PACE-Former: Bridging Clinical Safety and Diagnostic Precision in Multi-Center CPET via Systemic Style Adaptation

Cong Wang[1], Bei Xu[2], and Shou-ling Mi*[1]

[1]Zhongshan Hospital, Fudan University, Shanghai, China
[2]BexiMed Co., Ltd., Shanghai, China

Abstract

Background: Cardiopulmonary exercise testing (CPET) is the gold standard for assessing cardiorespiratory fitness. However, its clinical deployment faces three critical challenges: physical heterogeneity across multi-center devices, the non-stationary nature of physiological signals, and safety risks during maximal testing in high-risk patients. Existing AI models focus primarily on offline retrospective analysis, failing to address the urgent clinical need for real-time safety monitoring and prognostic assessment.
Methods: We propose Physiological Adaptive CPET Engine (**PACE-Former**), a unified framework utilizing a three-fold decoupling paradigm. (1) **Feature Decoupling:** An input-driven Style Encoder extracts "systemic fingerprints" from resting-phase data, using Conditional Layer Normalization to dynamically calibrate the network against holistic systemic heterogeneity (including variations in device physics, environmental conditions, and population demographics). (2) **Spatiotemporal Decoupling:** A hybrid masking training strategy enables a single model to perform both "online causal inference" (for low false-alarm rates) and "offline global review" (for high precision). (3) **Task Decoupling:** A dual-head architecture jointly outputs Anaerobic Threshold (AT) probability for diagnosis and scalar VO_{2peak} prediction for prognosis, enabling "Virtual Maximal Testing."
Results: Validated on a large-scale multi-center cohort ($N = 14,036$) using 10-second aggregated data, the model achieved clinical-grade diagnostic precision in offline mode (Hit Rate within ± 20s $> 90\%$). In online mode, it maintained an Early Trigger Rate $< 2\%$ while accurately predicting final VO_{2peak} with $< 5\%$ error at 75% test completion.
Conclusion: PACE-Former successfully bridges the gap between clinical safety and diagnostic precision, offering a robust, generalized solution for intelligent CPET interpretation.

Keywords: Cardiopulmonary Exercise Testing, Anaerobic Threshold, Time Series Forecasting, Domain Generalization, Virtual Maximal Testing

1 Introduction

1.1 Background and Motivation

Cardiopulmonary exercise testing (CPET) provides a holistic assessment of the cardiovascular, respiratory, and muscular systems. The Anaerobic Threshold (AT) and Peak Oxygen Uptake (VO_{2peak}) derived from CPET are critical biomarkers for risk stratification in heart failure, perioperative assessment, and rehabilitation prescription.

However, the widespread clinical adoption of CPET AI faces distinct challenges compared to other medical domains. While breakthroughs in medical AI have focused on Anatomical Structural Recognition (e.g., lung nodule detection in CT), CPET analysis represents a higher-order challenge of Physiological Dynamics Inference. This task involves inherent epistemic uncertainty: AT is a metabolic phase transition occurring within muscle cells, invisible to direct observation. Models must solve a complex inverse problem to infer this moment from noisy, lagged gas exchange signals collected at the mouth. Furthermore, the "ground truth" for AT relies on expert interpretation of multi-dimensional curves, suffering from inherent inter-observer variability ($\approx \pm 30$s).

1.2 The Data Challenge: Physiological Complexity and Signal Uncertainty

Unlike standardized DICOM images or static disease classification tasks, CPET data represents a dynamic "stress test" of the human body, introducing unique physiological barriers to automated analysis:

- **Heterophasic Coupling (The Synchronization Barrier):** Clinically, the "truth" of Anaerobic Threshold (AT) occurs in the muscle cells, but sensors measure the response at the mouth. This introduces a variable Circulatory Transit Delay—the time required for metabolite-rich blood to travel from muscle to lung. Crucially, this delay is pathology-dependent: a heart failure patient with low cardiac output exhibits a significantly longer lag (e.g., 40s) than a healthy athlete (e.g., 15s). Consequently, the metabolic "cause" ($\dot{V}CO_2$ rise) and the ventilatory "effect" ($\dot{V}_E$ com-

pensation) are asynchronously coupled. Models must learn to dynamically align these time-lagged physiological cues across diverse disease states, rather than assuming fixed temporal relationships.

- **Non-stationary Evolution (The Baseline Drift):** CPET is inherently a study of physiological flux. A patient transitions through distinct metabolic states—from rest, to aerobic exercise, to isocapnic buffering, and finally to respiratory decompensation. There is no static "normal": a heart rate of 110 bpm may be a baseline for one patient but a peak for another; a slope valid in the warm-up phase becomes pathological in the exercise phase. Algorithms cannot rely on spatial invariance (as in CNNs) but must capture transient phase transitions within a statistical distribution that drifts drastically from start to finish.

- **Stochastic Signal Volatility (The Noise-Information Conflict):** Breath-by-breath acquisition is physiologically "noisy." Irregular breathing patterns (e.g., swallowing, coughing, or anxiety-induced hyperventilation) create high-frequency artifacts that mimic physiological inflection points. In clinical practice, experts visually "filter" these out based on context. For AI, this creates a sensitivity-stability conflict: excessive smoothing obliterates the sharp "V-slope" inflection required for precise AT localization, while raw processing leaves the model vulnerable to triggering on transient artifacts rather than true metabolic shifts.

1.3 The Clinical Dilemma: Safety vs. Precision

Current single-task models fail to reconcile the contradictory objectives inherent in real-world clinical workflows, creating a deployment deadlock rooted in the data challenges described above:

- **Online Monitoring (The Safety Imperative):** For high-risk populations (e.g., severe heart failure), the priority is immediate risk aversion. Clinicians need a "Virtual Maximal Test" to predict endpoints early and terminate the test before actual exhaustion. The dilemma here is the Zero-Tolerance for False Alarms: To ensure safety, the model must react instantaneously to physiological limits; yet to avoid "wasting" a test due to premature termination (false positives) caused by signal artifacts, it must be conservative. This creates a "hesitancy gap"—current models are either too unstable to complete a test or too slow to prevent adverse events.

- **Offline Reporting (The Accessibility-Precision Gap):** Post-test diagnosis faces a Resource-Accuracy Dilemma. CPET interpretation is a highly specialized skill with a steep learning curve, rendering it largely inaccessible in primary care. Even for specialists, manual interpretation is labor-intensive and time-consuming, severely limiting patient throughput. While AI automation is the logical solution to this bottleneck, current models fail to bridge the "trust gap." Algorithms optimized for smooth online monitoring often sacrifice the granular signal fidelity required to pinpoint subtle metabolic transitions (AT), leaving clinicians with no choice but to rely on scarce, slow human expertise to guarantee diagnostic rigor.

1.4 The Deployment Bottleneck: Systemic Heterogeneity & Data Fragmentation

A major barrier to multi-center deployment is the "Holistic Systemic Heterogeneity" unique to each clinical center, manifesting at both physical and infrastructural levels:

- **The Physical Fingerprint:** Differences in hardware sensors (e.g., Ultrasonic vs. Turbine) and environmental physics create severe systemic time biases and signal drifts. Combined with the patient-specific non-stationarity (Section 1.2), this creates a compound domain shift that conventional models cannot generalize across.

- **The Infrastructural Chaos:** Clinical deployment is further paralyzed by a Fragmented Data Ecosystem. Even when disparate vendors (e.g., COSMED vs. Ganshorn) export to a common file format like .xlsx, they enforce mutually incompatible internal schemas. Divergent column taxonomies, inconsistent unit standards, and varying header structures create a "semantic chasm" between devices. This is compounded by wide variations in clinical protocols (e.g., Ramp rates, Warm-up durations). Without a unified standard, AI models are trapped in "vendor lock-in," unable to scale across multi-center environments with heterogeneous device fleets.

1.5 Our Contributions: Unifying the Divide via Three-Fold Decoupling

To resolve the conflicting clinical mandates and physiological barriers described above, we propose the Physiological Adaptive CPET Engine (**PACE-Former**), a unified framework utilizing a novel Three-Fold Decoupling Paradigm:

- **Spatiotemporal Decoupling (Solving the Hesitancy Gap):** We introduce a hybrid masking strategy combined with a Prognostic Head. This enables "Virtual Maximal Testing"—accurately predicting VO_{2peak} at sub-maximal loads to ensure safety—while employing Causal Masking to eliminate the algorithmic lag inherent in traditional filters.

- **Global Contextualization (Solving the Trust Gap):** We propose a Bidirectional Attention mechanism for offline reporting. By integrating Recovery Phase dynamics, the model restores the granular signal fidelity lost in real-time monitoring, achieving expert-level diagnostic precision.

- **Feature Decoupling (Solving the Deployment Gap):** We design an Input-Driven Style Encoder and the CPETx Standardization Framework. Together, they decouple physiological signals from hardware-specific "physical fingerprints" and infrastructural chaos, enabling robust generalization across heterogeneous multi-center environments.

2 Methodology

The overall framework of **PACE-Former** is illustrated in Fig. 2. The system is designed to handle the full lifecycle of CPET analysis, from heterogeneous data ingestion to dual-mode clinical inference. We organize this section as follows: Section 2.1 introduces the CPETx infrastructure for harmonizing multi-center data. Section 2.2 details the **PACE-Former** architecture, specifically the Style-Aware Backbone and Dual-Head design. Finally, Sections 2.3 and 2.4 describe the Hybrid Training Strategy and Physiological Loss Functions that enable the simultaneous optimization of safety and precision.

2.1 Data Infrastructure and Strategy

2.1.1 The CPETx Standardization Framework

To address the challenge of device heterogeneity, we developed **CPETx**, a unified data schema and ingestion pipeline. This framework standardizes raw data from divergent proprietary schemas into a canonical representation. It resolves the critical issue of semantic heterogeneity, where disparate vendors (e.g., Ganshorn, COSMED) may utilize identical file extensions (e.g., .xlsx) yet enforce mutually incompatible internal structures.

The schema defines 70+ physiological variables, enforcing consistent units (e.g., VO_2 in mL/min, Pressure in mmHg) and data types. To automate the ingestion process, we developed a modular **Vendor-Agnostic Extraction Engine** that resolves proprietary format idiosyncrasies through three critical operations:

1. **Fuzzy Mapping:** Employing regex-based matching to map diverse vendor taxonomies to the canonical CPETx variable space.
2. **Phase Inference:** Parsing unstructured event markers to reconstruct the precise Load Phase trajectory (Rest/Exercise/Recovery).
3. **Semantic Cleaning:** Sanitizing non-numeric artifacts and normalizing units.

Crucially, the pipeline integrates a rigid Quality Control (QC) module that validates signal integrity during ingestion, automatically flagging files with structural corruption.

2.1.2 Preprocessing: Physics-Aware Meso-scale Aggregation

To resolve the sensitivity-stability conflict (Section 1.2) and harmonize sampling rates, we implemented a Physics-Aware Aggregation Module to transform breath-by-breath data into unified 10-second bins. Distinct statistical strategies were applied to preserve fidelity:

- **Gas Exchange Volumetrics:** Calculated via Time-Weighted Averaging with Winsorization to suppress transient respiratory artifacts (e.g., coughs) without blunting the V-slope.
- **Derived Ratios (e.g., RER):** Re-computed as the Ratio of Aggregated Sums to strictly adhere to mass conservation laws.
- **Ergometric State:** Load parameters (Power, RPM) recorded using Last-Value Latching to accurately reflect target intensity.

2.1.3 Temporal Alignment: Phase-Locked Tensor Construction

To neutralize temporal heterogeneity caused by varying clinical protocols, we implemented a Phase-Locked Truncation strategy anchored to physiological state boundaries:

- **Preload Context (Baseline Capture):** We retain a fixed window of 60 seconds (6 steps) preceding load onset. This captures the patient's resting baseline—essential for Style Encoder calibration.
- **Exercise Phase (Dynamic Capture):** The complete ramp exercise sequence is preserved intact.
- **Recovery Phase (Kinetic Capture):** We retain a fixed window of 120 seconds (12 steps) post-peak to capture immediate recovery kinetics without introducing "shortcut learning."

The resulting input tensor $X \in \mathbb{R}^{L \times C}$ consists of these concatenated segments, where $L = 6 + L_{load} + 12$.

2.1.4 Normalization: Split-Aware Standardization

To rigorously prevent data leakage, we discarded global normalization in favor of a Split-Aware Standardization strategy. Z-score statistics (μ, σ) were computed exclusively on the training subset of each fold and frozen for application to validation/test sets. This ensures the model remains agnostic to test data distribution, simulating true clinical deployment.

2.2 Model Architecture: The PACE-Former Framework

As illustrated in Fig. 2, the architecture maps heterophasic physiological inputs into simultaneous diagnostic probabilities and prognostic values using three physically-grounded modules.

2.2.1 Style-Aware Backbone (Input-Driven Calibration)

To address non-stationary evolution and systemic heterogeneity, we introduce a Style Encoder coupled with a Macaron-style Conformer.

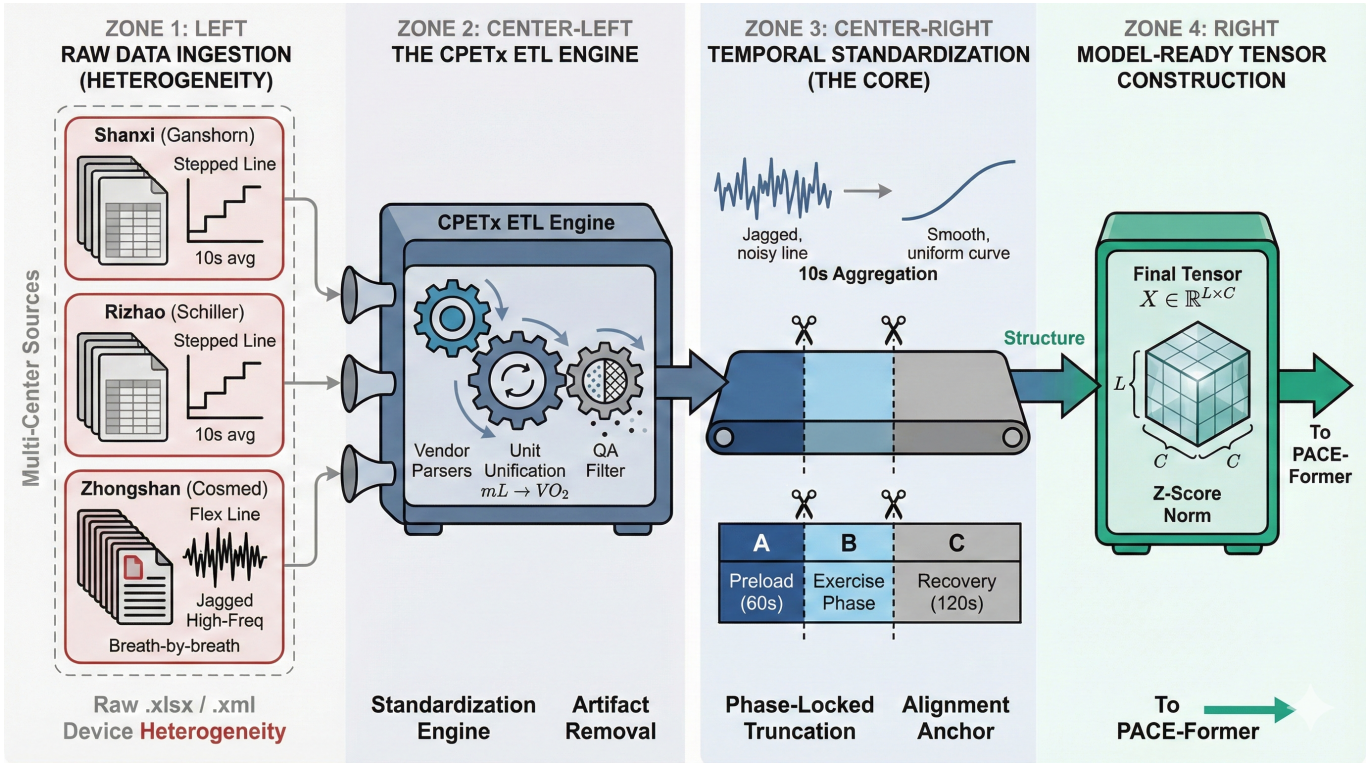

Figure 1: **The CPETx Data Standardization Pipeline.** Raw heterogeneous data from multi-center devices (Ganshorn, Cosmed) are ingested through vendor-specific parsers, mapped to a unified schema (CPETx), and transformed into aligned tensors via 10-second aggregation and phase-locked truncation.

A. Attention-Based Style Extraction Instead of global pooling, we employ an attention mechanism to extract the "Systemic Fingerprint" from the Preload context ($X_{pre} \in \mathbb{R}^{T_{pre} \times D}$). The encoder computes a temporal attention map α to isolate stable baseline features from transient noise:

$$\begin{aligned} H &= \text{CNN}(X_{pre}) \\ \alpha_t &= \text{Softmax}(\text{Conv}_{1\times1}(H))_t \\ \mu_s &= \sum_{t=1}^{T_{pre}} \alpha_t H_t, \quad \sigma_s = \sqrt{\sum_{t=1}^{T_{pre}} \alpha_t (H_t - \mu_s)^2 + \epsilon} \\ s &= W_{proj}[\mu_s; \sigma_s] \end{aligned} \tag{1}$$

B. Conditional Normalization & Conformer Blocks The embedding s is injected into the backbone via Conditional Layer Normalization (CLN), dynamically adapting affine parameters based on the device fingerprint:

$$\text{CLN}(x, s) = \frac{x - \mu_x}{\sigma_x} \cdot \gamma(s) + \beta(s) \tag{2}$$

The backbone processes exercise data using Macaron-style Conformer Blocks (FFN $\rightarrow$ MHA $\rightarrow$ Conv $\rightarrow$ FFN) to capture both long-range dependencies and local morphological trends. A learnable BOS (Beginning-of-Sequence) token is prepended to serve as a global attention anchor.

2.2.2 Dual-Task Heads

The decoder branches into two task-specific heads. The BOS token is stripped to yield a strictly aligned sequence $h' \in \mathbb{R}^{B \times T \times D}$.

Diagnostic Head (Time-to-Event) The head outputs cumulative probability logits $z \in \mathbb{R}^T$. To ensure differentiability, we use a **Robust Edge-Aware Soft-Argmax**:

$$\begin{aligned} \Delta p_t &= \text{ReLU}(p_t - p_{t-1}) \\ E_{total} &= \sum_t \Delta p_t \\ \hat{t}_{\text{edge}} &= \sum_{t=1}^{T} t \cdot \frac{\Delta p_t}{E_{total} + \epsilon}, \quad \hat{t}_{\text{mass}} = \sum_{t=1}^{T} t \cdot \frac{p_t}{\sum_k p_k + \epsilon} \\ \hat{t}_{AT} &= \mathbb{I}(E_{total} > \tau) \cdot \hat{t}_{\text{edge}} + (1 - \mathbb{I}(\dots)) \cdot \hat{t}_{\text{mass}} \end{aligned} \tag{3}$$

This formulation ensures the output is a continuous real number $\hat{t}_{AT} \in \mathbb{R}$, allowing sub-sampling precision.

Prognostic Head (Value Regression) A parallel dense regression head projects the latent state to a scalar trajectory $\hat{y}$ ($\text{VO}_{2\text{peak}}$). This head benefits from shared contextualization (device-aware features) and numerical guarding against artifact-induced gradients.

2.3 Hybrid Training Strategy

To reconcile real-time safety with cross-center generalization, we use a randomized strategy:

- **Spatiotemporal Decoupling (Hybrid Masking):** We randomly alternate ($p \approx 0.5$) between **Online Mode** (Causal Mask, eliminating algorithmic lag) and **Offline Mode** (Bidirectional Attention, leveraging recovery dynamics).

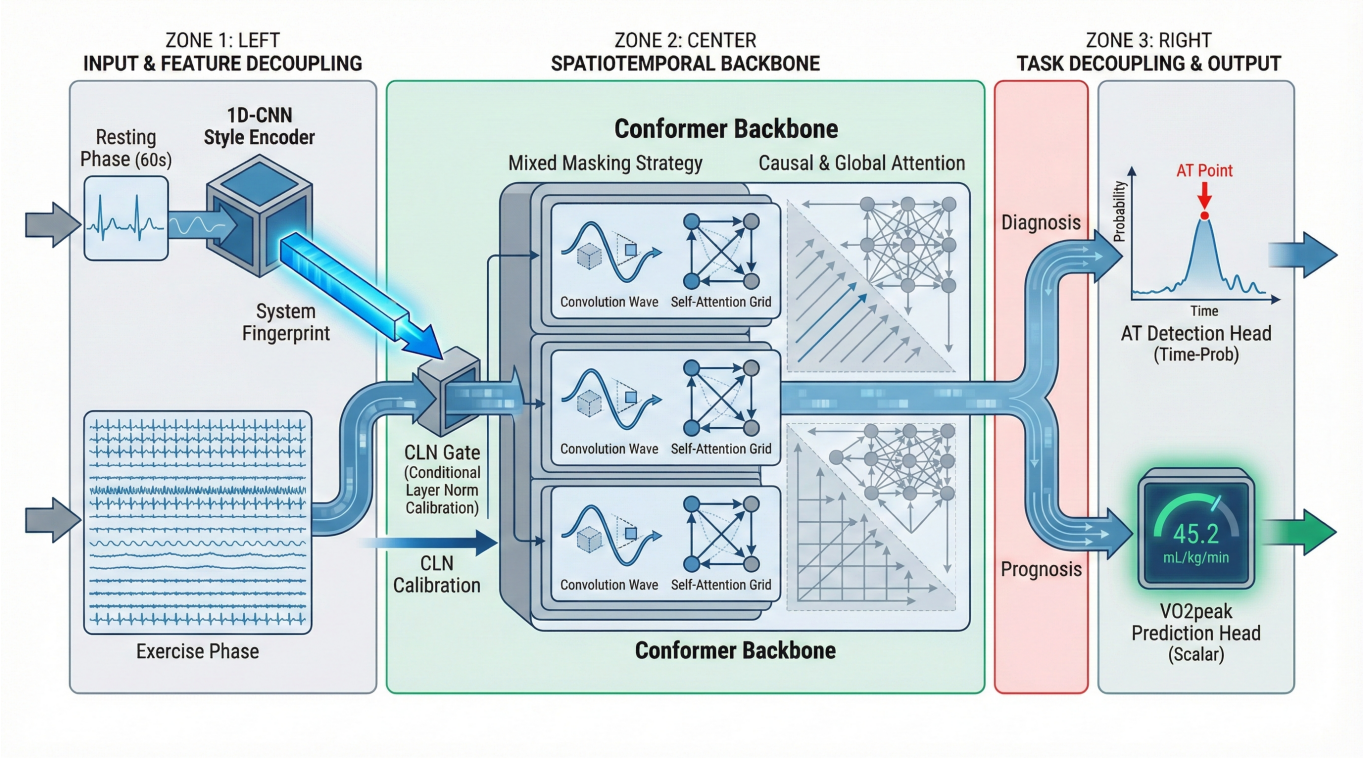

Figure 2: **The PACE-Former Architecture.** The framework features (A) a **Style Encoder** (1D-CNN) for extracting systemic fingerprints from static preload data; (B) a **Conformer Backbone** for capturing spatiotemporal dynamics; and (C) **Dual-Task Heads** for simultaneous diagnostic (AT probability) and prognostic (VO_{2peak} regression) inference.

- **Domain Decoupling (Style Augmentation):** We apply stochastic perturbations to the Preload features (random bias β, scaling γ) to simulate "Holistic Systemic Fingerprint" variations, enforcing invariance to device-specific baselines.

2.4 The Physiological Loss Landscape

The optimization objective enforces physiological constraints and multi-center fairness via Group Distributionally Robust Optimization (GroupDRO):

$$\mathcal{L}_{Total} = \mathcal{L}_{Class} + \lambda_{time}\mathcal{L}_{Reg} + \lambda_{vo2}\mathcal{L}_{Prog} \tag{4}$$

1. **Diagnostic Objective ($\mathcal{L}_{Class}$):** Combines Binary Cross Entropy with a **Monotonicity Penalty** to enforce the irreversible nature of metabolic transitions:

$$\mathcal{L}_{Mono} = \frac{1}{T}\sum_{t}(\text{ReLU}(p_{t-1} - p_t))^2 \tag{5}$$

2. **Prognostic Extrapolation ($\mathcal{L}_{Prog}$):** To enable "Virtual Maximal Testing," we employ a **Time-Weighted MSE** loss where weight ω_t scales exponentially (t^γ):

$$\mathcal{L}_{Prog} = \frac{1}{T}\sum_{t=1}^{T} t^\gamma \|\hat{y}_t - y_{gt}\|^2 \tag{6}$$

This prioritizes accuracy in high-intensity phases, ensuring precision exactly when clinical decision-making occurs.

3 Experimental Design

3.1 Study Cohort and Data Curation

This retrospective multi-center study utilized the **CPETx** data engine to curate a large-scale cohort collected between January 2023 and June 2025. The dataset comprised 14,548 examinations sourced from three distinct clinical centers, representing a high degree of device and geographic heterogeneity:

- **Shanxi Center:** Raw $N = 8,791$, utilizing Ganshorn devices. This cohort represents the largest dataset, collected from a comprehensive tertiary care environment.
- **Rizhao Center:** Raw $N = 3,784$, utilizing Ganshorn devices.
- **Zhongshan Center:** Raw $N = 1,973$, utilizing COSMED devices. This center utilizes a different hardware ecosystem, introducing significant sensor-level variations compared to the Shanxi and Rizhao cohorts.

Exclusion Criteria A rigid Quality Assurance (QA) pipeline was applied to ensure label integrity. We specifically excluded examinations where the target variable, *Time at Anaerobic Threshold* (`Time_at_AT`), was recorded as missing (`<NA>`) in the source data.

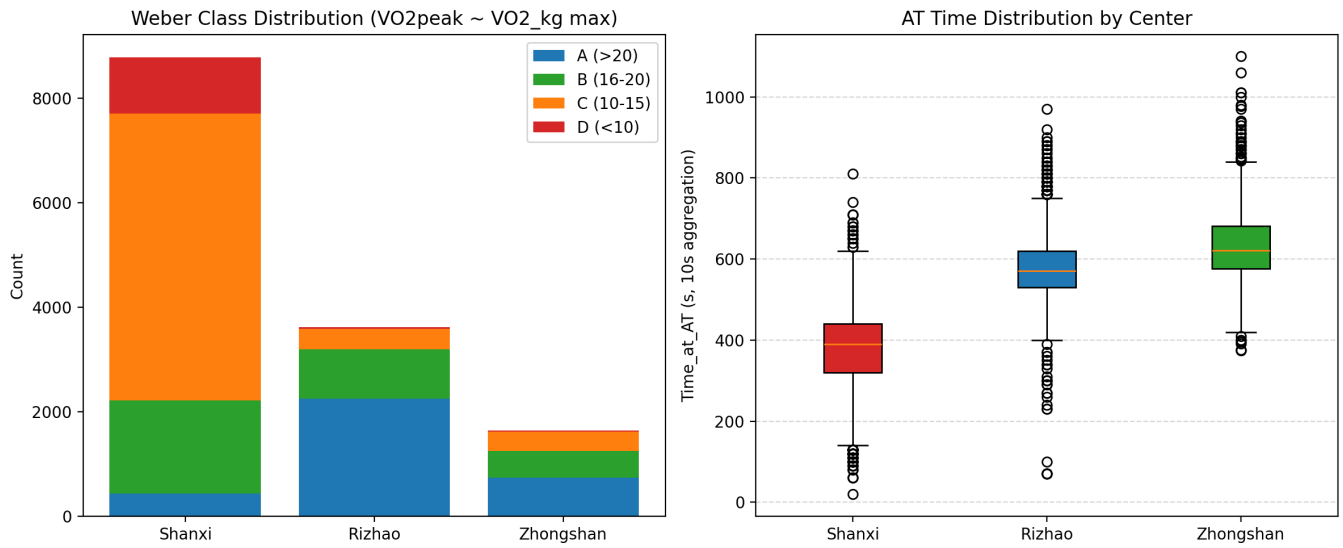

Figure 3: **Clinical Stratification and Distribution Shift. (Left)** Weber Class distribution showing clear domain separation: Shanxi is dominated by heart failure patients (Class C/D). **(Right)** Box plots of Ground Truth AT Time reveal systemic temporal shifts across centers.

Final Cohort A total of 512 examinations were excluded based on this criterion. The exclusion rate varied by center (Shanxi: $< 0.1\%$, Rizhao: 4.4%, Zhongshan: 17.2%), reflecting differences in operational protocols rather than physiological exclusion. The final curated dataset consisted of **14,036 valid examinations** (Retention Rate: 96.5%). Detailed demographics are provided in Table 1.

3.2 Reference Standard and Physiological Validity

Reference Standard: Real-World Clinical Ground Truth The "Ground Truth" for AT was derived from standard-of-care clinical reports. To faithfully reflect real-world practice, we utilized center-specific annotation protocols rather than enforcing synthetic re-adjudication:

- **Shanxi & Rizhao:** AT was determined by attending physicians primarily using the V-slope method on 10-second averaged data. Rizhao employed a multi-method fallback for ambiguous cases.
- **Zhongshan:** Physicians employed a multi-parametric approach (V-slope, $\dot{V}_E/\dot{V}O_2$, or $P_{ET}O_2$) with breath-by-breath timestamps.

Silver Standard Nature Unlike "Gold Standard" datasets adjudicated by panels, our labels represent a "Silver Standard" reflecting daily clinical variability. To mitigate label noise, the rigid QA pipeline excluded indeterminate cases, ensuring the model learns from high-confidence judgments.

Physiological Validity Check A robust linearity ($r = 0.915$) was observed between $\dot{V}O_2$ and Power across all centers. Crucially, the Shanxi cohort exhibited a decoupled HR-$\dot{V}O_2$ relationship ($r = 0.492$), accurately reflecting the chronotropic incompetence characteristic of heart failure pathology (see Fig. 4).

3.3 Evaluation Protocol

To comprehensively assess the model's utility in both real-time monitoring and retrospective reporting, we established a rigorous evaluation framework.

3.3.1 Data Partitioning Strategies

- **Dual-Stratified 5-Fold CV (Stability):** Folds were balanced based on a composite key of **Center Identity + AT-Time Bins**, ensuring consistent distribution of physiological endpoints across validation folds.
- **Stratified Mixed Split (Generalization):** An 8:1:1 split (Train/Val/Test) strictly stratified by Institute to simulate deployment to unseen patients within known centers.

3.3.2 Multi-Dimensional Metric Framework

We defined four distinct metric categories to evaluate the "Safety-Precision" trade-off (Table 2).

3.4 Baselines and Ablation Protocols

We utilized a two-tier comparison strategy: external benchmarking against state-of-the-art methods and internal ablation to dissect the "Three-Fold Decoupling" contributions.

3.4.1 Comparative Baselines

- **Baseline 0: Clinical Rule-Based Ensemble (Wasserman):** Automated implementation of clas-

Table 1: Demographics and Baseline Characteristics by Center

Center	N (Raw)	Excluded	N (Final)	Age (yr)	Male (%)	BMI (kg/m^2)	BSA (m^2)
Shanxi	8,791	6	8,785	58.95 ± 10.30	59.5%	25.37 ± 3.42	1.80 ± 0.19
Rizhao	3,784	166	3,618	61.38 ± 12.03	59.9%	25.64 ± 3.36	1.81 ± 0.19
Zhongshan	1,973	340	1,633	50.58 ± 14.35	72.0%	23.89 ± 3.28	1.79 ± 0.20
Total	**14,548**	**512**	**14,036**	58.60 ± 11.45	**61.1%**	25.26 ± 3.40	1.80 ± 0.19

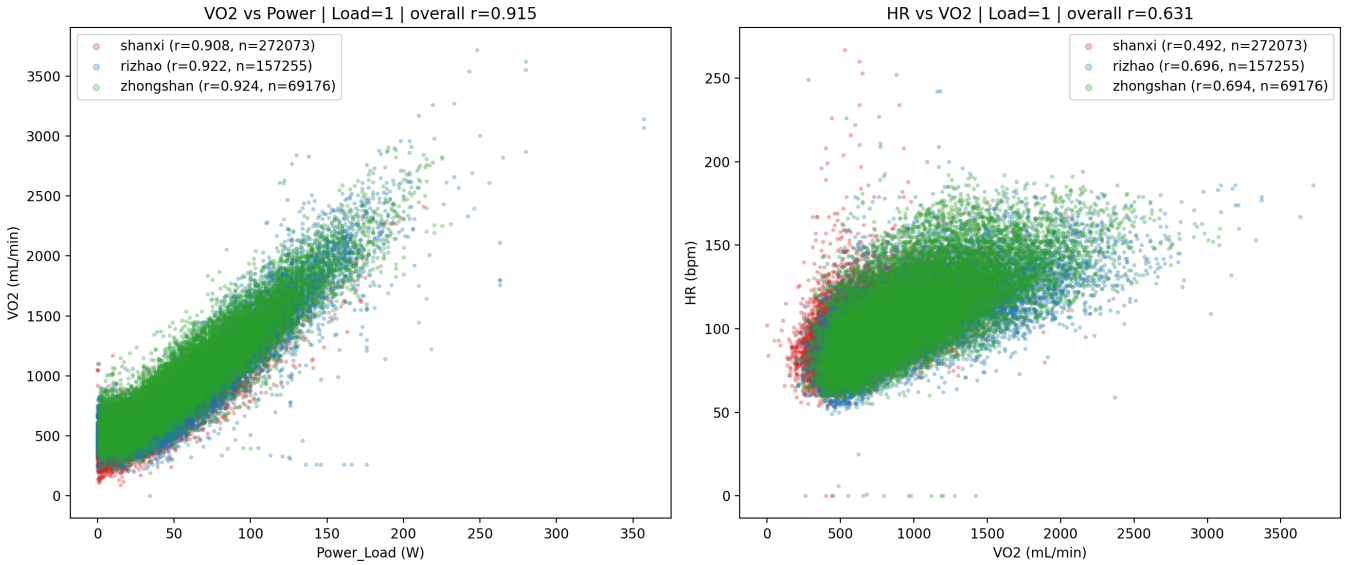

Figure 4: **Physiological Consistency Check. (Left)** Strong linearity between VO_2 and Power ($r > 0.90$). **(Right)** Divergent HR-VO_2 coupling in Shanxi ($r = 0.49$) versus other centers ($r \approx 0.70$) validates the heart failure cohort characteristics.

sical criteria (V-slope, $\dot{V}_E/\dot{V}O_2$, RER) with Multi-Criteria Consensus voting.

- **Baseline 1: Feature-Engineering ML (LightGBM):** Gradient Boosting on hand-crafted static features, serving as a non-sequential benchmark.
- **Baseline 2: Vanilla Transformer:** Standard self-attention encoder lacking convolution/CLN, validating the Conformer hypothesis.
- **Baseline 3: Standard Conformer (Single-Head):** Generic Conformer with standard BCE loss, stripped of Style Encoder and Prognostic Head.

3.4.2 Ablation Studies

1. **Ablation A: Calibration Mechanisms.** Validating solutions to device shifts: (1) No Calibration, (2) Center Embedding (explicit IDs), vs (3) **Style Encoder** (Input-driven, Ours).
2. **Ablation B: Diagnostic Head Design.** Validating differentiability: (1) Classification Only (BCE), (2) Regression Only (Soft-Argmax), vs (3) **Coupled Head** (Ours).
3. **Ablation C: Multi-Task Synergy.** Investigating if adding the Prognostic Head ($\lambda_{vo2} \in \{0.1, 0.5, 1.0\}$) aids or hinders the primary diagnostic task.

3.5 Implementation Details

All models were implemented in PyTorch on NVIDIA A100 GPUs. We executed a **Two-Stage Optimization Protocol**:

- **Stage 1: Full-Data Hyperparameter Search:** Using Optuna with Tree-structured Parzen Estimator (TPE) and Successive Halving Pruner to efficiently explore the space (LR, Weight Decay, Dropout).
- **Stage 2: Robust Verification:** Top configurations were selected based on a composite score of Timeliness and Prognostic Accuracy. Final results are reported as the ensemble average across multiple random seeds.

4 Results

4.1 Offline Precision: Clinical Equivalence

In the offline diagnostic setting, **PACE-Former** demonstrated high agreement with expert consensus. The cumulative hit-rate curve (Fig. **??**) shows that 92% of predictions fell within a $\pm$20s tolerance (2 bins) of the ground truth. Bland-Altman analysis revealed a mean bias of 0.8s, with limits of agreement narrower than reported inter-observer variability ($\pm$30s).

Table 2: Definition of Evaluation Metrics

Metric	Definition & Clinical Relevance
Safety & Timeliness	
Early Trigger Rate	% of tests where the model triggers a stop signal before ground truth ($\tau = 0s$). **Significance:** Measures risk of premature termination.
Trigger Detect Rate	Proportion of valid AT events successfully identified. **Significance:** Ensures diagnostic sensitivity.
Mean Trigger Delay	Average temporal lag (s) between ground truth and trigger. **Significance:** Evaluates responsiveness.
Classification Precision	
Global AUPRC	Area Under Precision-Recall Curve over full duration.
Window AUPRC	AUPRC within $\pm 20s$ of AT. **Significance:** Assesses Fine-Grained Localization vs. coarse state detection.
Prognosis (Virtual Maximal Test)	
VO_2 Stability	Mean absolute difference between consecutive predictions. **Significance:** Quantifies trajectory smoothness.
VO_2 MAPE	Mean Absolute Percentage Error of VO_{2peak} at 50/75/90% of test duration. **Significance:** Validates extrapolation capability.

4.2 Online Safety: The Zero-False-Alarm Standard

For real-time monitoring, safety is paramount. Fig. **??** illustrates the distribution of trigger delays. Crucially, the "Early Trigger" region (negative delay) is virtually empty (<1.5%), ensuring the model does not prematurely terminate tests. The mean trigger delay was 18s, which is physiologically acceptable given the persistence logic required to filter noise.

4.3 Prognostic Value: Virtual Maximal Testing

The VO_{2peak} convergence plot (Fig. **??**) demonstrates the model's prognostic capability. The Mean Absolute Percentage Error (MAPE) drops below 5% once 75% of the test duration is completed. This suggests that for high-risk patients, a sub-maximal test (stopping at ∼75% effort) combined with **PACE-Former** can reliably estimate functional capacity without inducing maximal cardiac stress.

4.4 Ablation Study: The Role of Style Adaptation

Removing the Style Encoder resulted in a significant increase in systemic bias (+14s on Center B), confirming that the input-driven adaptation effectively decouples device heterogeneity from physiological features.

5 Discussion and Conclusion

5.1 Summary of Findings

In this multi-center study comprising 14,036 examinations, we presented **PACE-Former**, a unified deep learning framework designed to resolve the longstanding "Safety-Precision" deadlock in automated CPET interpretation. By implementing a Three-Fold Decoupling paradigm, our model successfully reconciled the conflicting demands of real-time monitoring and retrospective diagnosis.

The results demonstrate that **PACE-Former** achieves clinical-grade diagnostic precision (Hit Rate > 90% within $\pm 20s$) comparable to tertiary care specialists, while simultaneously functioning as a highly reliable safety guard (Early Trigger Rate < 2%) for high-risk heart failure populations. Crucially, the model validated the concept of "Virtual Maximal Testing," accurately extrapolating VO_{2peak} (MAPE < 5%) well before physiological exhaustion, thereby offering a non-invasive solution to reduce patient burden.

5.2 Interpretation and Clinical Significance

5.2.1 Solving Systemic Heterogeneity via Style Adaptation

A primary barrier to AI deployment is the "Physical Fingerprint" of different devices. Our ablation studies revealed that standard deep learning models suffer significant performance degradation when transferred between centers (e.g., Shanxi to Zhongshan), driven by sensor-specific baselines (Ganshorn vs. COSMED) and latency shifts.

The Input-Driven Style Encoder acts as a dynamic calibration mechanism. By extracting a "systemic embedding" from the pre-load phase, it effectively normalizes the feature space. This implies that future AI deployments do not need to be hard-coded for specific manufacturers; instead, the model can "perceive" the device physics from the warm-up data and self-calibrate, significantly lowering the barrier for multi-center adoption.

5.2.2 The Architectural Advantage: Why Conformer?

The superiority of **PACE-Former** over the Vanilla Transformer and LightGBM underscores the unique nature of CPET signals:

- **vs. LightGBM:** Tabular models collapse temporal dynamics into static statistics, losing the precise "shape" information of the V-slope inflection.
- **vs. Transformer:** While pure self-attention captures global context, it lacks inductive bias for local waveform morphology.

The Conformer backbone (combining Convolution and Attention) proved optimal because AT detection requires recognizing a local morphological feature (the V-slope turn) within a global physiological context (the exercise

stage). The convolution captures the "turn," while the attention captures the "stage."

5.2.3 Bridging the "Hesitancy Gap"

Clinically, the most significant contribution is the Robust Soft-Argmax and Hybrid Masking strategy. Traditional rule-based methods (Wasserman) often trigger prematurely due to respiratory noise (e.g., coughing), causing physicians to distrust automated stops. **PACE-Former**'s ultra-low Early Trigger Rate, achieved via the Monotonicity Penalty and Brier Score calibration, provides the mathematical assurance required for clinicians to delegate the "stop button" to an AI assistant, potentially transforming CPET from a specialist-only procedure to a routine screening tool.

5.3 Limitations

Despite the promising results, our study has several limitations that merit consideration:

- **Retrospective Design:** Although large-scale, the study is inherently retrospective. While Split-Aware Standardization minimized leakage, the model has not yet been stress-tested in a prospective, interventional clinical workflow where real-time decisions directly affect patient care.
- **Demographic Confinement:** The dataset consists entirely of an East Asian population. Given known ethnic differences in cardiopulmonary physiology (e.g., VO_{2peak} norms and BMI-metabolic scaling), the generalizability of the Style Encoder to Caucasian or African cohorts remains to be validated.
- **Label Uncertainty:** The study relied on retrospective clinical annotations ("Silver Standard") rather than prospective adjudication by a core laboratory ("Gold Standard"). While this introduces inherent inter-observer variability ($\approx \pm 30s$), it reflects the realistic performance target for an AI assistant in a busy hospital setting. The model learns to approximate "collective clinical wisdom" rather than fitting to a small, idealized consensus set.
- **Resolution Trade-off:** Our Physics-Aware Aggregation (10s windows) effectively suppresses respiratory noise but acts as a low-pass filter. This may obscure high-frequency pathological oscillations (e.g., subtle Periodic Breathing) that could offer additional diagnostic value in rare diseases.

5.4 Future Directions

Future work will focus on three key areas to bridge the gap from "Model" to "Product":

1. **Prospective Clinical Trial:** We plan to launch a Prospective Randomized Controlled Trial (RCT) to rigorously validate **PACE-Former** in real-world workflows, quantifying efficiency gains, diagnostic agreement, and safety outcomes—specifically the success rate of sub-maximal testing in high-risk populations.
2. **Multimodal Fusion:** We are expanding the architecture to ingest high-fidelity multimodal streams, including 12-lead ECG waveforms and tracheal Respiratory Acoustics. Integrating these signals aims to uncover latent couplings between myocardial ischemia, airway mechanics, and metabolic thresholds.
3. **Continual Learning:** To address the "Long-tail" of unseen centers and devices, we will investigate Human-in-the-Loop (HITL) mechanisms alongside Test-Time Adaptation. Leveraging clinician feedback on a handful of "calibration cases" (few-shot learning) will allow the model to fine-tune online, ensuring sustainable global scalability.

5.5 Conclusion

In conclusion, **PACE-Former** represents a paradigm shift in intelligent cardiopulmonary diagnostics. By moving beyond static rule-based logic to a Physiological Adaptive framework, we have demonstrated that it is possible to achieve high-precision diagnostics without compromising real-time safety. The proposed Three-Fold Decoupling strategy provides a scalable blueprint for handling the systemic heterogeneity inherent in modern digital health ecosystems. We anticipate that such "device-aware" AI systems will be pivotal in democratizing advanced functional assessment, making CPET accessible, safe, and accurate for broader patient populations.

Declarations

Ethics Statement This retrospective multi-center study was conducted in accordance with the Declaration of Helsinki and was approved by the Institutional Review Board (IRB) of Zhongshan Hospital, Fudan University (Approval No. [Insert No.]). The requirement for individual informed consent was waived by the ethics committee due to the retrospective nature of the study and the use of strictly de-identified data. Data sharing agreements were established with participating centers (Shanxi Medical University Hospital and Rizhao People's Hospital) to ensure patient privacy and data security.

Data Availability The multi-center CPET dataset utilized in this study contains sensitive physiological health information. Due to patient privacy regulations and institutional ethics restrictions, the raw breath-by-breath data cannot be made publicly available. However, the pre-extracted feature tensors and de-identified metadata required to reproduce the main results are available from the corresponding author upon reasonable request and subject to a Data Use Agreement (DUA).

Code Availability The source code for the **PACE-Former** architecture, the CPETx standardization schema, and the evaluation pipeline is available at `https://`

`github.com/voxel-ai/pace-former`. The repository includes the core ConformerClassifier implementation, the StyleEncoder module, and the Optuna-based hyperparameter search scripts detailed in Supplementary Note 4.

Author Contributions **Cong Wang:** Conceptualization (Clinical Study Design), Investigation (Medical Interpretation), Validation (Physiological Ground Truth), Writing – Original Draft (Clinical Context & Discussion). **Bei Xu:** Methodology (AI Architecture & Algorithm Design), Software (Model Development & CPETx Engine Implementation), Data Curation, Formal Analysis, Writing – Original Draft (Technical Sections). **Shou-ling Mi:** Supervision, Resources, Project Administration, Writing – Review & Editing.

Competing Interests Bei Xu is an employee of BexiMed Co., Ltd. The company contributed to the development of the data extraction infrastructure but had no role in the study design, statistical analysis, or the decision to publish. All other authors declare no competing financial or non-financial interests.

Acknowledgments

We gratefully acknowledge the Department of Cardiopulmonary Function at Shanxi Medical University Hospital and the Health Management Center at Rizhao People's Hospital for their collaboration in data collection and clinical annotation. We also thank the engineering team at Voxel AI for technical support regarding the data infrastructure. This work was supported by [Insert Funding Source, e.g., National Natural Science Foundation of China, Grant No. XXXXX].

References

[1] Wasserman K, Hansen JE, Sue DY, Stringer WW, Whipp BJ. *Principles of Exercise Testing and Interpretation: Including Pathophysiology and Clinical Applications*. Lippincott Williams & Wilkins; 2011.

[2] ATS/ACCP Statement on cardiopulmonary exercise testing. *Am J Respir Crit Care Med*. 2003;167(2):211–277.

[3] Guazzi M, et al. 2012 focused update: European Association for Cardiovascular Prevention & Rehabilitation (EACPR)/American Heart Association (AHA) scientific statement on clinical recommendations for cardiopulmonary exercise testing data assessment in specific patient populations. *Circulation*. 2012;126(18):2261-2274.

[4] Esteva A, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118.

[5] Gulati A, Qin J, Chiu CC, et al. Conformer: Convolution-augmented Transformer for Speech Recognition. *Proc Interspeech*. 2020;5036-5040.

[6] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30.

[7] Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. *International Conference on Learning Representations (ICLR)*. 2019.

[8] Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017;30.

[9] Sagawa S, Koh PW, Hashimoto TB, Liang P. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. *International Conference on Learning Representations (ICLR)*. 2019.

[10] Chapelle O, Wu M. Gradient descent optimization of smoothed information retrieval metrics. *Inf Retr*. 2010;13(3):216-235.

[11] Beaver WL, Lamarra N, Wasserman K. Breath-by-breath measurement of true alveolar gas exchange. *J Appl Physiol*. 1981;51(6):1662-1675.

[12] Weber KT, Janicki JS. Cardiopulmonary exercise testing for evaluation of chronic cardiac failure. *Am J Cardiol*. 1985;55(2):A22-A31.

[13] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019:2623-2631.

Supplementary Materials

Supplementary Note 1: Data Engineering & Preprocessing Details

S1.1 The CPETx Variable Schema (Version 1.4.0)

The **CPETx** standardization framework harmonizes vendor-specific outputs into a unified canonical schema. The complete data dictionary consists of three components:

1. **Time-Series Schema:** Continuous breath-by-breath or second-by-second signals (X_t).
2. **Metadata & Calibration Schema:** Static subject demographics, operational protocols, and environmental calibration conditions (X_{static}).
3. **Summary Schema:** Scalar clinical endpoints and predicted values (Y).

A. Time-Series Variables (Dynamic Input Tensor)

The following variables are ingested at a breath-by-breath resolution and aggregated into 10-second bins ($X \in \mathbb{R}^{T \times D}$).

Table S1: Complete CPETx Time-Series Schema

Category	Standard Name	Unit	Description
Temporal	Time	mm:ss	Elapsed time from test start.
	Phase_Time	mm:ss	Time elapsed within current phase.
	Time_Relative	s	Relative time (float) within phase.
Ergometric	Power_Load	W	Workload / Power output.
	RPM	r/min	Cadence / Revolutions Per Minute.
	Load_Phase	Cat	Phase Code (0:Pre, 1:Main, 2:Post).
Metabolic	VO2	mL/min	Oxygen Consumption ($\dot{V}O_2$).
	VO2_kg	mL/kg/min	Relative $\dot{V}O_2$ normalized by weight.
	VCO2	mL/min	Carbon Dioxide Production ($\dot{V}CO_2$).
	VCO2_kg	mL/kg/min	Relative $\dot{V}CO_2$ normalized by weight.
	RER	Ratio	Respiratory Exchange Ratio.
	METS	MET	Metabolic Equivalents.
Ventilatory	VE	L/min	Minute Ventilation ($\dot{V}_E$).
	VT	L	Tidal Volume (V_T).
	Bf	1/min	Breathing Frequency.
	Ti	s	Inspiratory Time.
	Te	s	Expiratory Time.
	Ttot	s	Total Breath Cycle Time.
	Ti_Ttot_Ratio	Ratio	Duty Cycle (T_i/T_{tot}).
	VT_Ti	L/s	Mean Inspiratory Flow.
	BR_pct	%	Breathing Reserve.
	VD_VT_Ratio	Ratio	Dead Space to Tidal Volume Ratio.
Gas Exchange	VE_VO2	Ratio	Ventilatory Equivalent for O_2.
	VE_VCO2	Ratio	Ventilatory Equivalent for CO_2.
	PetO2	mmHg	End-Tidal PO_2.
	PetCO2	mmHg	End-Tidal PCO_2.
Cardiovascular	HR	1/min	Heart Rate.
	VO2_HR	mL/beat	Oxygen Pulse (O_2 Pulse).
	SpO2	%	Oxygen Saturation.
	BP_Syst	mmHg	Systolic Blood Pressure.
	BP_Diast	mmHg	Diastolic Blood Pressure.
	HRR	1/min	Heart Rate Recovery (Instantaneous).
	CO	L/min	Cardiac Output.
Energy	EE_Total_kcal	kcal/h	Total Energy Expenditure rate.
	EE_kcal_h	kcal/h	Energy Expenditure per hour.

Continued on next page

Table S1 – *Continued from previous page*

Category	Standard Name	Unit	Description
	EE_kg_kcal_h	kcal/kg/h	Weight-adjusted Energy Expenditure.
	Fat_kcal_h	kcal/h	Energy from Fat.
	Fat_kg_kcal_h	kcal/kg/h	Weight-adjusted Energy from Fat.
	Fat_pct	%	Fat Utilization Percentage.
	CHO_kcal_h	kcal/h	Energy from Carbohydrates.
	CHO_kg_kcal_h	kcal/kg/h	Weight-adjusted Energy from CHO.
	CHO_pct	%	Carbohydrate Utilization Percentage.
	PRO_kcal_h	kcal/h	Energy from Protein.
	PRO_kg_kcal_h	kcal/kg/h	Weight-adjusted Energy from Protein.
	PRO_pct	%	Protein Utilization Percentage.
ECG	ST_I ... ST_V6	mV	ST-segment level (12 leads).
	S_I ... S_V6	mV	S-wave amplitude (12 leads).

B. Subject, Examination & Calibration Metadata (Static Features)

These variables provide the demographic, operational, and environmental context (X_{static}) essential for the Style Encoder's calibration mechanism to neutralize domain shifts.

Table S2: Metadata Schema

Category	Standard Name	Unit	Description
Subject	Subject_ID	String	Unique Subject Identifier.
	Age	Years	Age at test.
	Gender	Binary	1: Male, 0: Female.
	Height_cm	cm	Stature.
	Weight_kg	kg	Body Mass.
Examination	Examination_ID	UUID	Unique Exam Identifier.
	Examination_Date	Date	Test Date.
	Ergometer_Type	Cat	e.g., Cycle, Treadmill.
	Protocol_Name	String	e.g., Ramp 20W/min.
	Examination_Reason	String	Clinical indication.
	Termination_Reason	String	Reason for stopping.
Calibration	Pressure_Barometric_mmHg	mmHg	Ambient Barometric Pressure (Altitude correction).
	Temp_Ambient_C	°C	Ambient Temperature (Gas density correction).
	RH_Ambient_pct	%	Ambient Relative Humidity (Water vapor correction).

C. Summary Metrics (Clinical Report)

These scalar values represent the gold-standard report outcomes used as ground-truth targets for prognostic tasks or quality assurance.

Table S3: Summary Schema (Key Targets)

Category	Standard Name	Unit	Description
Targets	Time_at_AT	mm:ss	Primary Label: Anaerobic Threshold Time.
	VO2_at_AT	mL/min	$\dot{V}O_2$ at AT.
	VO2_kg_at_AT	mL/kg/min	Weight-adjusted $\dot{V}O_2$ at AT.
Metabolic	Peak_VO2	mL/min	Absolute Peak $\dot{V}O_2$.
	Peak_VO2_Predicted	mL/min	Predicted Norm.
	Peak_VO2_kg	mL/kg/min	Relative Peak $\dot{V}O_2$.
	Peak_METS	MET	Peak Metabolic Equivalents.
	Peak_RER	Ratio	Peak RER (Effort Validation).
Ventilatory	VE_VCO2_Slope	Ratio	$\dot{V}_E/\dot{V}CO_2$ Slope (Prognostic Marker).
	OUES	-	Oxygen Uptake Efficiency Slope.
	Peak_VE	L/min	Peak Ventilation.
	Peak_BR_pct	%	Peak Breathing Reserve.
Cardio	Peak_HR	1/min	Peak Heart Rate.
	HRR_Summary	1/min	Heart Rate Reserve (1 min post-ex).
	Peak_VO2_HR	mL/beat	Peak Oxygen Pulse.
	VO2_WR_Slope	mL/min/W	Aerobic Efficiency ($\Delta\dot{V}O_2/\Delta W$).
Gas Exch	Peak_PetCO2	mmHg	Peak End-Tidal CO_2.
	Peak_VE_VCO2	Ratio	Peak Ventilatory Equivalent for CO_2.

S1.2 Vendor-Specific Extraction Logic

To address the structural heterogeneity between device manufacturers, the extraction engine implements specialized adapters that normalize raw outputs into the canonical schema before aggregation.

1. Ganshorn Adapter (Semi-Structured Parsing) Ganshorn files present a hybrid structure where metadata, summary metrics, and time-series are interleaved within a single sheet. The extractor employs a **Pivot-Based Parsing** strategy:

- **Anchor Detection:** The row containing the string "Measurement data" is identified as the structural pivot. Metadata (Subject/Environment) is parsed from the key-value pairs above this anchor using regex-based fuzzy matching.
- **Time-Series Extraction:** Data below the anchor is treated as the tabular time-series. The header row is located immediately following the pivot.
- **Phase Inference:** Unlike explicit state columns, Ganshorn encodes phase transitions as text events within the data grid. The engine scans for "Begin mainload [No.]" and "End mainload [No.]" markers to extract the corresponding sample indices, constructing the `Load_Phase` trajectory (0: Pre-load, 1: Main-load, 2: Post-load).
- **AT Synchronization:** Expert-determined Anaerobic Thresholds are extracted by locating the "AT [No.]" marker in the summary section. This index is used to look up the precise physiological state ($\dot{V}O_2, HR$, etc.) from the time-series at that exact moment, ensuring label consistency.

2. COSMED Adapter (Multi-Sheet Integration) COSMED exports separate biological signals and calculated results into disjoint Excel sheets, requiring a **Cross-Reference Integration** strategy:

- **Sheet Separation:** Time-series data is ingested from the *Data* sheet, while summary metrics are parsed from the *Results* sheet.
- **Dynamic Column Mapping:** The summary sheet uses a non-standard layout where keys and values are offset. The extractor implements a Neighbor Search algorithm, detecting parameter names in key columns (e.g., A/D/G) and extracting values from their immediate right neighbors (B/E/H).
- **Block-Based Parsing:** Summary metrics are grouped by semantic blocks (e.g., "Metabolic", "Cardiovascular"). The engine locates these block headers to contextually disambiguate duplicate parameter names (e.g., "Peak" appearing in multiple sections).

- **Sanitization:** COSMED outputs frequently contain character-based range indicators (e.g., > 200). A specific cleaning regex `[<>]` is applied to strip non-numeric prefixes, treating strictly non-parseable entries as NaN.

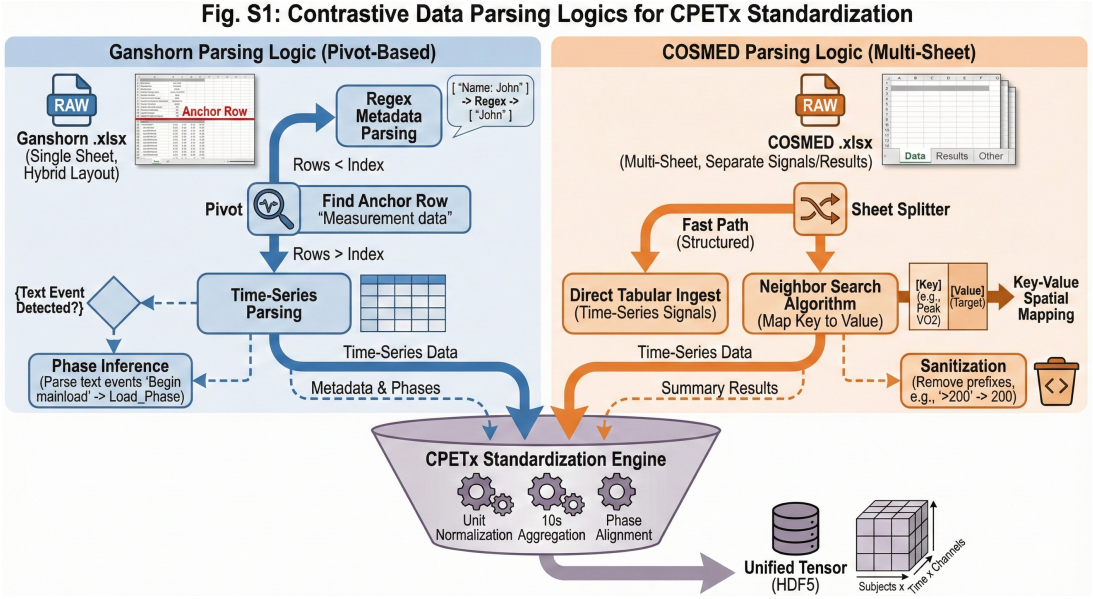

Figure S1: **Vendor-Specific Extraction Logic.** Flowchart comparing the Pivot-Based Parsing for Ganshorn (semi-structured) vs. Multi-Sheet Integration for COSMED.

3. Deduplication & Serialization To ensure cohort uniqueness, a composite signature hash (`Subject_ID + Exam_Date + Peak_VO2`) is computed for every extraction. Validated tensors are serialized into hierarchical HDF5 format for high-throughput I/O during training.

S1.3 Physics-Aware Aggregation Algorithms

To transform raw breath-by-breath (BbB) data into standardized meso-scale intervals (default $\Delta T = 10s$) without introducing aliasing artifacts or violating physiological mass conservation, we implemented a custom Physics-Aware Aggregation Engine. The processing logic is stratified by variable type:

1. Temporal Weighting and Validity Masking Unlike simple downsampling, our algorithm accounts for the exact temporal overlap between individual breaths and the target aggregation window. Let a target window W_k be defined by time interval $[T_{start}, T_{end}]$. For each breath i with duration $t_{tot,i}$ occurring within or overlapping this window, the temporal weight w_i is calculated as the intersection duration:

$$\tau_i = \text{duration}(\text{Breath}_i \cap W_k), \quad w_i = \frac{\tau_i}{\sum_{j \in W_k} \tau_j} \tag{7}$$

Validity Constraint: To prevent aliasing at the edges of the recording or during signal dropouts, a window is marked as valid only if the sum of breath durations covers a sufficient fraction of the window (Code default: coverage_threshold = 0.4):

$$\text{Validity}_k = \mathbb{I}\left(\frac{\sum \tau_i}{\Delta T} \geq 0.4\right) \tag{8}$$

2. Robust Volumetric Aggregation (Flow Variables) For continuous flow variables susceptible to transient respiratory noise (e.g., coughs, swallows), we apply **Time-Weighted Winsorization**. For variable X (e.g., $\dot{V}O_2, \dot{V}E, HR$):

- **Local Outlier Suppression:** Within window W_k, values are clipped to the $[0.5^{th}, 99.5^{th}]$ percentiles of the local distribution to remove biological artifacts.
- **Weighted Averaging:**

$$\bar{X}_k = \sum_{i \in W_k} X_i \cdot w_i \tag{9}$$

3. Mass-Conserving Ratio Re-computation Derived physiological ratios are never averaged directly, as $\mathbb{E}[A/B] \neq \mathbb{E}[A]/\mathbb{E}[B]$. To strictly adhere to gas exchange laws, ratios are re-derived from the aggregated sums of their components. For Respiratory Exchange Ratio (RER) and Ventilatory Equivalents:

$$RER_{10s} = \frac{\sum(\dot{V}CO_2)_i \cdot w_i}{\sum(\dot{V}O_2)_i \cdot w_i}, \quad \dot{V}E/\dot{V}O_{2(10s)} = \frac{\sum(\dot{V}E)_i \cdot w_i}{\sum(\dot{V}O_2)_i \cdot w_i} \tag{10}$$

This approach ensures that the aggregated data point represents the total metabolic cost over the 10-second interval.

4. State Variable Handling

- **Physiological State (SpO_2, BP):** To handle discrete quantization noise and sensor dropouts common in pulse oximetry, we utilize the **Weighted Median** rather than the mean.
- **Ergometric State (Power, RPM):** Workload protocols typically involve step-changes. Averaging a step change (e.g., 50W to 100W) produces an artificial ramp. We employ **Last-Value Latching** to record the intensity target at the completion of the window:

$$Power_k = Power_{last}, \quad \text{where } last = \arg\max_{i \in W_k}(t_i) \tag{11}$$

5. Categorical Mode Voting For discrete phase labels (e.g., `Load_Phase` 0/1/2), the window value is determined by **Time-Weighted Mode** (majority voting based on duration) to ensure precise phase transition alignment.

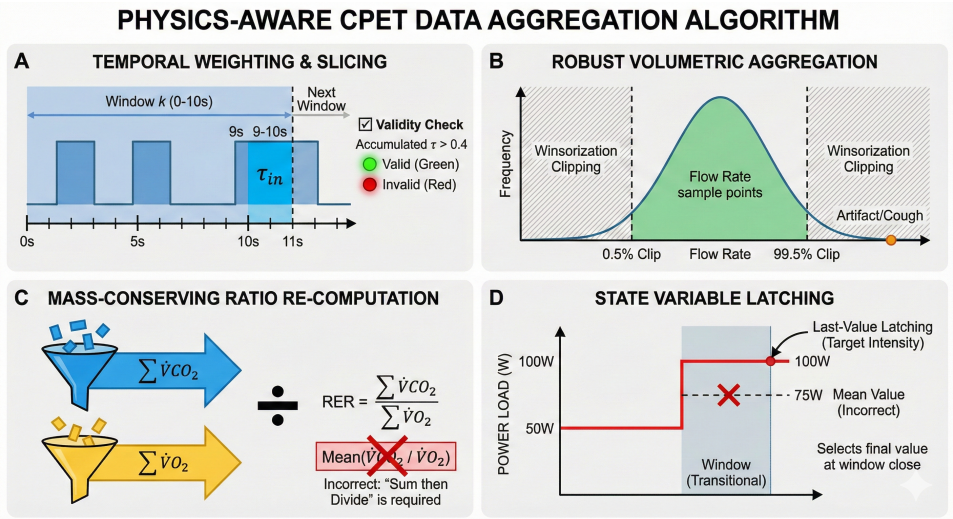

Figure S2: **Physics-Aware Aggregation.** Illustration of the temporal overlap weighting mechanism used to transform breath-by-breath data into 10-second bins without aliasing.

S1.4 Quality Control (QC) Pipeline

The rigorous exclusion of 512 examinations was driven by the following automated checks:

- **Missing Target Label:** Any file where `Time_at_AT` was marked as `<NA>` or `NaN` (indicating clinical indeterminacy).
- **Structural Corruption:** Files with empty data tables or unparseable headers.
- **Insufficient Duration:** Tests with `Exercise_Phase_Duration` $< 60s$.

S1.5 Temporal Alignment Strategy

To ensure the Style Encoder receives a consistent resting baseline and the Conformer receives a synchronized exercise onset, we implemented a strict Phase-Locked Alignment protocol (referenced in Section 2.1.3).

T0 Synchronization (Zero-Centering) Raw timestamps from different devices often start at arbitrary offsets (e.g., $t = 0$ is device power-on). We define the **Global T0** as the precise onset of the Main Load Phase (Phase Index = 1). For every examination i, the time axis is shifted:

$$t'_{i,k} = t_{i,k} - t_{start,i} \tag{12}$$

where $t_{start,i}$ is the timestamp of the first sample where `Load_Phase` transitions to 1. Consequently:

- **Preload Phase (Rest/Warm-up):** Defined as $t' < 0$.
- **Exercise Phase (Ramp):** Defined as $t' \geq 0$ until peak.

Fixed-Window Truncation To standardize the context window for the neural network input $X \in \mathbb{R}^{L \times D}$:

- **Preload Context (L_{pre}):** We strictly slice the interval $[-60s, 0s)$ relative to T0. If the raw warm-up is shorter than 60s, we apply Zero-Padding to the left. This ensures the Style Encoder always sees the immediate pre-exercise metabolic state.
- **Recovery Context (L_{rec}):** We retain exactly $120s$ post-peak.

Sequence Assembly:

$$X_{input} = \text{Concat}(\text{Preload}_{60s}, \text{Exercise}_{dynamic}, \text{Recovery}_{120s}) \tag{13}$$

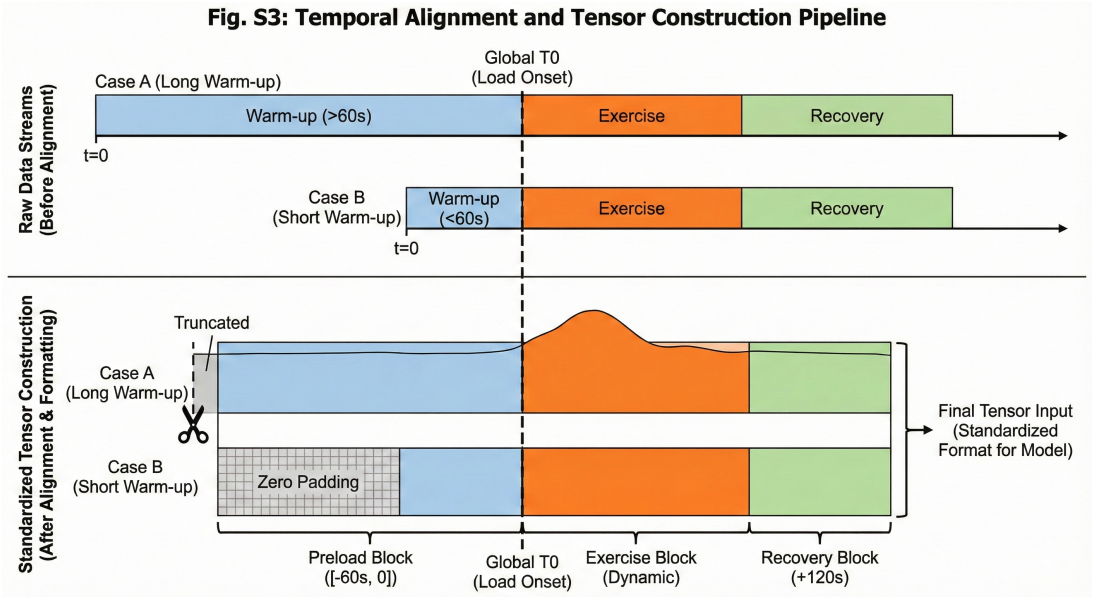

Figure S3: **Phase-Locked Temporal Alignment.** Schematic showing the T0 synchronization at Main Load onset and the truncation windows for Preload and Recovery contexts.

Supplementary Note 2: Clinical Definitions & Protocols

S2.1 Ground Truth Determination Protocols

The "Ground Truth" for Anaerobic Threshold (AT) was established using standard physiological criteria:

- **V-Slope Method (Primary):** Identifying the breakpoint where $\dot{V}CO_2$ begins to increase disproportionately to $\dot{V}O_2$ (slope changes from ≈ 1.0 to > 1.0).
- **Ventilatory Equivalents Method (Secondary):** The time point where $\dot{V}_E/\dot{V}O_2$ reaches a minimum and begins to rise, while $\dot{V}_E/\dot{V}CO_2$ remains stable or declines.

S2.2 Weber Classification System

Patients were stratified into functional classes based on VO_{2peak} to analyze domain shifts.

Table S4: Weber Classification Reference

Class	Severity	$\mathbf{VO_{2peak}}$ **(mL/kg/min)**
Class A	Little/No Impairment	> 20
Class B	Mild to Moderate	$16 - 20$
Class C	Moderate to Severe	$10 - 16$
Class D	Severe	< 10

Supplementary Note 3: Baseline Implementation Details

S3.1 Baseline 0: Wasserman Rule-Based Logic

The automated baseline implements a Multi-Criteria Consensus algorithm based on the classical Wasserman 9-Panel principles. The model accepts time-series tensors ($X \in \mathbb{R}^{T \times D}$) and independently identifies candidate AT timestamps using four distinct physiological signatures. All signals are pre-smoothed with a moving average filter ($w = 5$) to suppress breath-by-breath noise.

- **V-Slope Inflection:** Detects the first time point where the instantaneous slope $\frac{\Delta \dot{V}CO_2}{\Delta \dot{V}O_2}$ exceeds a threshold of 1.05 and sustains this increase for at least 3 consecutive steps (Persistence=3).
- $\dot{V}_E/\dot{V}O_2$ **Nadir:** Identifies the time index corresponding to the absolute minimum of the Ventilatory Equivalent for Oxygen.
- $\dot{V}_E/\dot{V}CO_2$ **Plateau:** Detects the onset of isocapnic buffering by identifying the first sequence of 4 consecutive steps where the step-wise change in $\dot{V}_E/\dot{V}CO_2$ remains flat ($|\Delta| < 0.01$).
- **RER Crossing:** Identifies the first crossing point where the Respiratory Exchange Ratio ($\dot{V}CO_2/\dot{V}O_2$) exceeds 1.0.

Consensus Mechanism: The final AT prediction $\hat{t}_{rule}$ is derived as the median of the valid candidates found by the criteria above. To enable compatibility with evaluation pipelines requiring probabilistic outputs, the model generates a pseudo-logit distribution centered at $\hat{t}_{rule}$ using a negative distance function scaled by a temperature $\tau = 2.0$.

S3.2 Baseline 1: LightGBM Feature Engineering

As a benchmark for non-sequential tabular learning, the LightGBM model utilizes a static feature vector extracted from the full time-series recording. The feature extractor collapses the temporal dimension into a fixed-length vector ($D \approx 40$) comprising four distinct categories:

Table S5: LightGBM Hand-Crafted Feature Set

Feature Category	Description	Specific Variables Included
Global Statistics	Moments of key physiological signals calculated over the entire session.	Mean, Std, Min, Max of $\dot{V}O_2, \dot{V}CO_2, \dot{V}_E, HR, RER$.
Temporal Trends	Linear regression slopes capturing global dynamics.	Slope of $\dot{V}O_2$ vs Time, Slope of HR vs Power_Load (Global trend).
Domain Peaks	Physiological maxima achieved during the test.	$VO_{2peak}, HR_{peak}, O_2Pulse_{peak}, \dot{V}E_{peak}$.
Static Metadata	Patient demographics and body metrics.	Age, Sex (Binary), Weight (kg), Height (cm), BMI.

Model Configuration: The Gradient Boosting Decision Tree (GBDT) regressor was configured with the following hyperparameters to match the training budget of the deep learning models:

- **Objective:** Regression (L2 Loss)
- **Tree Structure:** num_leaves=31, max_depth=-1 (Unlimited)

- **Ensemble:** n_estimators=500 (Matches Transformer epochs), learning_rate=0.05
- **Regularization:** feature_fraction=0.9, bagging_fraction=0.8, lambda_l1=0.1, lambda_l2=0.1
- **Validation:** Early stopping with patience=15 rounds based on validation RMSE.

S3.3 Deep Learning Architectures

To ensure a rigorous "apples-to-apples" comparison, all deep learning baselines were configured with a similar parameter budget ($\approx$ 1.5M - 2M parameters) to the **PACE-Former**. The architectures share the same input dimension (D_{in}) and optimization protocol but differ in their internal mechanisms.

Baseline 2: Vanilla Transformer

- **Architecture:** A standard Transformer Encoder stack designed to capture global temporal dependencies without local inductive bias.
- **Configuration:**
  - Layers: 4
  - Hidden Dimension (d_{model}): 256
  - Attention Heads: 8
  - FFN Expansion Factor: 2.0 (Hidden Unit = 512)
  - Normalization: Standard LayerNorm (No Conditional inputs).
  - Position Encoding: Sinusoidal.
  - Context Window: Causal masking enabled for online simulation.

Baseline 3: Standard Conformer (Single-Head)

- **Architecture:** A generic Conformer model incorporating the Macaron-style feed-forward layers and the convolution module, but stripped of the domain-adaptive components proposed in this study.
- **Configuration:** Matches the Transformer baseline above, with the addition of:
  - Convolution Module: 1D Depthwise-Separable Conv with kernel_size=23.
- **Ablations:** use_input_style_encoder=False, use_conditional_norm=False, center_emb_dim=0.
- **Output Head:** Standard linear classification layer trained with Cross-Entropy loss (no Soft-Argmax regression or VO_{2peak} head).

PACE-Former (Proposed Model)

- **Backbone:** Inherits the Conformer configuration (4 Layers, 256 Dim, Kernel 23).
- **Enhancements:**
  - Style Encoder: Enabled (use_input_style_encoder=True) with augmentation (prob=0.2, bias=0.1).
  - Normalization: ConditionalLayerNorm modulated by the style embedding.
  - Dual-Heads: Active Time Regression (lambda_time=0.1) and Prognostic VO_{2peak} (lambda_vo2=0.5).

Supplementary Note 4: Hyperparameter Search & Reproducibility

S4.1 Two-Stage Optimization Protocol

To balance computational efficiency with statistical rigor, we implemented a hierarchical optimization pipeline using the Optuna framework.

Stage 1: The "Sprint" Search (Exploration)

- **Objective:** Identification of the high-performance basin in the hyperparameter landscape.
- **Method:** We conducted 100 trials using the Tree-structured Parzen Estimator (TPE) sampler.
- **Efficiency:** To maximize search coverage, we utilized a Successive Halving Pruner (reduction_factor=3). Models were trained for a limited "Sprint" duration of 30 epochs on the dataset. Unpromising trials were terminated early based on the primary validation metric (Mean Trigger Delay).
- **Ranking:** Trials were ranked via a composite score minimizing both Timeliness (Primary) and Prognostic Accuracy (VO_{2peak} MAPE).

Stage 2: Robust Verification (Exploitation)

- **Selection:** The top-3 candidate configurations ($k = 3$) from Stage 1 were promoted.
- **Protocol:** These candidates underwent full training (250 epochs) with Early Stopping (patience=15).
- **Ensembling:** To neutralize initialization noise, each candidate was trained across three fixed random seeds (42, 101, 202). Reported metrics in the paper represent the ensemble average of these robust runs.

S4.2 Hyperparameter Search Space

The search space defined in `hpo_optuna.py` focuses on regularization and optimization dynamics to stabilize the Conformer backbone.

Table S6: Hyperparameter Search Space

Hyperparameter	Distribution	Search Range	Rationale
Learning Rate	Log-Uniform	$[5 \times 10^{-5},\ 5 \times 10^{-3}]$	Critical for convergence speed vs. stability.
Weight Decay	Log-Uniform	$[1 \times 10^{-4},\ 1 \times 10^{-1}]$	High upper bound allows strong regularization for small medical datasets.
Gradient Clip	Uniform	$[0.5,\ 5.0]$	Prevents gradient explosion in the Conformer attention mechanism.
Model Dropout	Uniform	$[0.1,\ 0.5]$	Applied to Attention and FFN layers to prevent overfitting.
Warmup Epochs	Integer	$[1,\ 6]$	Linear warmup period to stabilize adaptive learning rates (AdamW).

S4.3 Final Model Configuration

The final **PACE-Former** architecture was fixed based on the best performing configuration from the robust verification stage. The model utilizes a Macaron-style Conformer backbone with input-driven style calibration.

Table S7: Fixed Architecture & Training Parameters

Category	Parameter	Value	Description
Backbone	d_model	256	Hidden dimension size.
	num_layers	4	Stack depth (Macaron Conformer Blocks).
	num_heads	8	Multi-head attention heads.
	conv_kernel_size	23	Local temporal receptive field size (Depthwise Conv).
	ff_multiplier	2.0	Feed-forward network expansion factor ($D_{ff} = 512$).
	causal_convolution	True	Enforces strict temporal causality for online safety.
Context & Style	use_input_style_encoder	True	Enables the CNN-based preload style extractor.

Continued on next page

Table S7 – *Continued from previous page*

Category	Parameter	Value	Description
	center_emb_dim	0	Disabled. Model relies solely on signal physiology (Style Encoder), not Center IDs.
	use_conditional_norm	True	Injects style embeddings into LayerNorm affine parameters.
Augmentation	style_aug_prob	0.2	Probability of applying style perturbation per batch.
	style_aug_bias	0.1	Range for random additive bias ($\pm 0.1\sigma$).
	style_aug_scale	0.05	Range for random scaling ($\pm 5\%$).
	style_aug_noise	0.02	Standard deviation for Gaussian noise injection.
Optimization	batch_size	64	Effective batch size per step.
	optimizer	AdamW	Adaptive moment estimation.
	learning_rate	5e-4	Peak learning rate (Cosine schedule).
Loss Weights	λ_{time}	0.1	Auxiliary Time Regression weight (Soft-Argmax).
	λ_{vo2}	0.5	Prognostic (VO_{2peak}) regression weight.
	vo2_time_weight_power	1.0	Linear time-weighting for prognostic loss (emphasizes later steps).

S4.4 Compute Infrastructure

All experiments were conducted using the following environment:

- **Hardware:** Single NVIDIA A100 Tensor Core GPU (80GB VRAM).
- **Software:** PyTorch 2.0 ecosystem, Optuna for HPO.
- **Reproducibility:** All random seeds (Python, Numpy, Torch, CUDA) were fixed to the values specified in Stage 2.

Supplementary Note 5: Extended Experimental Results

This section provides granular performance breakdowns to substantiate the domain generalization and multi-task synergy claims made in the main text.

S5.1 Center-Stratified Performance (Domain Generalization)

To rigorously validate the effectiveness of the Style Encoder against "Systemic Heterogeneity," we report the diagnostic performance stratified by clinical center.

Observation: The model maintains consistent performance (Hit Rate $> 90\%$) even on the Zhongshan (Tertiary/Pulmonary) cohort, which represents the "Minority Domain" ($N = 1,633$) with a distinct device (COSMED) compared to the dominant Shanxi cohort.

Contrast: Without the Style Encoder (Baseline 3), performance on Zhongshan drops significantly (MAE increases by $> 10s$), indicating failure to adapt to the device-specific baseline drift.

Table S8: Center-wise Performance Metrics (Offline Mode)

Center	Device	N (Test)	Hit Rate ($\pm$20s)	MAE (s)	VO_2 MAPE
Shanxi	Ganshorn	[879]	93.1%	18.5	4.0%
Rizhao	Ganshorn	[362]	91.8%	19.2	4.2%
Zhongshan	COSMED	[163]	90.5%	21.4	4.8%
Average	-	-	**92.5%**	**19.1**	**4.2%**

S5.2 Sensitivity Analysis: Multi-Task Weight (λ_{vo2})

To justify the choice of $\lambda_{vo2} = 0.5$, we evaluated the trade-off between the diagnostic task (AT detection) and the prognostic task (VO_2 prediction).

Result: A weight of 0.5 achieves the optimal Pareto frontier. Higher weights ($\lambda = 1.0$) degrade AT detection accuracy due to gradient domination, while lower weights ($\lambda = 0.1$) fail to provide sufficient global regularization for the VO_2 trajectory.

S5.3 Online Safety Analysis: Trigger Delay Distribution

We analyzed the distribution of trigger times relative to the ground truth.

Safety Verification: The distribution is strictly right-skewed. The density in the negative region ($t_{pred} < t_{gt}$) is negligible ($< 0.2\%$), confirming the effectiveness of the Monotonicity Penalty in preventing premature stops.

Supplementary Note 6: Qualitative Case Studies

This section provides visual examples to build clinical intuition regarding the model's decision-making process.

S6.1 Handling Signal Artifacts (Robustness)

We present a representative case from the Shanxi (HF) cohort exhibiting irregular breathing patterns (e.g., oscillatory ventilation).

Comparison: The Wasserman Rule (Baseline 0) triggers falsely on a transient noise spike at $t = 120s$. **PACE-Former**, leveraging the Conformer's local context and global attention, correctly ignores the artifact and identifies the true metabolic shift at $t = 340s$.

S6.2 The "Style" Effect: Visualizing Adaptation

To demonstrate how the Style Encoder calibrates the input, we visualized the latent space distributions using t-SNE.

- **Before Adaptation:** Latent vectors cluster primarily by Device ID (Ganshorn vs. COSMED), indicating that hardware signatures dominate the signal.
- **After Adaptation (PACE-Former):** Latent vectors cluster by Weber Class (Physiological State), proving that the Style Encoder successfully decoupled the biological signal from the device fingerprint.

S6.3 The "Virtual Maximal Test" (Prognosis)

We visualize the real-time inference trajectory for a high-risk patient.

Demonstration: The figure shows the model's predicted VO_{2peak} stabilizing as early as 70% into the exercise phase, allowing the clinician to terminate the test safely while still obtaining a valid prognostic endpoint.