

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

Cong Wang<sup>1</sup>, Bei Xu<sup>2</sup>, and Shou-ling Mi\*<sup>1</sup>

<sup>1</sup>Zhongshan Hospital, Fudan University, Shanghai, China

<sup>2</sup>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 **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 device-specific bias. (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  $\text{VO}_{2\text{peak}}$  prediction for prognosis, enabling “Virtual Maximal Testing.”

**Results:** Validated on a multi-center cohort using 10-second aggregated data, the model achieved expert-level diagnostic precision in offline mode (Hit Rate within  $\pm 20\text{s} > 90\%$ ). In online mode, it maintained an Early Trigger Rate  $< 2\%$  while accurately predicting final  $\text{VO}_{2\text{peak}}$  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 ( $\text{VO}_{2\text{peak}}$ ) derived from CPET are critical biomarkers for risk stratification in heart failure, perioperative assessment, and rehabilitation prescription [1, 2].

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 [3]. Furthermore, the “ground truth” for AT relies on expert interpretation of multi-dimensional

curves, suffering from inherent inter-observer variability ( $\approx \pm 30\text{s}$ ).

### 1.2 The Data Challenge: Heterophasic Coupling

Unlike standardized DICOM images, CPET data are multivariate, non-stationary time series characterized by complex dynamics:

- **Heterophasic Coupling:** AT determination relies on the decoupling of linear relationships between  $\dot{V}O_2$ ,  $\dot{V}CO_2$ , and  $\dot{V}E$ . However, due to differences in chemoreceptor sensitivity and gas transport rates, these variables exhibit natural phase lags (e.g., ventilatory compensation  $\dot{V}E$  lags behind metabolic acidosis). Models must align these asynchronous cues.
- **Non-stationary Evolution:** From rest to exhaustion, the statistical distribution (mean, variance) of physiological signals drifts drastically. Models cannot rely on spatial invariance (as in CNNs for images) but must capture transient phase-change features within a dynamically evolving manifold.

### 1.3 The Clinical Dilemma: Safety vs. Precision

Current single-task models fail to address the contradictory dual needs of clinical workflows:

- **Online Monitoring (Safety First):** For high-risk patients, clinicians need a “Virtual Maximal Test”—predicting  $\text{VO}_{2\text{peak}}$  early (e.g., at 75% load) to terminate the test safely. This requires an extremely low **Early Trigger Rate**; false alarms causing premature termination are unacceptable.
- **Offline Reporting (Precision First):** Retrospective diagnosis requires unbiased temporal localization ( $\text{Bias} \approx 0$ ) to match expert consensus.

### 1.4 The Deployment Bottleneck: Systemic Heterogeneity

A major barrier to multi-center deployment is the **Holistic Systemic Fingerprint**. Differences in hardware (Cortex vs. Cosmed), environmental physics (barometric pressure), and protocols (mask dead space) create severe systemic time biases ( $> 40\text{s}$ ) across centers, hindering generalization.

To address these challenges, we introduce **PACE-Former**, a Conformer-based framework that utilizes systemic style adaptation and hybrid task learning to unify real-time safety and offline precision.

## 2 Methodology

### 2.1 Data Strategy: Meso-scale Aggregation and Physical Truncation

To mitigate high-frequency noise caused by hyperventilation near the limits of tolerance, we employ a standard operating procedure (SOP) for data processing:

1. **Meso-scale Aggregation:** Breath-by-breath data is resampled into 10-second bins (0.1 Hz). This smoothing highlights physiological trends over breath-to-breath noise while maintaining sufficient resolution for clinical diagnosis.

#### 2. Physical Truncation:

- **Preload:** We retain the final 180 seconds of the resting/warm-up phase. This window contains minimal physiological flux and is used solely to capture the “systemic fingerprint” (device noise floor, baseline drift).
- **Postload:** We retain only a 120-second buffer after the load phase ends. This covers the physiological lag of  $\text{VO}_{2\text{peak}}$  while physically removing redundant recovery data to prevent the model from shortcut learning (i.e., detecting AT solely by looking far into the recovery phase).

### 2.2 Model Architecture: The PACE-Former

The proposed architecture (Fig. ??) integrates three key components:

#### 2.2.1 Style-Aware Backbone

We introduce a lightweight 1D-CNN **Style Encoder** that processes the Preload stream. It extracts statistical moments (mean, variance, texture) representing the device and patient baseline. These style embeddings are injected into the main network via **Conditional Layer Normalization (CLN)**. Let  $x$  be the feature input. CLN adapts the normalized features using affine parameters  $(\gamma, \beta)$  generated from the style embedding  $s$ :

$$\text{CLN}(x, s) = \frac{x - \mu}{\sigma} \cdot \gamma(s) + \beta(s) \quad (1)$$

This allows the network to perform “adaptive normalization,” effectively removing site-specific bias before physiological feature extraction.

The backbone utilizes **Conformer Blocks** [4], combining Convolutional modules (to capture local morphological trends like V-slope deflection) and Self-Attention (to capture long-range heterophasic coupling between metabolic production and ventilatory response).

#### 2.2.2 Dual-Task Heads

- **Time Head (Diagnostic):** Outputs a sequence of probabilities indicating if AT has occurred. We use **Soft-Argmax** during inference to regress sub-bin continuous time indices, overcoming the quantization error of 10s binning.
- **Value Head (Prognostic):** A scalar regression head that predicts the final  $\text{VO}_{2\text{peak}}$  at every time step. This enables the assessment of the “Virtual Maximal Test” capability.

### 2.3 Hybrid Training Strategy

To unify online and offline capabilities in a single model, we employ **Dynamic Mask Sampling**:

- **Online Mode ( $p = 0.5$ ):** A Causal Mask (upper triangular) is applied to the Self-Attention mechanism. The model can only attend to historical data, optimizing for low latency and monotonic probability progression.
- **Offline Mode ( $p = 0.5$ ):** The mask is removed. The model utilizes bidirectional attention, leveraging recovery phase features (e.g., rapid drop in HR) to refine AT localization.

### 2.4 Loss Function

The total loss combines classification, regression, and regularization terms:

$$\mathcal{L} = \mathcal{L}_{BCE} + \lambda_1 \mathcal{L}_{Mono} + \lambda_2 \mathcal{L}_{Time} + \lambda_3 \mathcal{L}_{VO2} \quad (2)$$

Where  $\mathcal{L}_{Mono}$  enforces monotonic non-decreasing probabilities for the AT event, and  $\mathcal{L}_{VO_2}$  uses time-weighted MSE (heavier weights near the end) to encourage early convergence of prognostic predictions.

## 3 Experimental Design

### 3.1 Cohorts and Cross-Validation

We utilized a multi-center retrospective cohort comprising data from tertiary hospitals (Zhongshan Hospital) and health checkup centers, covering diverse devices (Cosmed, Cortex). We employed Leave-One-Center-Out (LOCO) cross-validation to rigorously test generalization. Ground truth AT labels were generated via a double-blind expert consensus process.

### 3.2 Dual-Mode Evaluation Metrics

We established distinct metric sets for the two deployment scenarios:

Table 1: Dual-Mode Evaluation Metrics

Mode	Key Metric	Target
Online	<b>Early Trigger Rate</b>	< 2% (Safety)
	Mean Trigger Delay	< 30s
	VO <sub>2</sub> MAPE @ 75%	< 5% (Prognosis)
	VO <sub>2</sub> Stability	Low Variance
Offline	<b>Hit Rate @ 20s</b>	> 90% (Precision)
	Time Bias	≈ 0s
	Bland-Altman LoA	Clinical limits

## 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 ±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 (±30s).

### 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<sub>2peak</sub> 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

This study presents the first CPET analysis framework to explicitly decouple and optimize for the conflicting requirements of safety and precision. By leveraging meso-scale aggregation (10s bins) and Conditional Layer Normalization, **PACE-Former** overcomes the noise and heterogeneity inherent in multi-center respiratory data.

The physiological significance of the **Conformer backbone** is evident in its ability to handle heterophasic coupling; the attention mechanism naturally aligns the lagged ventilatory response with metabolic events. Furthermore, the **Dual-Head design** validates the feasibility of "Virtual Maximal Testing," potentially transforming CPET protocols for heart failure and perioperative populations.

## 6 Conclusion

**PACE-Former** establishes a new technical standard for automated CPET interpretation. It provides a clinically safe, diagnostically precise, and universally applicable tool that requires minimal calibration, paving the way for large-scale, standardized cardiopulmonary phenotyping.

## References

- [1] Guazzi M, et al. 2016 European Guidelines on cardiovascular disease prevention in clinical practice. *Eur Heart J*. 2016;37:2315-2381.
- [2] Wasserman K, et al. *Principles of Exercise Testing and Interpretation*. 5th edn. Lippincott Williams & Wilkins; 2012.
- [3] Beaver WL, et al. A new method for detecting anaerobic threshold by gas exchange. *J Appl Physiol*. 1986;60:2020-2027.

- [4] Gulati A, et al. Conformer: Convolution-augmented Transformer for Speech Recognition. *Interspeech*. 2020.

## Appendix: Standardized Data Model

The following tables describe the standardized data schema used to harmonize multi-center CPET data for **PACE-Former**.

**Table S1.** Time-series Variables (Aggregated per 10s)

Variable	Unit	Type	Description
Time	s	float	Relative time from exercise start.
$\dot{V}O_2$	mL/min	float	Oxygen uptake.
$\dot{V}O_2/\text{kg}$	mL/kg/min	float	Relative oxygen uptake.
$\dot{V}CO_2$	mL/min	float	Carbon dioxide production.
$\dot{V}_E$	L/min	float	Minute ventilation.
RER	ratio	float	Respiratory Exchange Ratio ( $\dot{V}CO_2/\dot{V}O_2$ ).
PetO <sub>2</sub>	mmHg	float	End-tidal Oxygen Tension.
PetCO <sub>2</sub>	mmHg	float	End-tidal Carbon Dioxide Tension.
$\dot{V}_E/\dot{V}O_2$	ratio	float	Ventilatory equivalent for O <sub>2</sub> .
$\dot{V}_E/\dot{V}CO_2$	ratio	float	Ventilatory equivalent for CO <sub>2</sub> .
HR	bpm	int	Heart Rate.
Power	Watts	float	Ergometer workload.
RR	1/min	float	Respiratory Rate.
VT	L	float	Tidal Volume.

**Table S2.** Static & Target Variables

Variable	Unit	Type	Description
Age	years	int	Patient age.
Sex	binary	int	0: Female, 1: Male.
BMI	kg/m <sup>2</sup>	float	Body Mass Index.
Center_ID	cat	int	Anonymized center identifier.
Device_Type	cat	int	Device manufacturer code (e.g., 0:Cortex, 1:Cosmed).
AT_Time	s	float	Consensus ground truth time of AT.
VO <sub>2peak</sub>	mL/min	float	Global maximum VO <sub>2</sub> (including recovery buffer).