Supplementary Material

S1 Further Dataset Information

The dataset used in this study includes recordings from 10 subjects, each performing six types of physical activities: walking, running, incline walking, backward walking, cycling, and stair climbing. These activities were divided into two main experimental sessions:
Session 1: Sitting, standing, level walking, incline walking, and backward walking on a treadmill
Session 2: Sitting, standing, running on a treadmill, cycling on a stationary bike, and stair climbing on a stairmill.
For each activity, subjects followed a protocol in which they stood quietly for 6 minutes, performed each speed/resistance condition for 6 minutes in randomized order, and then sat quietly for another 6 minutes. All sensor data were recorded with synchronized timestamps, and each time step was manually annotated with the corresponding activity type.

Below, we mentioned the utilized signals:
(i) acceleration magnitudes from the left/right ankles and wrist, waist, and chest (For all acceleration signals, the vector magnitude across the three axes was calculated).
(ii) right/left wrist electrodermal activity (EDA) and skin temperature
(iii) a composite lower-limb signal derived from normalized sEMG envelopes
(iv) respiratory and cardiovascular measures, including $\dot{V}_{O_2}$, $\dot{V}_{CO_2}$, SpO_2, breath frequency, minute ventilation, and heart rate. Respiratory measures were collected breath-by-breath with a portable respirometer, while heart rate and SpO_2 were measured with a chest strap and earlobe oximeter. All signals were synchronized with the respirometer and stored on a breath-by-breath basis. Because these signals were measured breath-by-breath, their sampling frequency varied. We maintained a constant sample rate by averaging over the frequency of breath for each activity and its specific condition (e.g., backward walking at $1\,\mathrm{m/s}$ vs. $0.7\,\mathrm{m/s}$) [20].
The ground truth energy expenditure: It was computed using the Brockway equation [1], which relies on measurements of $\dot{V}_{O_2}$a and $\dot{V}_{CO_2}$. The resulting values were normalized to body weight for comparability across subjects. Steady-state EE was estimated by averaging the final three minutes of each six-minute activity condition. To obtain the net energetic cost, the standing baseline value recorded at the start of each trial was subtracted from the steady-state estimate. (Further details on data collection and processing can be found in [8].)
Input formatting and fusion: In all experiments, input signals were segmented into fixed-length windows of 10 or 20 time steps. The final choice of time step size was selected based on preliminary performance tuning. For multi-signal inputs, we applied early fusion by concatenating the signals along the feature dimension.

S2 Detailed Model Architecture and Training Procedure

In this study, we tested six models: Linear Regression, CNN, LSTM, ResNet, ResNet+Attention, and Transformer. For each model, we provide details on the network architecture, training configuration, and implementation choices, including layer design and optimization settings.

S2.1 Linear Regression:

We implemented both single and multiple linear regression models for the energy expenditure (EE) estimation. The general form of the model is:

$$\hat{y} = b_0 + \sum_{i=1}^{n} b_i x_i = Xb \tag{1}$$

where $\hat{y}$ represents the vector of predicted EE values, the variable n denotes the number of input signals included in the model. The input matrix X consists of a column of ones to account for the bias term and n columns representing the input signals. The vector b contains the learned regression coefficients.

S2.2 CNN:

The CNN model consists of three 1D convolutional blocks followed by fully connected layers. Each convolutional block includes a 1D convolution layer (kernel size = 3), batch normalization, ReLU activation, max pooling (kernel size = 2), with dropout applied in the second and third blocks. The number of filters decreases across the layers (64, 32, and 16). The convolutional output is flattened and passed through two fully connected layers: The first is a dense layer with 40 units, ReLU activation, and dropout. The second is an output layer with linear activation to match the target dimension. We used 20 time steps, a batch size of 8, and the Adam optimizer (learning rate = 0.0005) for training this network.

S2.3 LSTM:

We implemented a stacked LSTM-based regression network. The model consists of two sequential Long Short-Term Memory (LSTM) layers. The first LSTM layer has 128 hidden units, followed by dropout regularization. Its output is passed to a second LSTM layer with 64 hidden units and an additional dropout. The final LSTM output is flattened and passed through a fully connected layer with 64 units, followed by batch normalization and dropout. We set the number of time steps to 20, used a batch size of 32, and used the Adam optimizer with a learning rate of 0.0005.

S2.4 ResNet:

The original ResNet architecture is adapted for 1D time-series input. The model architecture begins with a 1D convolution using 64 filters (kernel size = 7), followed by batch normalization, ReLU activation, and max pooling. Next, there are three residual blocks with increasing output dimensions (64 to 128 channels, 128 to 256 channels, and 256 to 512 channels). Each block contains two Conv1D layers (kernel size = 3) with batch normalization layers and skip connections (including a convolution to match the input and output dimensions). After the final residual block, global average pooling is applied over the time dimension, followed by a linear layer mapping the pooled features to the desired output size. The network was trained with 10 time steps, a batch size of 32, and the Adam optimizer with a learning rate of 0.001.

S2.5 ResNet+Attention:

In this architecture, there is an attention block added to the residual blocks in our ResNet architecture. This block computes a self-attention mechanism over the temporal dimension.

Within this block, three distinct 1×1 convolutions are applied to derive the query, key, and value representations of the input.

The attention score computed using the attention function equation from [24], based on a scaled dot-product attention mechanism over the temporal dimension. After re-weighting the value features, a residual connection integrates the attention output with the original input, ensuring that the initial features are preserved. The network was trained with 10 time steps, a batch size of 8, and the Adam optimizer (learning rate = 0.0005).

S2.6 Transformer:

This model is based on the Transformer encoder framework [24], adapted for sequential signal modeling. The raw input signal is first projected into a higher-dimensional representation using a 1D convolution with kernel size 3, which captures local temporal patterns. Since the Transformer architecture does not contain recurrence or convolutional structure, temporal order is incorporated through sinusoidal positional encodings as introduced in [24]. The projected sequence is then processed by a stack of two Transformer encoder layers, each consisting of 8-head self-attention, a feedforward network with hidden dimension 256, residual connections, and layer normalization. Finally, the encoded sequence is passed through a lightweight feedforward output head composed of two fully connected layers with a ReLU activation in between to produce predictions at each time step. For training, we used a time step length of 10, a batch size of 4, and the Adam optimizer with a learning rate of 0.0009.

S3 Additional Per-activity Evaluation Tables

In "Per-Activity Evaluation" (Section 3.4) of the paper, we discussed how model performance varied across activities. To complement that analysis, we provide a detailed summary of pairwise signal combinations here. Table S1 presents the worst-performing pairs (left) and the best partner for each signal when minute ventilation was excluded (right). These results highlight which modalities provide complementary information and which pairs lead to consistently poor predictions.

Signal 1	Signal 2	Model	RMSE (W/kg))
L_Wrist_Temp	R_Wrist_Temp	ResNet	3.10
Waist_ACCL	L_Wrist_ACCL	ResAtt	3.12
R_Wrist_Elec	R_Wrist_Temp	ResNet	3.15
L_Wrist_Elec	L_Wrist_Temp	Lin-Reg	3.19
L_Wrist_Temp	R_Wrist_Temp	Trans	3.21
L_Wrist_Elec	R_Wrist_Temp	Lin-Reg	3.23
L_Wrist_Temp	R_Wrist_Temp	LSTM	3.25
EMG_M_L	SpO2	CNN	3.35
Waist_ACCL	Chest_ACCL	ResAtt	3.36
EMG_M_R	SpO_2	CNN	3.44
Waist_ACCL	R_Wrist_ACCL	ResAtt	3.81
EMG_M_R	SpO_2	LSTM	4.36
EMG_M_L	R_Wrist_Elec	CNN	4.51
EMG_M_R	R_Wrist_Elec	CNN	5.28
EMG_M_R	L_Wrist_Elec	CNN	7.90
EMG_M_L	L_Wrist_Elec	CNN	8.05

Signal	Best Pair	Model	RMSE (W/kg))
Waist_ACCL	EMG_M_L	Trans	1.64
Chest_ACCL	HR	CNN	1.67
L_Ankle_ACCL	HR	CNN	1.51
R_Ankle_ACCL	HR	CNN	1.49
L_Wrist_ACCL	HR	CNN	1.79
L_Wrist_Elec	HR	CNN	1.63
L_Wrist_Temp	R_Ankle_ACCL	CNN	1.93
R_Wrist_Elec	HR	CNN	1.64
R_Wrist_Temp	R_Ankle_ACCL	CNN	1.87
R_Wrist_ACCL	HR	CNN	1.81
EMG_M_L	Waist_ACCL	Trans	1.64
EMG_M_R	L_Ankle_ACCL	ResAtt	1.53
HR	R_Ankle_ACCL	CNN	1.49
SpO_2	HR	CNN	1.86
Breath_Freq	R_Ankle_ACCL	CNN	1.91
Min_Vent	EMG_M_L	ResAtt	0.90

Table S1: **Left**: Worst signal pair combinations. **Right**: Best pair for each signal (in absence of Minute Ventilation).

S4 Additional Per-Subject Evaluation Plots

In "Per-Subject Evaluation" (Section 3.5) of the paper, we focused on the Transformer and CNN models and their corresponding plots. In Figure S1, we provide the per-subject performance plots for the four remaining models. The trends largely mirror those observed for the Transformer and CNN models, confirming the key observations in the main paper.

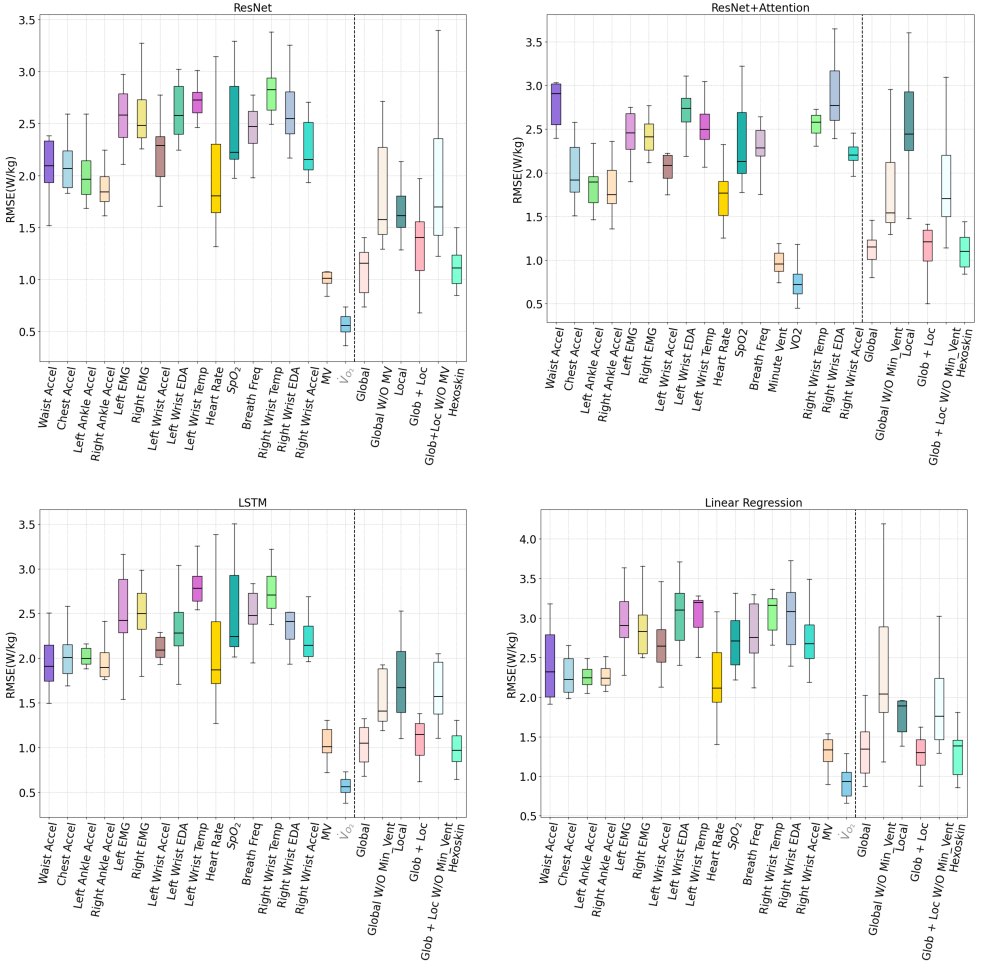

Figure S1: Performance of the remaining four models for single and grouped signals across 10 subjects (complementing Section 3.5 of the main paper). The dashed line separates single- from grouped-signals in each plot. Boxplots represent the distribution of RMSE values across subjects: median (line), 25th–75th percentiles (box), and whiskers to 1.5×IQR. (MV: Minute Ventilation)