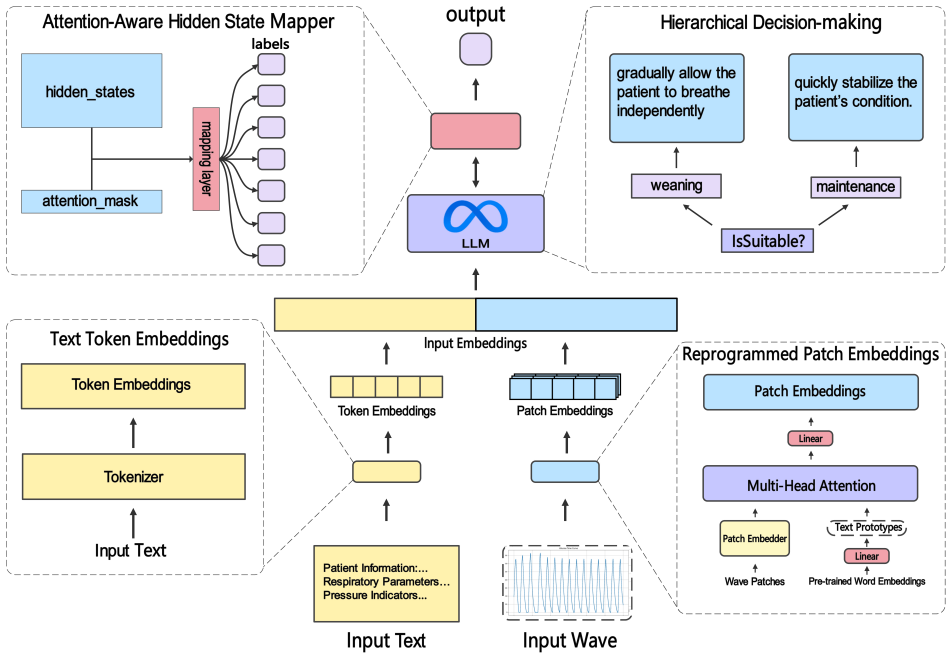


**Figure 2: The overall framework diagram of the TLA model.**

autocorrelation. The prompt tensor is constructed as:

$$P_{\text{stat}} = [\text{Embed}(x_{\min}); \text{Embed}(x_{\max}); \text{Embed}(x_{\text{median}})] \quad (3)$$

The final input sequence $\hat{O}$ to the LLM is constructed by concatenating the statistical prompt tensor $P_{\text{stat}}$ and the waveform patch representations $O$ processed by cross-attention:

$$\hat{O} = \text{Concat}(P_{\text{stat}}, O) \in \mathbb{R}^{(3x+P)\times d_{\text{llm}}} \quad (4)$$

This hierarchical encoding strategy preserves local temporal structures while enabling global semantic alignment. Through parameter-efficient cross-modal adaptation, it effectively "reprograms" raw waveform data into LLM-interpretable token sequences. This process aims to bridge the gap between continuous time series data and the LLM's processing of discrete tokens, and to activate the LLM's time series understanding and reasoning capabilities.

## 3.2 Ventilation Strategy Selection

The **TLA** model formalizes clinical decision-making through a hierarchical optimization framework that dynamically adjusts ventilator parameters based on a nuanced understanding of patient status. At the core of this hierarchy is an initial decision-making phase where the model assesses the patient's current condition using integrated multimodal input, which includes both discrete physiological parameters $\hat{D}$ and reprogrammed waveform data $\hat{O}$.

Our **TLA** model incorporates a dedicated action head that processes the combined representation derived from $\hat{D}$ and $\hat{O}$. This head determines the appropriate high-level ventilation strategy by generating a policy $A_t$, which represents the model's strategic decision regarding the patient's state and the general direction of intervention. The policy $A_t$ is a binary decision selecting between two primary strategies: Maintenance (Patient not suitable for weaning, $A_t = 0$) and Weaning (Patient suitable for weaning, $A_t = 1$).

$$A_t = \begin{cases} 0 & \text{Maintenance} \\ 1 & \text{Weaning} \end{cases} \quad (5)$$

This initial decision ($A_t$) is a critical step, acting as a high-level control signal within the hierarchical framework. It dictates the subsequent phase of parameter optimization by setting the overall goal and context.

Depending on the selected strategy $A_t$, the input for the detailed parameter adjustment process is augmented with a specific, guiding prompt tensor $\hat{P}$. These prompts are designed to steer the LLM's reasoning towards generating parameters consistent with the chosen clinical strategy, leveraging insights from multimodal learning [24, 31]. Specifically, if $A_t = 0$ (Maintenance), indicating the patient's current condition requires stabilization, the input is augmented with the prompt $\hat{P}_0$: *"The patient's current condition is not suitable. I need to use larger parameters to quickly stabilize the patient's condition."*. Conversely, if $A_t = 1$ (Weaning), indicating the patient's condition is suitable for reducing support, the input is augmented with the prompt $\hat{P}_1$: *"The patient's current condition is suitable. I need to set smaller parameters to gradually allow the patient to breathe independently."*. These prompts encourage the model to focus on settings aimed at providing robust support ($\hat{P}_0$) or facilitating gradual liberation from mechanical ventilation ($\hat{P}_1$).

These enhanced inputs, comprising the original discrete data $\hat{D}$, the reprogrammed waveform data $\hat{O}$, and the strategy-specific prompt $\hat{P}_{A_t}$ (where $A_t$ determines the specific prompt), are then processed by the core **TLA** model. The model generates an output

in the form of hidden states:

$$\text{hidden\_states} = \text{TLA}(\hat{D}, \hat{O}, \hat{P}_{A_t}) \tag{6}$$

These hidden states encapsulate the model's comprehensive understanding of the patient's state in the context of the chosen strategy and serve as the foundation for the subsequent parameter optimization performed by the Attention-Aware Hidden State Mapper. This two-step hierarchical process, starting with a strategic decision based on multimodal input and then using that decision to inform detailed parameter generation via prompting, allows the **TLA** model to provide clinically informed and contextually relevant ventilator adjustments.

## 3.3 Attention-Aware Hidden State Mapper

The Attention-Aware Hidden State Mapper is a critical component responsible for transforming the aggregated hidden state representation from the language model into structured outputs suitable for downstream decision-making.

The input to the mapper is the pooled hidden state representation (denoted as $\mathbf{h}_{\text{pooled}} \in \mathbb{R}^{d_p}$). This state is obtained by taking the mean across the sequence dimension of the LLM's last hidden state after processing the multimodal input. This aggregated representation $\mathbf{h}_{\text{pooled}}$ captures the distilled semantic information from the input context.

The pooled hidden state $\mathbf{h}_{\text{pooled}}$ is then processed through a two-layer feed-forward network. The first linear layer transforms the state from its input dimension $d_p$ to a hidden dimension $d_h = d_p/2$, followed by a ReLU activation function. The second linear layer maps $\mathbf{h}_1$ from dimension $d_h$ to an output dimension $d_y$. The dimensionality and interpretation of this output $\mathbf{Y} \in \mathbb{R}^{d_y}$ are adapted based on the mapper's specific role within the **TLA** model's hierarchical decision-making process. For the **Ventilation Strategy Selection task (Stage 1)**, $d_y$ is set to $d_{y,strat}$, which is typically a low dimension (e.g., $d_{y,strat} = 1$ if the strategy decision is represented along a single axis). The sigmoid output $\mathbf{Y} \in [0,1]^{d_{y,strat}}$ from the mapper is then utilized to determine the discrete clinical strategy. This determination involves mapping the continuous output to predefined numerical representations for each strategy class; for instance, by assigning the output to the strategy class whose predefined numerical representation (e.g., 0 for Maintenance, 1 for Weaning in a binary setup, or other equidistant points in [0,1] for multi-class scenarios) is closest to the mapper's output value. In contrast, for the **Ventilator Parameter Optimization task (Stage 2)**, $d_y$ is set to $d_{y,param}$, directly corresponding to the number of ventilator parameters being optimized. A sigmoid activation function $\sigma$ is applied to the output of this second linear layer. This function constrains the output values $\mathbf{Y}$ to the range $[0, 1]$. This characteristic is crucial, rendering the output suitable for the proximity-based interpretation for strategy selection in Stage 1, and for representing normalized parameter values in Stage 2. The transformations are as follows:

$$\mathbf{h}_1 = \text{ReLU}(\mathbf{W}_1\mathbf{h}_{\text{pooled}} + \mathbf{b}_1) \tag{7}$$

$$\mathbf{Y} = \sigma(\mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2) \tag{8}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_h \times d_p}$ and $\mathbf{b}_1 \in \mathbb{R}^{d_h}$ are the weights and bias of the first linear layer. The weights $\mathbf{W}_2 \in \mathbb{R}^{d_y \times d_h}$ and bias $\mathbf{b}_2 \in \mathbb{R}^{d_y}$ of the second linear layer are dimensioned according to the task-specific $d_y$. The final output $\mathbf{Y}$ of the mapper thus serves as a flexible interface for various downstream decisions within the **TLA** framework.

## 3.4 TLA Model Architecture

The Temporal-Language-Action model architecture is designed to facilitate a hierarchical decision-making process for optimizing ventilator parameters by effectively integrating diverse clinical data streams. The foundational component is a Large Language Model , specifically Llama 3.2 [11], which serves as the backbone. Its input system is engineered for **dual-modal collaborative encoding**: discrete physiological parameters $\hat{D}$ (e.g., gender, age, Tidal Volume, Respiratory Rate, Peak Pressure, Mean Airway Pressure) are embedded using the Llama 3.2 native tokenizer, while ventilator waveforms are "reprogrammed" into LLM-interpretable temporal features $\hat{O}$, as described in Section 2.1. The LLM's inherent self-attention mechanisms are leveraged to process these combined variable-length inputs, effectively capturing temporal dynamics and integrating raw data with learned semantic information to enhance clinical interpretability.

The problem is framed as a hierarchical decision-making process during inference (detailed in Algorithm 1), which is mirrored by a two-stage training methodology that optimizes distinct aspects of the model:

**Algorithm 1** TLA Model Inference Process

**Input:** Raw discrete clinical data $D_{raw}$, Raw ventilator waveforms $W_{raw}$
**Output:** Optimized ventilator parameters $V_{final}$
// — Input Preprocessing —
$\hat{D} \leftarrow$ TokenizeAndEmbedDiscreteData($D_{raw}$) ▷ Embed discrete data using LLM native tokenizer
$\hat{O} \leftarrow$ ReprogramWaveformData($W_{raw}$) ▷ Process waveforms into LLM-interpretable $\hat{O}$
// — Stage 1: Ventilation Strategy Selection —
$X_{S1} \leftarrow$ Concatenate($\hat{D}, \hat{O}$)
$H_{S1} \leftarrow$ TLA_CoreLLM($X_{S1}$) ▷ Process combined input through LLM backbone
$\mathbf{h}_{pooled,S1} \leftarrow$ PoolHiddenStates($H_{S1}$)
$\mathbf{Y}_{S1} \leftarrow$ AHSM($\mathbf{h}_{pooled,S1}$, config='strategy') ▷ AHSM for strategy prediction
$A_t \leftarrow$ DetermineDiscreteStrategy($\mathbf{Y}_{S1}$) ▷ Map $\mathbf{Y}_{S1}$ to discrete strategy $A_t \in \{0, 1\}$
// — Stage 2: Ventilator Parameter Adjustment —
$\hat{P}_{A_t} \leftarrow$ GenerateAndEmbedStrategyPrompt($A_t$) ▷ Create embedded prompt $\hat{P}_{A_t}$ from $A_t$
$X_{S2} \leftarrow$ Concatenate($\hat{D}, \hat{O}, \hat{P}_{A_t}$) ▷ Augment input with strategy prompt
$H_{S2} \leftarrow$ TLA_CoreLLM($X_{S2}$) ▷ Re-process augmented input through LLM
$\mathbf{h}_{pooled,S2} \leftarrow$ PoolHiddenStates($H_{S2}$)
$V_{final} \leftarrow$ AHSM($\mathbf{h}_{pooled,S2}$, config='parameters') ▷ AHSM for parameter prediction
**return** $V_{final}$

**Stage 1: Ventilation Strategy Selection Training.** The first training stage focuses on accurately predicting the overarching ventilation strategy ($A_t \in \{\text{Maintenance, Weaning}\}$), using the combined reprogrammed waveform data $\hat{O}$ and discrete clinical data $\hat{D}$ to achieve robust cross-modal alignment. To efficiently preserve the LLM's extensive pre-trained knowledge and focus learning on task-specific components for this initial classification, the LLM backbone is kept **frozen**. This freezing strategy allows for dedicated optimization of the waveform reprogramming pathways and the Attention-Aware Hidden State Mapper (AHSM), enabling them to effectively bridge the modality gap and align features for the strategic decision task by leveraging the LLM's powerful fixed representations. Consequently, the AHSM (configured as described in Section 2.3) is trained to classify the patient's status and thereby determine $A_t$ from these aligned, LLM-extracted features.

**Stage 2: Ventilator Parameter Adjustment Training.** The determination of $A_t$ relies on ground-truth strategies during this training phase (whereas Stage 1's output is used during inference). Subsequently, the second training stage addresses fine-grained ventilator parameter adjustment. The input to the TLA model is augmented with a strategy-specific prompt $\hat{P}_{A_t}$, forming an enhanced input comprising $\hat{D}$, $\hat{O}$, and $\hat{P}_{A_t}$. In this stage, the LLM backbone undergoes **low-rank adaptation (LoRA)** [12] fine-tuning. LoRA is employed for its parameter efficiency, enabling robust adaptation of the LLM to this specialized downstream task with reduced computational demands while substantially mitigating the risk of catastrophic forgetting of its valuable pre-trained knowledge. Such targeted fine-tuning is crucial for imbuing the LLM with the nuanced reasoning capabilities required for precise, strategy-aligned parameter adjustments based on the integrated multimodal input. The AHSM (configured as per Section 2.3 for parameter output) is then trained to map the LoRA-fine-tuned LLM's context-aware hidden states to precise, normalized ventilator parameter values.

In essence, TLA's synergistic integration of textual prompts, physiological parameters, and reprogrammed waveform data within a two-stage hierarchical framework (with the inference process detailed in Algorithm 1) significantly enhances predictive accuracy, clinical adaptability, and interpretability. This structured approach facilitates a clinically coherent progression from high-level strategic decisions to fine-grained parameter adjustments in complex scenarios.

# 4 Experiments and Results

## 4.1 Dataset

The clinical dataset used in this study was collected and curated by multiple physicians guided by a senior attending physician. This dataset comprises detailed clinical records from over 1000 patients, encompassing a wide range of physiological indicators and ventilator-related data throughout the course of mechanical ventilation. The clinical labels for 'Ventilation Strategy Selection' (defined as the `IsSuitable` variable) were reviewed and annotated by the senior attending physician, ensuring high clinical relevance and data quality.

This multimodal dataset integrates diverse patient information relevant to ventilator parameter optimization decisions. Specifically, it includes fundamental demographic characteristics such as gender and age, along with key physiological indicators and ventilator parameters reflecting the patient's instantaneous state—for example, inhaled tidal volume (VTI), actual respiratory rate (RATE), peak airway pressure (PPEAK), and mean airway pressure (PMEAN). Furthermore, the dataset contains high-temporal-resolution respiratory mechanics waveforms, continuously recorded at a sampling frequency of 200 Hz. These high-fidelity waveforms primarily consist of airway pressure, flow, and volume waveforms, providing a foundation for the proposed model to perform an in-depth analysis of subtle respiratory dynamics.

Leveraging this dataset, the core prediction tasks are twofold. The first is 'Ventilation Strategy Selection' (`IsSuitable`), which aims to determine whether the patient's current condition is appropriate for initiating the weaning process or requires maintenance of the existing ventilatory support; this task is framed as a binary classification problem. The second aspect involves recommending a series of optimized ventilator parameter settings. These parameters specifically include `SET_SIMVRR` (Synchronized Intermittent Mandatory Ventilation Respiratory Rate), `SET_VENTMODE` (Ventilation Mode), `SET_TRIGGERFLOW` (Flow Trigger Sensitivity), `SET_OXYGEN` (Oxygen Concentration), `SET_PEEP` (Positive End-Expiratory Pressure), and `SET_PSUPP` (Pressure Support Level).

By learning from this comprehensive multimodal clinical dataset, the **TLA** model proposed in this paper aims to achieve precise, dynamic, and personalized optimization of ventilator parameters and to provide reliable decision support for clinicians in selecting ventilation strategies.

## 4.2 Experimental Setup

**Table 1: Key hyperparameters for TLA model configuration and training.**

| Parameter | Value |
|---|---|
| *Data Processing* | |
| Input Sequence Length (`seq_len`) | 1000 |
| Waveform Patch Length (`patch_len`) | 200 |
| Waveform Patch Stride (`stride`) | 50 |
| *Reprogram Wave Data (MHCA) Specifics* | |
| MHCA: Input Dimension (`enc_in`) | 7 |
| MHCA: Internal Model Dimension (`d_model`) | 16 |
| MHCA: Attention Heads (`n_heads`) | 8 |
| MHCA: Encoder Layers (`e_layers`) | 2 |
| MHCA: Feed-Forward Dimension (`d_ff`) | 32 |
| *Core Training Parameter* | |
| Learning Rate (`learning_rate`) | $1 \times 10^{-4}$ |
| Dropout Rate | 0.1 |

All experiments were performed on a single NVIDIA A100 GPU with 80GB of memory, utilizing a PyTorch-based framework. The dataset, as detailed in Section 4.1, was partitioned into training and validation sets, with 10% of the data allocated for validation purposes, employing patient-level splits to prevent data leakage where appropriate.

The **TLA** model was trained following the two-stage methodology described in our Method section: Stage 1 (Ventilation Strategy Selection with a frozen LLM) and Stage 2 (Ventilator Parameter Adjustment via LLM fine-tuning). The AdamW optimizer [19] was employed for parameter updates. Key hyperparameters governing data processing, general LLM configuration, specific settings for the Reprogram Wave Data component's multi-head cross-attention , and the learning rate are enumerated in Table 1. Other aspects of the training process, such as the number of epochs for each stage, batch size, loss function, learning rate schedule, and early stopping criteria, were set to standard values suitable for the respective tasks and model architecture, and can be found in our supplementary materials or codebase.

For Stage 2 LLM fine-tuning, Low-Rank Adaptation (LoRA) was employed to efficiently adapt the model. The LoRA configuration featured a rank (`r`) of 8, an alpha scaling factor (`lora_alpha`) of 32, and a LoRA-specific dropout rate (`lora_dropout`) of 0.1. LoRA was applied to key projection layers within the LLM's attention and feed-forward network modules (e.g., `q_proj`, `k_proj`, `v_proj`, `o_proj`, and MLP-related projections).

## 4.3 Baselines

**Table 2: Baseline Model Performance on Key Predictive Metrics.**

| Model | Loss↓ | MAE↓ | $R^{2}$↑ |
|---|---|---|---|
| **TLA-3b(ours)** | **0.0843** | **0.2300** | **0.2541** |
| **TLA-1b(ours)** | 0.0847 | **0.2300** | 0.2300 |
| Llama-3.2-3B-full | 0.1488 | 0.3369 | -0.1078 |
| Llama-3.2-1B-full | 0.2086 | 0.4033 | -0.4966 |
| Llama-3.2-3B | 0.1503 | 0.3391 | -0.1187 |
| Llama-3.2-1B | 0.2061 | 0.4011 | -0.4782 |
| Llama3-Med42-8B | 0.1507 | 0.3431 | -0.1243 |

We conducted an extensive comparison with several categories of baseline models—including general-purpose large language models (Llama-3.2 variants), their "-full" counterparts, and a specialized medical LLM (Llama3-Med42-8B) to evaluate the clinical efficacy of our **TLA** model. Key performance metrics, including Loss, Mean Absolute Error (MAE), and $R^2$, are presented in Table 2. A focused comparison of Explained Variance is provided separately in Figure 3. In Table 2 (and other result tables where applicable), bold text indicates the optimal solution, while underlined text denotes the suboptimal solution. The "-full" suffix, applied to some baseline models, signifies that their input embeddings were filled with uniformly distributed random values to match the input length of the **TLA** model, simulating naive input padding.

The evaluation detailed in Table 2 demonstrates that conventional language models, when directly applied to this complex clinical prediction task, exhibit substantial performance degradation. Their Mean Absolute Errors (MAE) are considerably higher, exceeding that of our **TLA**-3b model by 32.1-43.0%. Furthermore, their consistently negative $R^2$ scores (e.g., Llama-3.2-3B achieving -0.1187) indicate a significant difficulty in effectively capturing the underlying physiological dynamics, implying their predictions are less reliable than simply using the mean of the target values. The performance degradation was also evident in the "-full" variants; the sharp 114.8% MAE increase observed in 1B models under these conditions underscores the importance of our clinical waveform reprogramming paradigm over simplistic or generic input handling strategies.

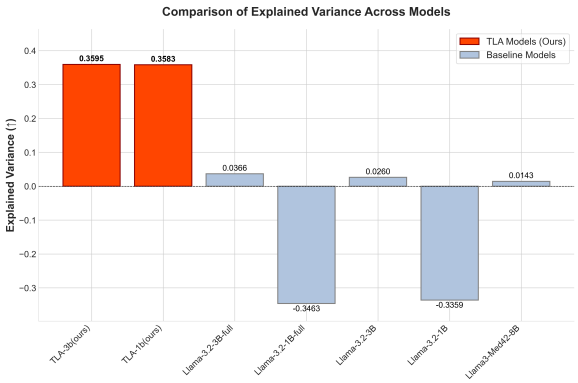


**Figure 3: Comparison of Explained Variance Across Models**

In contrast, the **TLA** framework demonstrates superior predictive accuracy and noteworthy parameter efficiency. Figure 3 specifically highlights the **TLA** models' ability to account for data variability (Explained Variance). For instance, the **TLA**-1B variant achieves an Explained Variance over 25-fold higher than that of the much larger 8-billion parameter Llama3-Med42-8B medical specialist model, despite **TLA**-1B utilizing approximately 89.6% fewer parameters. Building upon this efficiency, our **TLA**-3B model establishes state-of-the-art performance (Table 2). Crucially, our **TLA** model's safety-centric design, evidenced by its consistently positive $R^2$ scores (from Table 2, e.g., 0.2541 for **TLA**-3B) and Explained Variance scores (from Figure 3), ensures reliable clinical interpretability and mitigates the risk of negative or nonsensical predictions that were observed in all evaluated baseline models, a paramount consideration for trustworthy deployment in clinical settings.

## 4.4 Waveform Reprogramming Module

A key component of our **TLA** architecture is the dedicated waveform reprogramming module, designed to translate complex, high-resolution physiological waveforms into semantically rich features interpretable by the downstream language model. To evaluate the efficacy of this component, we conducted a series of focused experiments. In this evaluation phase, the primary parameters of the larger **TLA** model were kept frozen, allowing us to isolate the module's contribution to performance on clinically relevant upstream decision tasks. We hypothesized that this specialized reprogramming would yield significantly better results compared to models relying on more direct or less processed inputs.

Two distinct experimental schemes were designed to assess the module's performance under different clinical decision-making scenarios, with results presented in Table 3 and Table 4. In these experiments, models incorporating the waveform reprogramming

module are denoted with a "-wave" suffix (e.g., 3b-wave, 1b-wave), and are compared against baseline counterparts (3b, 1b) that utilize the same frozen LLM backbones but lack this specialized waveform processing.

**Table 3: Maintenance/Weaning Strategy Performance**

| Model | Loss↓ | Acc↑ | F1↑ |
|---|---|---|---|
| 3b-wave | 0.6713 | **69.44** | **0.4098** |
| 1b-wave | **0.6696** | **69.44** | **0.4098** |
| 3b | 0.6711 | 67.13 | 0.4041 |
| 1b | 0.7459 | 32.20 | 0.2436 |

**Table 4: Direct Mode Selection Performance**

| Model | Loss↓ | Acc↑ | F1↑ |
|---|---|---|---|
| 3b-wave | **1.0522** | **55.25** | **0.2372** |
| 1b-wave | 1.0525 | **55.25** | **0.2372** |
| 3b | 1.0973 | 43.21 | 0.2011 |
| 1b | 1.0918 | 43.21 | 0.2011 |

Maintenance/Weaning Strategy Selection (Table 3): This task evaluates the model's ability to make a key clinical decision: whether a patient's condition warrants continuation of current ventilatory support or if they are suitable for gradual liberation towards spontaneous breathing. Such decisions are pivotal for patient outcomes and resource management. As shown in Table 3, the models equipped with waveform reprogramming markedly outperformed their counterparts. For instance, the 1b-wave model achieved a loss of 0.6696 and an F1-score of 0.4098, substantially exceeding the standard 1b model (Loss: 0.7459, F1-score: 0.2436). Similarly, the 3b-wave model (F1-score: 0.4098) demonstrated improved performance compared to the 3b model (F1-score: 0.4041). Both the 1b-wave and 3b-wave models achieved an identical accuracy of 69.44%. These results highlight the enhanced discriminative power provided by the reprogrammed waveform features for this nuanced clinical assessment.

Direct Ventilator Mode Selection (Table 4): This scheme assesses the quality of learned waveform representations on a more direct classification task—selecting an appropriate ventilator mode. While perhaps less complex than the dynamic Maintenance/Weaning strategy, accurate mode selection is fundamental to effective ventilation. The results in Table 4 further corroborate the benefits of our reprogramming approach. The 3b-wave model yielded the best performance with a loss of 1.0522 and the highest F1-score of 0.2372. Both the 3b-wave and 1b-wave models (Accuracy: 55.25%) surpassed the 3b and 1b models (Accuracy: 43.21%, F1-scores of approximately 0.2011). This suggests that even for more straightforward classification tasks, the features extracted by the waveform reprogramming module offer a distinct advantage.

## 4.5 Ablation Study

In summary, across both evaluation scenarios, the "-wave" variants consistently demonstrated superior performance in terms of loss, accuracy, and F1-score. The substantial improvements observed when the waveform reprogramming module was incorporated validate our hypothesis and underscore the module's effectiveness in extracting clinically salient information from raw physiological signals. This capability is crucial for enhancing the **TLA** model's overall ability to make informed, dynamic adjustments in complex ventilation management scenarios, ultimately paving the way for more robust and reliable automated clinical decision support.

To dissect the contributions of individual components and key design choices within the **TLA** framework, we conducted a comprehensive series of ablation studies. These experiments systematically evaluate the impact of clinical knowledge integration, the hierarchical decision-making structure, the utilization of waveform data, and model scalability under varied data conditions. The results of these studies are presented in Table 5. An asterisk (*) in the table indicates metric values that were observed to be significantly influenced by the reduced data scale in specific experiments.

First, we investigated model scalability and the influence of data volume. The **TLA**-3B model consistently outperformed its 1B counterpart across most metrics; for example, **TLA**-3B exhibited a 10.5% higher validation $R^2$ score (0.2541) compared to **TLA**-1B (0.2300). To assess performance under data constraints, we trained variants on a smaller dataset of 200 samples. These models, denoted with an '-s' suffix (e.g., **TLA**-3b-s and **TLA**-1b-s), demonstrated the architecture's ability to learn effectively from limited data. For instance, while **TLA**-3b-s achieved a validation $R^2$ of 0.2099 (approximately 82.6% of the full **TLA**-3B's $R^2$), its Mean Absolute Error (MAE) was notably lower (further detailed in Table 5), an outcome attributed to data-scale dependencies as indicated by an asterisk in Table 5 for relevant metrics. This suggests that while larger datasets are beneficial for overall model generalization, **TLA** can achieve respectable performance even with limited data.

Next, we examined the impact of clinical knowledge integration. Ablation models were created by removing the **TLA** model's direct incorporation of heuristic information, resulting in variants designated with an '-Msg' suffix (e.g., 3b-Msg and 1b-Msg). These models, lacking this explicit guidance, generally exhibited inferior validation $R^2$ and Explained Variance compared to the full **TLA** models. Specifically, the 3b-Msg variant showed a 1.39% increase in validation MAE (0.2332) compared to the **TLA**-3B's MAE of 0.2300, highlighting the benefits of **TLA**'s structured approach to integrating clinical knowledge.

The efficacy of **TLA**'s hierarchical decision architecture was then evaluated against alternative, non-hierarchical approaches. One set of variants, identified by the '-mode' suffix (e.g., 3b-mode), implemented a single-phase process where the ventilator mode was selected first, followed by parameter optimization, contrasting with **TLA**'s strategy-first methodology. The full **TLA** model significantly outperformed these '-mode' variants; for instance, the 3b-mode variant exhibited a 19.6% lower validation $R^2$ (0.2042) compared to **TLA**-3B (0.2541). This underscores the effectiveness of **TLA**'s strategy of first determining the overall ventilation strategy (Maintenance or Weaning) before optimizing specific parameters. Furthermore, other variants employing LoRA for direct parameter adjustments, thereby bypassing the full **TLA** hierarchy (e.g., 3b-lora and 1b-lora), also demonstrated substantial parameter adjustment errors and were outperformed by both the '-mode' variants and

**Table 5: Comprehensive Ablation Study Results for TLA Model Components and Variants.**

| | Train | | | | Val | | | |
|---|---|---|---|---|---|---|---|---|
| **Model** | **Loss**$^{\downarrow}$ | **MAE**$^{\downarrow}$ | **R²**$^{\uparrow}$ | **Explained**$^{\uparrow}$ | **Loss**$^{\downarrow}$ | **MAE**$^{\downarrow}$ | **R²**$^{\uparrow}$ | **Explained**$^{\uparrow}$ |
| **TLA-3b** | **0.0823** | **0.2250** | **0.2563** | **0.3769** | **0.0843** | **0.2300** | **0.2541** | **0.3595** |
| **TLA-1b** | **0.0823** | **0.2250** | 0.2562 | 0.3767 | 0.0847 | **0.2300** | 0.2300 | 0.3583 |
| TLA-3b-s | 0.0871 | 0.2186* | 0.1725 | 0.3226 | 0.0865 | 0.2187* | 0.2099 | 0.3468 |
| TLA-1b-s | 0.0859 | 0.2193* | 0.1869 | 0.3354 | 0.0863 | 0.2198* | 0.2168 | 0.3538 |
| 3b-Msg | 0.0835 | 0.2254 | 0.2443 | 0.3702 | 0.0855 | 0.2332 | 0.2086 | 0.3575 |
| 1b-Msg | 0.0828 | 0.2251 | 0.2507 | 0.3722 | 0.0851 | 0.2311 | 0.2119 | 0.3585 |
| 3b-lora | 0.0861 | 0.2263 | 0.2223 | 0.3456 | 0.0865 | 0.2339 | 0.2047 | 0.3586 |
| 1b-lora | 0.0857 | 0.2326 | 0.2069 | 0.3633 | 0.0873 | 0.2352 | 0.2009 | 0.3550 |
| 3b-mode | 0.0853 | 0.2255 | 0.2392 | 0.3528 | 0.0871 | 0.2337 | 0.2042 | 0.3579 |
| 1b-mode | 0.0846 | 0.2259 | 0.2358 | 0.3581 | 0.0869 | 0.2346 | 0.2030 | 0.3572 |
| 3b-lora-f | 0.0857 | 0.2251 | 0.2250 | 0.3485 | 0.0876 | 0.2279 | 0.1712 | 0.3260 |
| 1b-lora-f | 0.0848 | 0.2254 | 0.2341 | 0.3566 | 0.0860 | 0.2329 | 0.2037 | 0.3562 |

the complete **TLA** models, reinforcing the value of the hierarchical structure.

Finally, the importance of processed waveform data was assessed. A variant designated 3b-lora-f was tested, where the input embeddings were padded (similar to the "-full" baseline variants) with uniformly distributed random values to match the **TLA** model's input length, effectively ablating the contribution of meaningful waveform temporal features. The removal of these features resulted in a significant performance decline. The Explained Variance for 3b-lora-f on the validation set dropped to 0.3260, a decrease of 9.3% compared to the full **TLA**-3B model. This finding confirms the critical role of the reprogrammed waveform features in **TLA**'s predictive capabilities.

Collectively, these ablation studies validate the individual and synergistic contributions of **TLA**'s key design elements: its effective integration of clinical knowledge, the robust hierarchical decision-making process, and the crucial role of reprogrammed waveform data. The results also highlight the model's scalability and its ability to perform effectively even when trained on limited data.

## 5 Discussion

Our Temporal-Language-Action model presents a hierarchical framework for clinical ventilator management decision-making, integrating multimodal data like reprogrammed physiological waveforms with large language models, which is crucial in critical care. Though showing promise, it needs improvements. Key development areas include enlarging and diversifying the clinical dataset for better robustness and generalizability, and enhancing the model's ability to process long physiological time series to capture subtle patient dynamics. Maintaining high model interpretability is also vital for clinician confidence. Future work will focus on refining the TLA framework in ventilator management. Priorities are curating larger multicenter datasets and integrating advanced sequence modeling techniques for longer physiological waveform data processing, which can improve predictive accuracy and capture complex patient trajectories. Although the TLA principles might apply to other medical device control scenarios, the immediate goal is to enhance its respiratory care application. Rigorous evaluation, including prospective studies, is essential to assess its real-world performance and ensure responsible development as a clinical support tool. AI-driven tools like TLA aim to shift critical care towards more proactive and individualized patient management.

## 6 Conclusion

In this paper, we introduced the Temporal-Language-Action model, a novel, hierarchically structured framework designed to optimize ventilator parameters by integrating continuous respiratory waveform analysis, discrete clinical data, and domain knowledge, leveraging the capabilities of large language models. The **TLA** model first determines an appropriate high-level ventilation strategy—maintenance or weaning—and subsequently refines specific ventilator settings, mirroring expert clinical reasoning. Our experimental evaluations on real-world clinical data demonstrated the **TLA** model's effectiveness in both accurate strategy selection and precise parameter optimization, successfully addressing key challenges in multimodal data fusion and temporal dependency modeling. This research represents a significant advancement towards the development of intelligent, adaptive, and patient-centric respiratory support systems. Such advancements pave the way for a new paradigm where intelligent systems work in close synergy with medical professionals, augmenting their capabilities and helping to standardize optimal care practices, thereby enhancing the quality and efficiency of critical care.

## GenAI Usage Disclosure

In the preparation of this manuscript, Generative AI (GenAI) tools were utilized. The application of these tools was strictly limited to assisting with the editing and polishing of text written by the authors, such as improving grammar, spelling, and phrasing. The intellectual contribution, all presented research, and the final content of this paper are solely the work of the authors, who take full responsibility for its integrity and originality, in accordance with ACM's Authorship Policy.

# References


[1] Nimeesha Chan, Felix Parker, William Bennett, Tianyi Wu, Mung Yao Jia, James Fackler, and Kimia Ghobadi. 2024. MedTsLLM: Leveraging LLMs for Multimodal Medical Time Series Analysis. arXiv:2408.07773 [cs.LG] https://arxiv.org/abs/2408.07773

[2] Ching Chang, Wen-Chih Peng, and Tien-Fu Chen. 2023. Llm4ts: Two-stage fine-tuning for time-series forecasting with pre-trained llms. *CoRR* (2023).

[3] Pin-Yu Chen. 2024. Model reprogramming: Resource-efficient cross-domain machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 22584–22591.

[4] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. 2024. Towards a general-purpose foundation model for computational pathology. *Nature Medicine* 30, 3 (2024), 850–862.

[5] Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical llms. *arXiv preprint arXiv:2408.06142* (2024).

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[7] Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. 2024. TimeGPT-1. arXiv:2310.03589 [cs.LG] https://arxiv.org/abs/2310.03589

[8] Yingqiang Ge, Wenyue Hua, Kai Mei, Juntao Tan, Shuyuan Xu, Zelong Li, Yongfeng Zhang, et al. 2023. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems* 36 (2023), 5539–5568.

[9] Dimitris Georgopoulos, George Prinianakis, and Eumorfia Kondili. 2006. Bedside waveforms interpretation as a tool to identify patient-ventilator asynchronies. *Intensive care medicine* 32 (2006), 34–47.

[10] Daniel Gilstrap and Neil MacIntyre. 2013. Patient–ventilator interactions. Implications for clinical management. *American journal of respiratory and critical care medicine* 188, 9 (2013), 1058–1068.

[11] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv e-prints* (2024), arXiv–2407.

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[13] Hanyao Huang, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin, Chuan Xu, Renjie Yang, Qian Zheng, et al. 2023. ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. *International Journal of Oral Science* 15, 1 (2023), 29.

[14] Justin Jee, Christopher Fong, Karl Pichotta, Thinh Ngoc Tran, Anisha Luthra, Michele Waters, Chenlian Fu, Mirella Altoe, Si-Yang Liu, Steven B Maron, et al. 2024. Automated real-world data integration improves cancer outcome prediction. *Nature* (2024), 1–9.

[15] Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. 2024. Gpt4mts: Prompt-based large language model for multimodal time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 23343–23351.

[16] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728* (2023).

[17] Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns* 5, 3 (2024).

[18] Fenglin Liu, Tingting Zhu, Xian Wu, Bang Yang, Chenyu You, Chenyang Wang, Lei Lu, Zhangdaihong Liu, Yefeng Zheng, Xu Sun, et al. 2023. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine* 6, 1 (2023), 226.

[19] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101 [cs.LG] https://arxiv.org/abs/1711.05101

[20] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. arXiv:2211.14730 [cs.LG] https://arxiv.org/abs/2211.14730

[21] Arne Peine, Ahmed Hallawa, Johannes Bickenbach, Guido Dartmann, Lejla Begic Fazlic, Anke Schmeink, Gerd Ascheid, Christoph Thiemermann, Andreas Schuppert, Ryan Kindle, et al. 2021. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. *NPJ digital medicine* 4, 1 (2021), 32.

[22] Niranjani Prasad, Li-Fang Cheng, Corey Chivers, Michael Draugelis, and Barbara E Engelhardt. 2017. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *arXiv preprint arXiv:1704.06300* (2017).

[23] Martin J Tobin, Amal Jubran, and Franco Laghi. 2001. Patient–ventilator interaction. *American Journal of Respiratory and Critical Care Medicine* 163, 5 (2001), 1059–1063.

[24] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems* 34 (2021), 200–212.

[25] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. 2024. Towards generalist biomedical AI. *Nejm Ai* 1, 3 (2024), AIoa2300138.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[27] Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. 2024. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. *Advances in Neural Information Processing Systems* 37 (2024), 39249–39280.

[28] Chaoqi Yang, M Westover, and Jimeng Sun. 2023. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems* 36 (2023), 78240–78260.

[29] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, et al. 2022. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540* (2022).

[30] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178* (2023).

[31] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review* 11, 12 (Nov. 2024). doi:10.1093/nsr/nwae403

[32] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*. 558–567.

[33] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).