ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Application of deep learning model with multi-modal data fusion in clothing comfort prediction

Miao Yu[1,*]

[1] Faculty of Art and Design, Shanghai Business School, Shanghai, 200235, China

SUMMARY: *Aiming at the problems of insufficient representation of single data source, fixed modal fusion weight and low prediction accuracy of continuous comfort score in clothing comfort prediction, a multimodal data fusion deep learning prediction model was constructed. Taking clothing images, human heat and humidity pressure sensing sequences, fabric structural parameters and environmental variables as input, the model uses convolution branch, bidirectional gated recurrent branch and structured coding branch to extract heterogeneous features, and completes dynamic weight allocation through cross-modal attention layer to output comfort level and continuous comfort score. The validation was carried out based on the GarmentComfort-MM dataset, and the total number of samples was 7200 groups. The experimental results show that the Accuracy, Precision, Recall and F1-score of the model on the test set reach 93.8%, 93.4%, 92.9% and 93.1%, respectively. The MAE and RMSE of the continuous comfort score prediction are 0.041 and 0.058, respectively. The Accuracy decreases by 4.6 percentage points after removing the sensing branch, which indicates that the dynamic information of heat and humidity pressure contributes significantly to the comfort prediction. The research can provide technical support for the intelligent evaluation of clothing comfort.*

KEYWORDS: *Multimodal data fusion; Deep learning; Clothing comfort; Prediction model*

1 Introduction

Clothing comfort prediction involves many factors such as human body heat and humidity state, fabric structure parameters, clothing pressure distribution, environmental temperature and humidity, and movement behavior changes. Traditional evaluation methods mostly rely on subjective questionnaires, single physical tests or static thermal resistance indicators, which can reflect the comfort performance of a certain material or a certain wearing condition, but it is difficult to describe the dynamic coupling relationship between fabric contact, heat transfer, moisture diffusion and body surface perception during human activities. Previous studies have summarized the constitutive mechanism of thermal and physiological comfort of clothing from the perspective of textile science, and pointed out that thermal conductivity, moisture permeability, air layer structure and human metabolic state will all affect the final comfort feeling [1]. For complex wearing scenes such as tight-fitting sportswear, researchers begin to record human motion, clothing pressure and subjective comfort feedback through multi-sensor data acquisition, which gradually shifts clothing comfort evaluation from single-point measurement to multi-source data-driven analysis [2]. Fabric friction, contact electrical signals and surface behavior are also used to build a comfort perception system,

*sbssjcd@163.com
https://doi.org/10.65102/is2026969
Published: 30 April, 2026

which provides a new data entry for clothing wearing state recognition [3].

With the development of computer vision, intelligent sensing and deep learning methods, clothing comfort prediction has begun to have a stronger basis for automated modeling. Clothing insulation level classification method based on image recognition and transfer learning can convert wearing images into thermal comfort calculation input, and reduce the cost of manual annotation and on-site measurement [4]. After the combination of thermal imaging and deep learning, the model can extract the difference of comfort state from the body surface temperature distribution, and improve the adaptation ability of thermal comfort prediction for different genders [5]. According to the different dressing states of indoor personnel, the prediction model of local clothing thermal resistance further shows that clothing comfort is not a single overall indicator, but is formed by the joint effect of local coverage, posture changes and material differences [6]. Research on textile material properties and hand feel prediction also shows that artificial intelligence models have been able to learn the nonlinear relationship between fabric structure, material parameters and perceptual evaluation [7]. The application of data-driven decision system in the design of new textile fabrics shows that predictive analysis and optimization analysis can jointly support the research and development process of clothing materials [8].

Existing research provides a methodological basis for intelligent garment comfort prediction, but there are still three shortcomings: first, some studies focus on a single image, a single sensor signal or a single fabric parameter, and it is difficult to use human body state, fabric structure and environmental variables at the same time. Second, there are differences in sampling frequency, scale range and noise level in multi-modal data, and simple splicing is easy to cause feature redundancy and modal weight imbalance. Third, the prediction model often focuses on the result output, and the analysis of cross-modal feature alignment, fusion weight update and error sources is insufficient. Multimodal hybrid deep learning studies have shown that images, sensing, and structured variables can form more stable predictive representations through unified embedding, feature interaction, and attention fusion [10]. Deep multimodal data fusion methods can deal with the complementary relationship between heterogeneous information and provide model support for comfort prediction in complex scenes [11]. The attention fusion mechanism still has strong robustness under the condition of missing modalities and noise interference, which can provide reference for multi-source data modeling of clothing comfort [12].

Based on the above research, this paper constructs a deep learning clothing comfort prediction model based on multimodal data fusion. The model takes clothing images, fabric structural parameters, human heat and humidity sensor data and environmental variables as input, and realizes the synchronous prediction of clothing comfort level and continuous comfort score through unified tensor construction, feature embedding coding, cross-modal attention fusion and classification regression joint output. The research focus is not on the concept explanation of clothing comfort, but on the multimodal input organization, deep fusion network structure, end-to-end training process and prediction performance verification, and strive to form a computable, trainable, and comparably verified intelligent prediction method for clothing comfort.

2 Multimodal Data Fusion Deep Learning Prediction Model Construction

2.1 Modeling of Multimodal Input Data for Clothing Comfort Prediction

Clothing comfort prediction cannot rely solely on a single fabric parameter or a single wearing image. The human wearing sensation is influenced by the fabric structure, the fitting state of the clothing, the changes in body surface heat and moisture, local pressure, exercise intensity, and environmental conditions [13-15]. The input end organizes the clothing images, human sensor sequences, fabric structure parameters, and environmental variables into a unified sample unit. The image modality records the clothing outline, fold distribution, fitting degree, and local coverage area; the sensor modality collects body surface temperature, skin humidity, the microclimate temperature and humidity inside the clothing, and local pressure; the structured modality stores the fabric thickness, weight, thermal resistance, moisture permeability rate, elastic recovery rate, and bending stiffness; the environmental modality records the experimental temperature, relative humidity, wind speed, and exercise intensity. All types of data are aligned using the same wearing experiment number as the index to avoid sample mismatch between different modalities [16-18].

Let the i-th wearing experiment sample consist of clothing images, sensor sequence, structured parameters, environmental variables, and comfort labels. The multi-modal input sample is represented as:

$$X_i=\{I_i,S_i,P_i,E_i,y_i\},\quad S_i\in R^{T\times C_s},\ P_i\in R^{C_p},\ E_i\in R^{C_e} \tag{1}$$

where, Xi represents the ith multimodal input sample; Ii represents the clothing image; Si represents the human sensing timing matrix; Pi represents fabric and garment structural parameter vector; Ei represents the environment variable vector; yi denotes the comfort label; T represents the sensing sampling window length; Cs represents the number of sensing channels; Cp denotes structured parameter dimension; Ce represents the environment variable dimension.

The comfort labels are constructed in parallel with the rating labels using continuous scores. After the wearing experiment, the subjects gave the scores of heat, humidity, pressure, activity obstruction and overall comfort. The scores were normalized to the range [0,1], and the comprehensive comfort score was formed by combining the abnormal sensing intensity:

$$y_i=\alpha_1 q_i^{heat}+\alpha_2 q_i^{wet}+\alpha_3 q_i^{press}+\alpha_4 q_i^{motion}+\alpha_5 q_i^{overall}-\alpha_6 r_i^{sensor} \tag{2}$$

where, q_i^{heat} represents the normalized thermal sensation score; q_i^{wet} stands for wetness normalized score. q_i^{press} is the pressure normalized score; q_i^{motion} represents the normalized score of activity obstruction; $q_i^{overall}$ denotes the overall comfort normalized score; r_i^{sensor} represents the sensing anomaly term formed by temperature and humidity mutation, pressure peak and microclimate fluctuation. α1 to α6 represent the weight coefficients of each term. The continuous scores are used for regression prediction, which are further divided into three levels of "uncomfortable, general, and comfortable" according to the threshold interval for classification prediction tasks.

The multimodal variable configuration needs to correspond to the comfort prediction target to avoid feature redundancy caused by the stacking of irrelevant indicators. Table 1 organizes the input variables and their modeling roles.

Table 1: Configuration Table of Multimodal Input Variables for Clothing Comfort

Modality Type	Input Variables	Data Function
Image Modality	Front image, side image, local wrinkle image	Extract clothing contour, fitting state, and local deformation features
Sensor Modality	Body surface temperature, skin humidity, in-clothing temperature and humidity, local pressure	Represent thermal-humidity changes and pressure state during wearing
Structured Modality	Thickness, gram weight, thermal resistance, moisture permeability, elastic recovery rate, bending stiffness	Describe the basic physical properties of fabrics and clothing materials
Environmental Modality	Ambient temperature, relative humidity, wind speed, exercise intensity	Correct the influence of external conditions and human activity on comfort
Label Data	Comfort level, continuous comfort score	Support joint training of classification prediction and regression prediction

After completing the configuration of multimodal variables, it is necessary to further clarify the processing order of different data sources before entering the model. Figure 1 shows the input modeling process from raw data acquisition to multimodal sample packaging, in which the image, sensing, fabric parameters and environment variables are cleaned, aligned and bound to labels respectively, and then unified output as sample objects that can be read by the model.

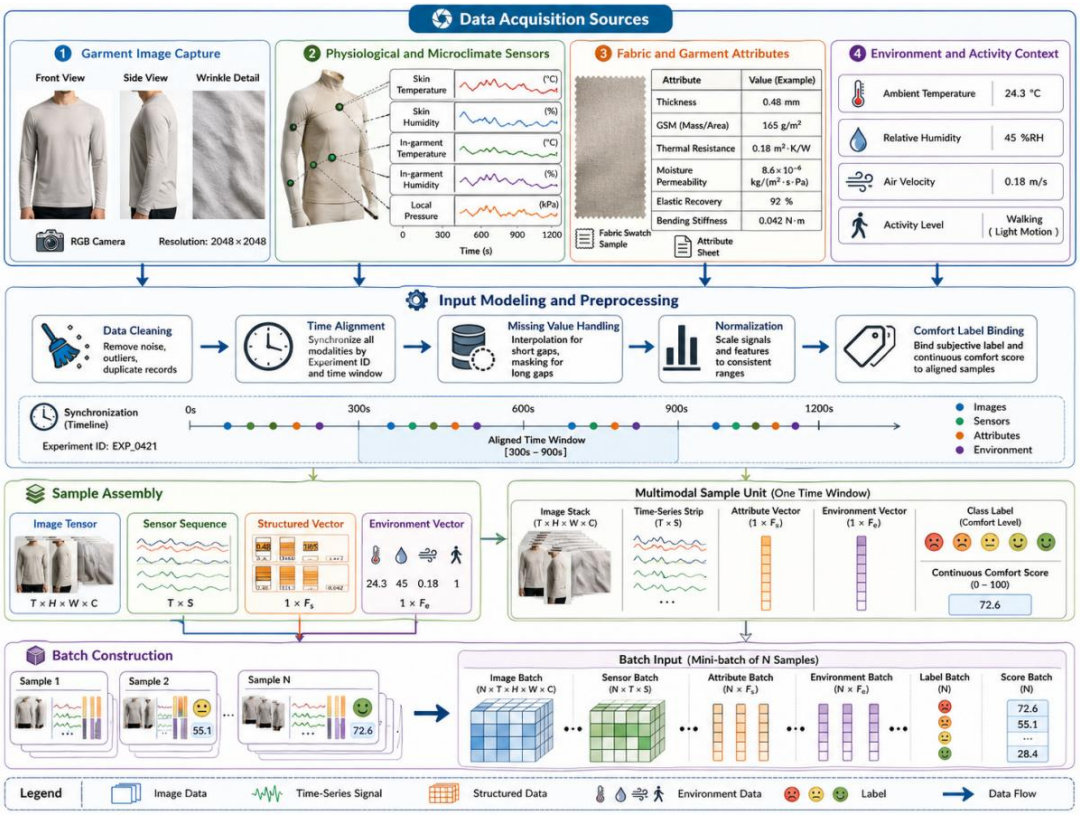

Figure 1: Flow chart of multimodal data acquisition and input modeling for clothing comfort

Through the above modeling process, clothing images, human sensing sequences, fabric structural parameters, and environmental variables are compressed into the same wearing experimental sample. The sample not only retains the independent information of each modality, but also provides a unified index for cross-modal feature fusion, so that the subsequent deep learning model can complete the joint prediction of classification and regression around the same comfort label.

2.2 Joint encoding of image, sensing and structural features

After the unified packaging of multi-modal samples, there are still differences in the expression forms of images, sensing sequences and structured variables. Clothing image belongs to two-dimensional spatial data, which can reflect contour, fold and fit state. The sensing sequence belongs to time continuous data, which can reflect the fluctuation of heat and humidity and the change of pressure during the wearing process. Fabric parameters and environmental variables belong to structured numerical data, which can describe material properties and external conditions. If the three types of data are directly concatenated into the model, it is easy to cause high-dimensional image features to suppress low-dimensional structured variables, or short-term sensing anomalies to be averaged out by the overall features [19]. In order to maintain the information independence of different modalities, the branching structure is adopted in the encoding stage, and the image, sensing and structured variables are respectively mapped to the latent feature space of the same dimension [20].

The image branch uses the convolutional feature extraction structure to share the garment front image, side image and local wrinkle image. The convolution layer extracts edges, textures, wrinkle densities, and body-fitting contours, the pooling layer compreszes spatial redundancy, and the normalization layer reduces the distribution shift caused by illumination and shooting distance. Let the clothing image of the ith sample be Ii, the parameter of the l-th convolution kernel be Wlimg, and the bias be blimg. The image coding result can be expressed as follows.

$$F_i^{img}=\mathrm{Pool}\left(\sigma\left(W_l^{img}*I_i+b_l^{img}\right)\right) \tag{3}$$

where, F_i^{img} represents the image feature vector of the ith sample; $\mathrm{Pool}(\cdot)$ represents the pooling operation. Let $\sigma(\cdot)$ denote the nonlinear activation function; $*$ denotes the convolution operation; W_l^{img} represents the weight of the image convolution kernel at the l-th layer; b_l^{img} represents the LTH bias term; Ii represents the input clothing image. This formula is used to convert the two-dimensional image into a compact feature vector, so that the local deformation and global fit information can enter the subsequent fusion network.

The sensing branch uses a bidirectional gated loop structure to process the body surface temperature, skin humidity, clothing temperature and humidity and local pressure in the time window. The forward hidden state records the change trend of the dressing process from the starting point to the end point, and the backward hidden state adds the explanation of the preceding sequence fluctuation in the subsequent segment, so that the accumulation of heat and humidity, the sudden increase of pressure and the change of activity obstruction can be preserved at the same time. Let the sensing sequence be Si={si,1,si,2,... si,T}, and its timing encoding result can be expressed as follows.

$$F_i^{sen}=\left[\vec{h}_{i,T};\overleftarrow{h}_{i,1}\right],\quad \vec{h}_{i,t}=\mathrm{GRU}_f(s_{i,t},\vec{h}_{i,t-1}),\ \overleftarrow{h}_{i,t}=\mathrm{GRU}_b(s_{i,t},\overleftarrow{h}_{i,t+1}) \tag{4}$$

where, F_i^{sen} represents the sensing timing feature vector of the ith sample. $\vec{h}_{i,t}$ denotes the

sensing input at the TTH sampling time; $\overleftarrow{h}_{i,t}$ denotes the forward hidden state; hi,t denotes the backward hidden state; GRUf(·) and GRUb(·) represent the forward and backward gated recurrent unit, respectively; [;] Representation vector stitching; T denotes the sampling window length. The formula is used to extract the dynamic change characteristics of the sensing sequence and avoid judging the comfort state only according to a single sampling point.

The structured branch concatenates the fabric parameters with the environment variables and inputs the multi-layer perceptron to complete the scale compression and nonlinear mapping. The variables such as thickness, gram weight, thermal resistance, moisture permeability, elastic response rate, and ambient temperature and humidity are standardized and entered into the fully connected network to output a structured representation consistent with the dimensions of the image branch and the sensing branch [21]. In order to avoid the three types of modalities still retaining different scales and different dimensions after coding, the coding results need to be further converted into alignable feature expressions. Figure 2 shows the output results of image feature maps, sensing timing features, and structured feature vectors from the perspective of feature morphology.

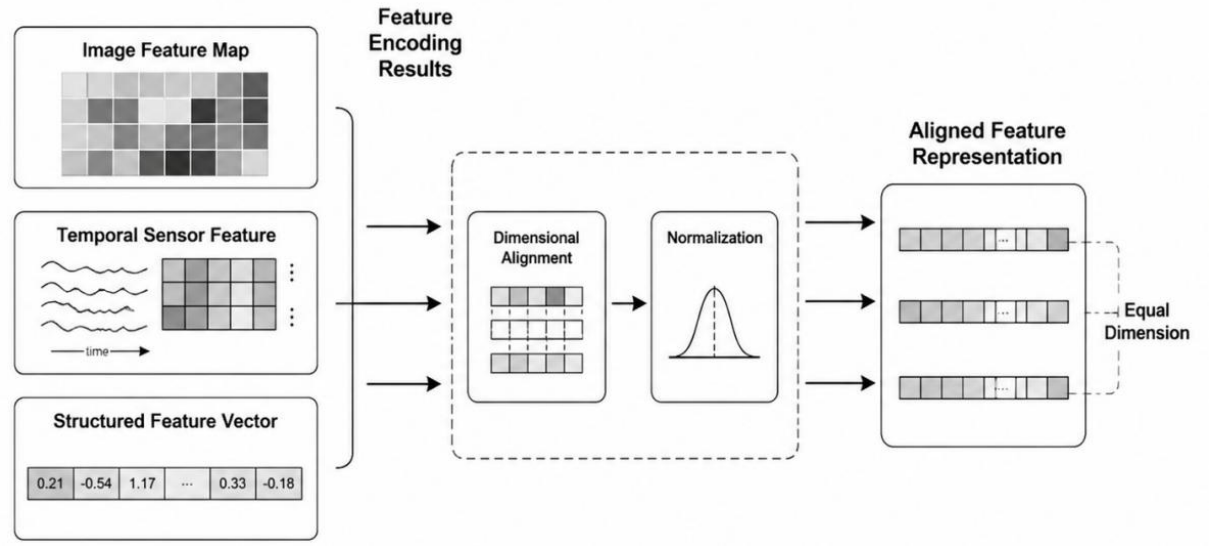

Figure 2: Schematic diagram of multi-modal feature coding results and representation dimensions

The result of image coding is a two-dimensional feature map, which mainly retains the garment contour, pleated texture and local fit differences. The sensing coding results show multi-channel temporal characteristics, which mainly retain the responses of body surface temperature, humidity, microclimate and local pressure over time. The structured coding results are represented as low-dimensional feature vectors, which mainly compress fabric attributes and environmental conditions. The three types of features are mapped by a unified dimension to form an equal-length representation, which not only retains the independent information of each modality, but also provides comparable input for subsequent cross-modal fusion.

2.3 Structure design of cross-modal attention fusion network

Image, sensing and structured variables have been transformed into computable feature vectors after branch coding, but the contribution of the three categories of features to comfort

prediction is not fixed. In the loose clothing sample, the image contour and fold distribution are more sensitive to the fit state judgment. In the experimental segments of high temperature and high humidity, the changes of body surface temperature, skin humidity and microclimate in clothing were more direct to the decrease of comfort. Under different fabric combinations, thermal resistance, moisture permeability and elastic recovery rate will affect the modification of comfort score by the model [22]. If the simple concatenation method is used, each mode can only enter the prediction layer in a fixed proportion, and it is difficult to dynamically adjust the weight according to the sample state. The model sets a cross-modal attention layer after joint coding results, so that different modalities complete weight allocation and information screening in the same prediction task [23].

Let the image coding feature of the ith sample be F_i^{img}, the sensing coding feature be F_i^{sen}, and the structured coding feature be F_i^{str}. Before fusion, the modal representation matrix is obtained by linear projection mapping to a unified dimensional space:

$$Z_i=\left[W_{img}F_i^{img};W_{sen}F_i^{sen};W_{str}F_i^{str}\right]+B_z \tag{5}$$

where, Zi represents the multi-modal projection feature matrix of the ith sample. W_{img},W_{sen},W_{str} represent the linear mapping matrices of image, sensing and structural features, respectively. F_i^{img}, F_i^{sen}, F_i^{str} represent three types of modal coding features, respectively. Bz represents the projection bias matrix; [;] Representations are concatenated by modal dimension.

In the unified feature space, the cross-modal attention layer calculates the modal correlation according to the current sample state. The model generates query matrix, key matrix and value matrix with the projected modal features, and calculates attention weights by scaling dot product:

$$A_i=\mathrm{Softmax}\left(\frac{(Z_iW_Q)(Z_iW_K)^T}{\sqrt{d_k}}\right) \tag{6}$$

where, Ai represents the cross-modal attention weight matrix of the ith sample; WQ represents the query mapping matrix; WK denotes the key mapping matrix; dk is the key vector dimension. Softmax(·) represents the normalization function; (·)T denotes the transpose of the matrix. This weight matrix is used to characterize the degree of correlation between different modes, so that the model can highlight the key modes and weaken the noise modes.

After obtaining the attention weights, the model applies the weight matrix to the value vector and makes residual connection with the original projection features to form the final fusion feature:

$$F_i^{fus}=\mathrm{LayerNorm}(A_iZ_iW_V+Z_i) \tag{7}$$

where, F_i^{fus} represents the fused feature of the ith sample; WV represents value mapping matrix; Ai represents cross-modal attention weight matrix; Zi represents the multi-modal projection feature matrix; LayerNorm(·) represents the layer normalization operation. The residual connection retains the original coding information of each modality, and the layer normalization reduces the feature fluctuation of different batches of samples.

The cross-modal fusion process not only includes feature stitching, but also needs to complete the shared space mapping, modal correlation judgment and dynamic weight adjustment. Table 2 is organized around the fusion configuration in the shared representation

space, which is used to illustrate the processing relationship of image, sensing, and structured features before and after entering the unified semantic representation.

Table 2: Multi-modal Shared Representation and Dynamic Fusion Configuration Table

Configuration Link	Processing Content	Function Description
Feature Mapping	Image, sensor, and structured features enter the shared representation space	Compress features from different sources into a unified semantic scale
Dimension Alignment	Perform unified dimension mapping on the three types of features	Eliminate the influence of modal dimension differences on fusion results
Correlation Calculation	Calculate the association strength among image, sensor, and structured features	Determine the effective contribution of different modalities in the current sample
Dynamic Weighting	Adjust the modal contribution ratio according to the correlation results	Improve key modal responses and weaken noise modal interference
Residual Retention	Retain the original modal encoding information	Avoid complete suppression of weak modal information during weighting
Fusion Output	Generate unified semantic fusion features	Provide input for comfort level classification and continuous score regression

The configuration links in the table need to be put into a unified feature space to observe, so as to present the relationship between different modalities from independent coding to dynamic fusion. Figure 3 takes the 3D shared representation space as the core, and organizes the mapping, correlation calculation, weight adjustment and fusion output of image features, sensing features and structured features into the same spatial structure.

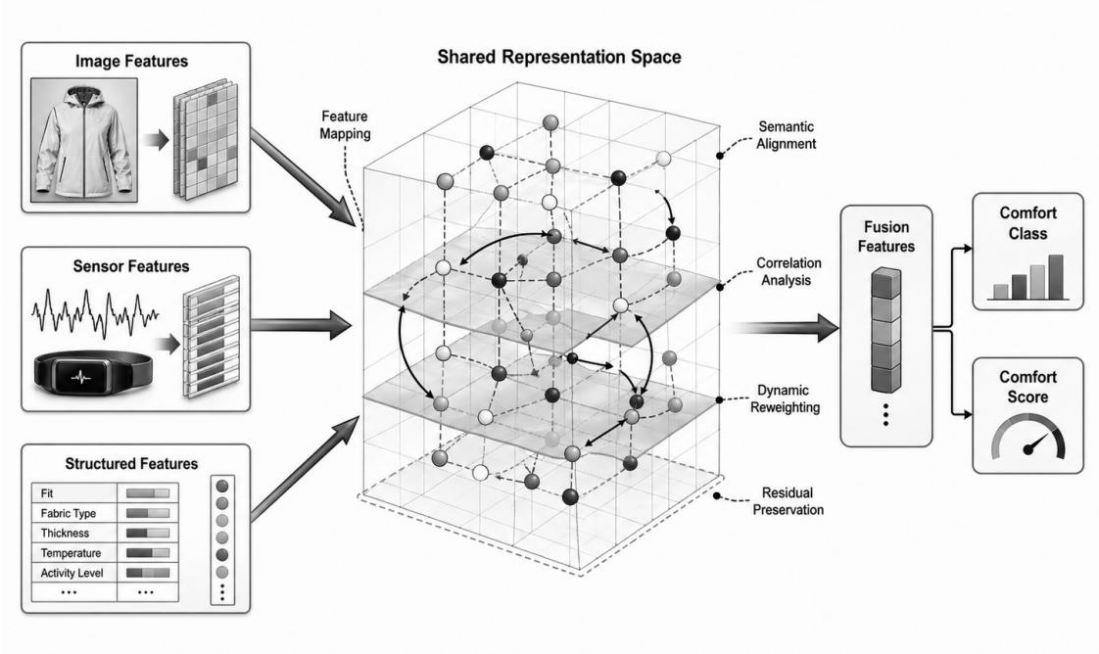

Figure 3: ultimodal shared representation space and dynamic fusion mechanism diagram

Image, sensing and structural features enter the shared representation space from different directions, and the three types of features are semantically aligned under a unified scale. The inter-modal correlation calculation is used to judge the effectiveness of different information sources in the current sample, and the dynamic weighting process adjusts the modal contribution according to the correlation results. The samples with obvious changes in heat, humidity and pressure are more dependent on sensing features, the samples with obvious differences in fit state or fold are more dependent on image features, and the samples with obvious differences in material properties need structural features to be corrected. The residual retention mechanism preserves the original encoded information, so that the fusion result can highlight the key modalities without completely losing the supplementary information in the weak modalities.

2.4 Comfort prediction output layer and objective function construction

After cross-modal attention fusion, the model obtains fusion features that include clothing appearance, human heat and humidity pressure changes, fabric structural parameters, and environmental conditions. The comfort prediction task should not only set a single classification output, because the wearing experience of clothing is not only represented by the grade difference of "uncomfortable, general, comfortable", but also represented by the slight change of continuous comfort score [24]. The output layer adopts the parallel structure of classification and regression, so that the model can give more fine-grained numerical results of comfort while judging the comfort level. The classification branch is used to output the probability distribution of different comfort levels, and the regression branch is used to output the normalized continuous comfort score, and the two types of results share the same fusion feature input.

Let the ith sample get the fused feature F_i^{fus} through the cross-modal attention network, the classification branch weight is Wc, the regression branch weight is Wr, and the corresponding bias terms are bc and br respectively. Then the comfort prediction output can be expressed as follows.

$$\hat{p}_i=\mathrm{Softmax}(W_c F_i^{fus}+b_c), \quad \hat{y}_i=\sigma(W_r F_i^{fus}+b_r) \tag{8}$$

where, $\hat{p}_i$represents the comfort level prediction probability of the ith sample; yi represents the predicted continuous comfort score; F_i^{fus} represents cross-modal fusion features; Wc and Wr represent the weight matrices of the classification and regression branches, respectively. bc and br represent the bias terms of the classification and regression branches, respectively. Softmax($\cdot$) represents the multi-class probability normalization function; $\sigma(\cdot)$ represents the Sigmoid activation function used to limit the prediction score to the interval [0,1].

The objective function is composed of classification loss, regression loss and regularization constraint. The classification loss constrains the model to correctly distinguish comfort levels, the regression loss constrains the continuous comfort scores to be close to the true labels, and the regularization term suppresses parameter overfitting. The joint loss function can be expressed as follows.

$$L=-\lambda_c \sum_{k=1}^{K} y_{i,k} \log(\hat{p}_{i,k})+\lambda_r(y_i-\hat{y}_i)^2+\lambda_w\|\Theta\|_2^2 \tag{9}$$

where, L represents the joint training loss; K represents the number of comfort level categories; yi,k denotes the true label of the ith sample at the KTH comfort level; $\hat{p}_{i,k}$ is the

predicted probability of the corresponding class; yi denotes the true continuous comfort score; $\hat{y}_i$ denotes the predicted continuous comfort score; Θ represents all the trainable parameters of the model; Let λc, λr, and λw denote the classification loss, regression loss, and regularization term weight, respectively. With joint goal constraints, the model is able to learn comfort level boundaries and continuous score changes simultaneously, avoiding focusing only on classification accuracy and ignoring subtle differences between adjacent levels.

3 Training and implementation of deep learning model for multimodal data fusion

3.1 Multi-modal sample tensor construction and batch input processing

Before model training, clothing images, sensing sequences, structured parameters and label data need to be uniformly packaged into batch tensors. The data form of each modality is different, the image retains channel, height and width, the sensing sequence retains time step and channel number, the fabric parameters and environment variables are merged into structured vectors, and the label data stores comfort level and continuous comfort score. The batch input is indexed by the experiment number, which ensures that different modalities in the same batch still correspond to the same wearing experimental sample, and avoids mismatching between images, sensing records and labels.

Let the BTH training batch contain Nb samples and the batch input tensor be denoted as follows.

$$B_b=\{I_b,S_b,U_b,Y_b\},\quad I_b\in R^{N_b\times C_i\times H\times W},\ S_b\in R^{N_b\times T\times C_s},\ U_b\in R^{N_b\times(C_p+C_e)} \tag{10}$$

where, Bb represents the BTH training batch; Ib stands for batch image tensor; Sb stands for batch sensing timing tensor; Ub represents the concatenation tensor of structural parameters and environment variables. Yb stands for batch label; Nb represents the number of samples in the batch; Ci denotes the number of image channels; H and W denote the height and width of the image; T represents the sensing window length; Cs represents the number of sensing channels; Cp and Ce represent the fabric parameter dimension and the environment variable dimension, respectively.

In order to reduce the impact of dimensional differences and missing values on training, sensory and structured variables are standardized before input, and masks are used to record the valid status of the variables:

$$\tilde{x}_{i,j}=m_{i,j}\cdot\frac{x_{i,j}-\mu_j}{\sigma_j+\varepsilon},\quad m_{i,j}\in\{0,1\} \tag{11}$$

where, $\tilde{x}_{i,j}$represent the normalized variable values; xi,j denote the original variable values; Let μj and σj denote the mean and standard deviation of the JTH variable in the training set, respectively. Let ε denote the smoothing term; mi,j denote the missing mask. Table 3 organizes the tensor dimensions and preprocessing methods of various training inputs.

Table 3: Dimension of Multimodal Sample Tensor and Preprocessing Configuration Table

Data Type	Tensor Dimension	Preprocessing Method
Clothing Image	Batch size × 3 × 224 × 224	Size scaling, normalization, random cropping
Sensor Sequence	Batch size × 120 × 8	Window segmentation, outlier smoothing, standardization
Structured Parameters	Batch size × 10	Missing value imputation, mean-variance standardization
Environmental Variables	Batch size × 4	Normalization, experimental condition encoding
Label Data	Batch size × 2	Level encoding, continuous score normalization

Before multimodal samples are input into the model, they need to undergo size standardization, window extraction, mask generation and batch packaging. Figure 4 organizes the above processing steps into the same input process.

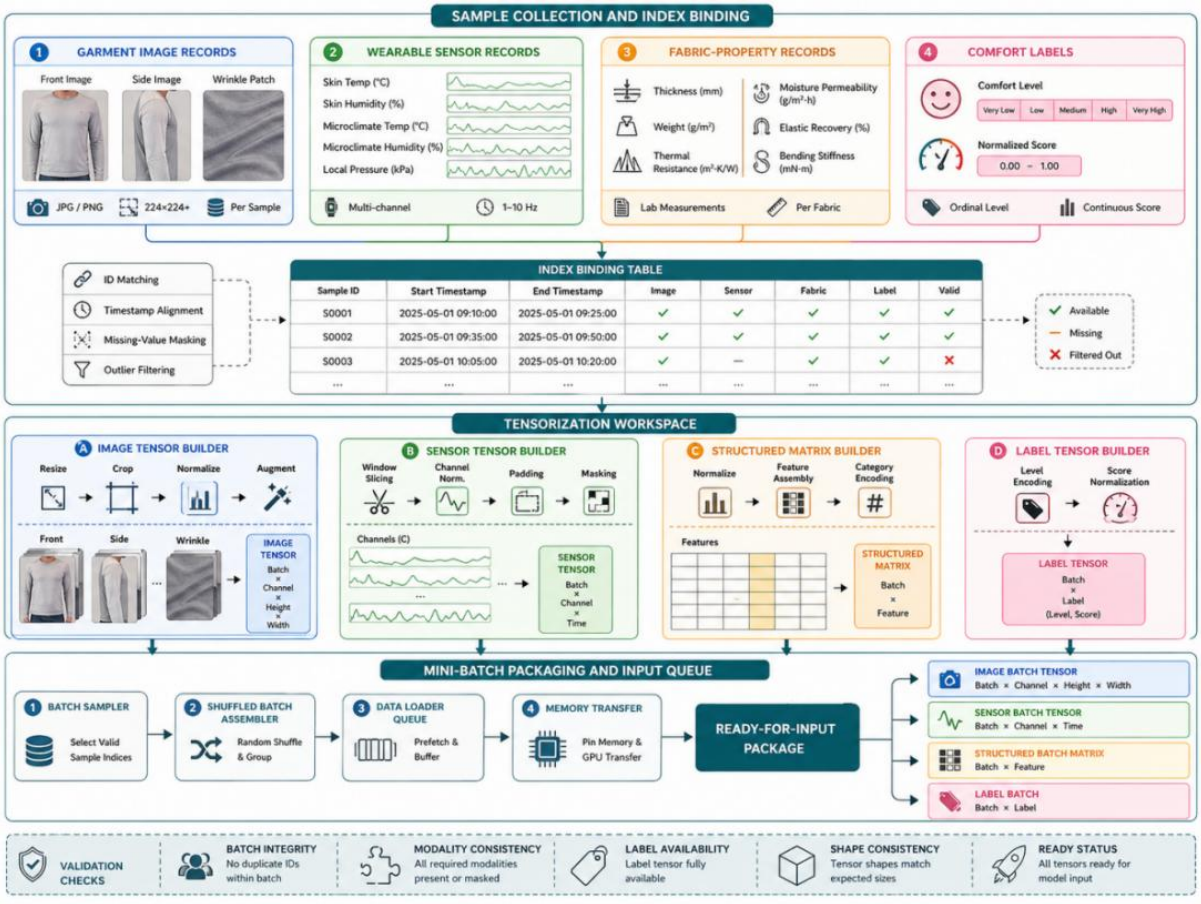

Figure 4: Flow chart of multi-modal sample tensor construction and batch input

After batch packaging, image tensors, sensing tensors, structured tensors, and label tensors can be input into the model synchronously to support subsequent end-to-end training.

3.2 End-to-end training process of fusion prediction model

After the batch tensors enter the model, the training process is executed in the order of "branch encoding-attention fusing - joint prediction - loss propagation". Firstly, the image tensor enters the convolutional encoding branch to extract the garment contour, fold and fit state. The sensing tensor enters the bidirectional gated cyclic branch to extract the heat and humidity pressure changes. The structured tensor enters the fully connected branch to extract fabric properties and environmental conditions [25]. The three types of features are fused by the cross-modal attention layer, and then the comfort level probability and continuous comfort score are output by the classification branch and regression branch, respectively.

Let the input of the BTH training batch be Bb, the encoding network parameters be Θe, the fusion network parameters be Θf, and the prediction layer parameters be Θo. The end-to-end forward propagation process can be expressed as follows.

$$(\hat{P}_b,\hat{Y}_b)=G_{\Theta_o}\left(M_{\Theta_f}\left(E_{\Theta_e}(I_b,S_b,U_b)\right)\right) \tag{12}$$

where, $\hat{P}_b$ represents the batch comfort level prediction probability; $\hat{Y}_b$ represents the predicted value of batch continuous comfort score; $E\Theta e(\cdot)$ denotes the multi-branch coding network; $M\Theta f(\cdot)$ represents the cross-modal attention fusion network; $G\Theta o(\cdot)$ represents the classification and regression joint prediction layer; Ib, Sb, Ub denote the bulk image tensor, sensing tensor, and structured tensor, respectively.

The training loss is calculated on the batch dimension and consists of a classification loss, a regression loss, and a parametric regularization term. The average training loss for the BTH batch can be expressed as follows.

$$L_b=\frac{1}{N_b}\sum_{i=1}^{N_b}\left[-\lambda_c\sum_{k=1}^{K}y_{i,k}\log(\hat{p}_{i,k})+\lambda_r(y_i-\hat{y}_i)^2\right]+\lambda_w\|\Theta\|_2^2 \tag{13}$$

where Lb represents the average training loss of the BTH batch; Nb represents the number of samples in the batch; K represents the number of comfort level categories; yi,k denotes the true rank label; $\hat{p}_{i,k}$ denotes the predicted grade probability; yi denotes the true continuous comfort score; yi denotes the predicted continuous comfort score; Θ represents all trainable parameters; λc, λr, λw denote the classification, regression, and regularization term weights, respectively.

To present the execution relationship between modules in end-to-end training, Figure 5 organizes batch input, feature encoding, cross-modal fusion, joint prediction, loss calculation and backpropagation into the same training link.

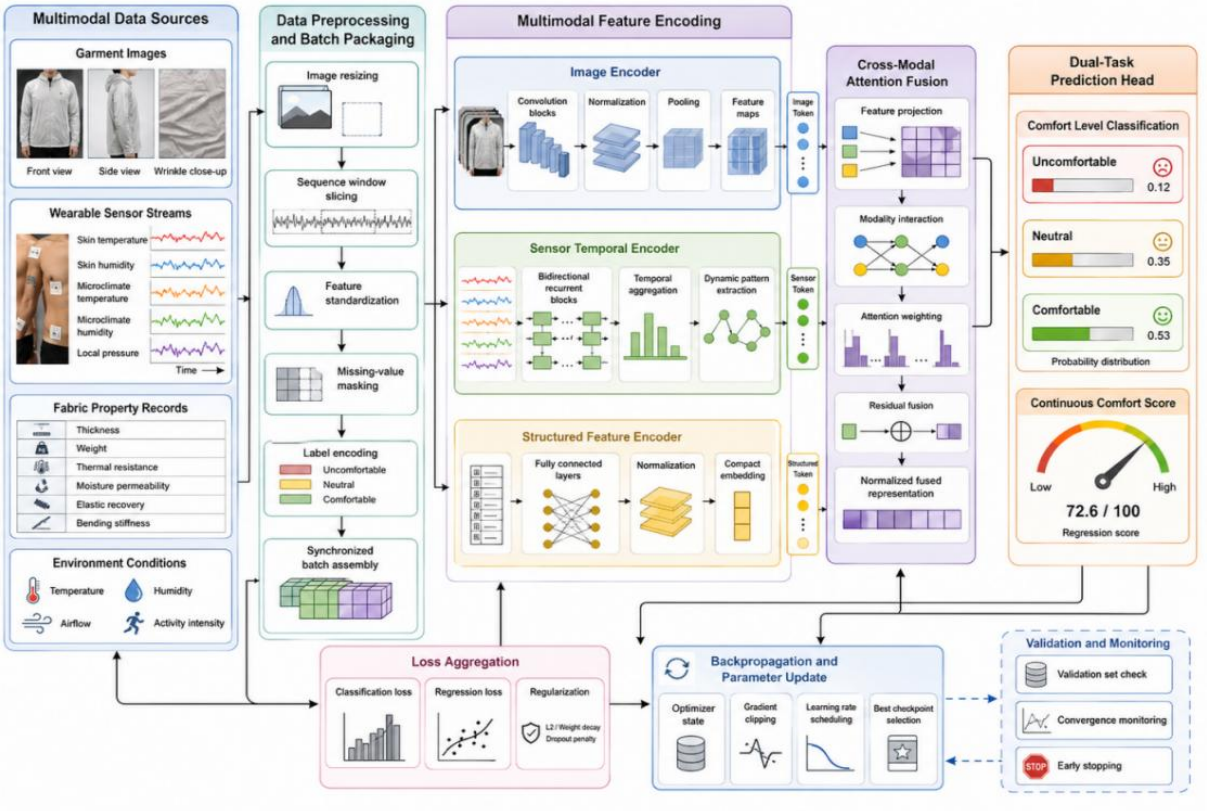

Figure 5: End-to-end training flow diagram of multimodal fusion deep learning model

During the training process, the loss value is passed back from the output to the prediction layer, the fusion layer and the coding branch, so that the image, sensing and structural feature extraction processes can be updated synchronously around the same comfort prediction target.

3.3 Parameter update and convergence control under loss function constraints

The fusion prediction model includes a convolution branch, a gated recurrent branch, an attention fusion layer, and a classification and regression output layer with a large parameter scale. If the training process only uses a fixed learning rate, it is easy to have problems such as fusion weight oscillation, sensing branch gradient fluctuation and validation set loss repeatedly rising. In the training phase, AdamW optimizer, learning rate dynamic attenuation and gradient clipping are used to jointly constrain parameter update, so that the model can maintain the convergence speed while reducing the risk of overfitting and gradient anomaly.

In order to make the parameter update consider gradient direction, adaptive step size and weight decay simultaneously, the AdamW update process is expressed as follows.

$$\Theta_{t+1}=\Theta_t-\eta_t\left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\varepsilon}+\omega\Theta_t\right) \tag{14}$$

where, $\Theta t+1$ represents the updated model parameters; Θt denotes the current model parameters; ηt represents the current iteration learning rate; $\hat{m}_t$ represents the bias-corrected first moment estimate; $\hat{v}_t$ represents the bias-corrected second moment estimate; Let ε denote the smoothing term; Let ω denote the weight decay coefficient. This update method combines adaptive gradient adjustment with parameter regularization constraints, which can reduce the

risk of overfitting in high-dimensional fusion layers.

The learning rate control adopts the combination of warmup and cosine attenuation. A small step size at the beginning of training can avoid excessive gradient in the stage of random initialization, and gradually reducing the learning rate in the stable stage can refine the parameter search range, which is calculated as follows.

$$\eta_t=\eta_{min}+\frac{1}{2}(\eta_{max}-\eta_{min})\left(1+\cos\frac{\pi(t-T_w)}{T_{max}-T_w}\right) \tag{15}$$

where, ηt represents the current iteration learning rate; ηmax represents the maximum learning rate; ηmin represents the minimum learning rate; Tw denotes the number of warmup iteration steps. Tmax is the maximum number of training iterations. t denotes the current iteration step. Through this scheduling method, the model can start smoothly in the early stage and continue to be optimized with a small step size in the later stage.

In order to prevent the gradient explosion in the back propagation of the sensing timing branch and the attention layer, the gradient norm is trimmed during the training process:

$$\tilde{g}_t=g_t\cdot\min\left(1,\frac{\tau}{\|g_t\|_2}\right) \tag{16}$$

where, $\tilde{g}_t$represents the gradient after clipping; gt represents the original gradient; Let τ denote the maximum gradient norm threshold; ‖gt‖2 denotes the two-norm of the original gradient. If the gradient norm exceeds the threshold, the gradient will be scaled; If the threshold is not exceeded, the original gradient is kept unchanged. Therefore, the interference of abnormal batches on the parameter update direction can be reduced, and the multi-modal fusion model can obtain a more stable convergence process.

3.4 Model Hyperparameter Configuration and Training Stability Control

The stability of model training is not only dependent on the loss function and optimizer, but also influenced by batch size, hidden layer dimensions, number of attention heads, Dropout ratio, number of training epochs, and loss weights. Configuring the hyperparameters too small will limit the feature expression ability, resulting in insufficient learning of the correlation between image textures, sensor fluctuations, and structured variables; while configuring them too large will increase the risk of overfitting and cause fluctuations in the validation set loss in the later stages. The training process integrates model capacity, regularization constraints, and feedback from the validation set into a unified control, enabling the multimodal fusion network to maintain stable updates across different batches of samples.

To comprehensively observe the model training status, a training stability constraint indicator is constructed to simultaneously measure the extent of loss reduction, fluctuations in the validation set, and the intensity of gradient changes:

$$\Phi_t=\rho_1|L_t-L_{t-1}|+\rho_2|V_t-V_{t-1}|+\rho_3\frac{\|g_t-g_{t-1}\|_2}{\|g_{t-1}\|_2+\varepsilon} \tag{17}$$

where, Φt represents the stability constraint index in the TTH round of training. Lt represents the training loss in round t. Lt−1 represents the training loss in the previous round; Vt represents the validation loss at round t; Vt−1 represents the validation loss in the previous round; gt is the t-th gradient vector. gt−1 denotes the gradient vector of the previous round; Let ε denote the smoothing term; ρ1, ρ2 and ρ3 denote the stability index weights. If the value

of this index is too high, it means that there may be loss shock or gradient mutation in the training process, and it is necessary to reduce the learning rate, increase the Dropout ratio, or stop the training early.

The model hyperparameters are set around the input scale, network capacity, regularization constraints, and training process, and Table 4 organizes the main configurations.

Table 4: Model Parameters and Training Configuration Table

Parameter Name	Parameter Setting	Function Description
Image Size	224 × 224	Unify the input scale of clothing images
Sensor Window	120	Control the time window length of sensor sequences
Hidden Dimension	128	Control the representation dimension of encoded features
Attention Heads	4	Provide multi-head cross-modal correlation calculation
Dropout Rate	0.3	Suppress overfitting in the fusion layer and output layer
Batch Size	64	Balance training efficiency and gradient stability
Learning Rate	0.001	Control the initial step size of parameter update
Weight Decay	0.0005	Limit excessive expansion of model parameters
Max Epoch	150	Set the maximum number of training epochs
Early Stop Patience	15	Stop training when there is no improvement on the validation set

During the training process, if the validation loss does not decrease for multiple consecutive rounds, the early stopping mechanism is triggered and the model parameters with the best performance in the validation set are saved. Dropout and weight decay jointly limit the parameter inflation of the high-dimensional fusion layer, and the batch size matches the learning rate to avoid excessive gradient noise in small batches. Through the joint control of hyperparameter configuration and stability index, the model training process can maintain a balance between feature expression ability and generalization ability.

4 Application verification of clothing comfort prediction

4.1 Experimental data set, operating environment and evaluation index setting

In order to verify the effectiveness of multimodal data fusion deep learning model in clothing comfort prediction task, the GarmentComfort-MM multimodal clothing comfort dataset is constructed. The dataset contains 7200 groups of wearing experiment samples from 60 subjects and 12 types of clothing fabric and style combinations. Each group of samples contains front image, side image, local fold image of clothing, surface temperature, skin humidity, internal temperature and humidity of clothing, shoulder pressure, waist pressure, elbow and knee pressure, and structural parameters such as fabric thickness, gram weight, thermal resistance, moisture permeability, elastic recovery rate, and bending stiffness. The experimental environment is set to four temperature conditions and three humidity conditions, the temperature is 18 ℃, 22 ℃, 26 ℃ and 30 ℃ respectively, and the relative humidity is 40%, 60% and 80% respectively, which is used to cover the low temperature, normal

temperature, high temperature and high humidity wearing states.

The sample label consists of both comfort level and continuous comfort score. The comfort level was divided into three categories: "uncomfortable, general, and comfortable", and the continuous scores were normalized to the 0-1 range. The dataset is divided into training set, validation set and test set according to the proportion of 70%, 15% and 15%, and the corresponding sample numbers are 5040 groups, 1080 groups and 1080 groups, respectively. The training set is used for model parameter learning, the validation set is used for hyperparameter tuning and early stop judgment, and the test set is only used for final performance evaluation. The evaluation metrics cover both classification prediction and regression prediction, where accuracy, precision, recall and F1 score are used for classification task, and mean absolute error and root mean square error are used for regression task. In order to compare the improvement of multi-modal fusion with single-modal or simple splicing model, the fusion gain index is further set.

The classification performance index is used to evaluate the recognition ability of the model for different comfort levels, which is calculated as follows.

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}},\quad \text{F1}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{18}$$

where, Accuracy represents the accuracy rate; F1 is the F1 value. Precision is the accuracy rate. Recall is the recall rate. TP is the number of samples correctly predicted as the target class. TN is the number of samples correctly predicted as non-target class. FP is the number of samples incorrectly predicted as the target class. FN represents the number of samples incorrectly predicted as non-target class.

Continuous comfort score prediction is evaluated by mean absolute error and root mean square error, which are calculated as follows.

$$\text{MAE}=\frac{1}{\text{N}}\sum_{\text{i}=1}^{\text{N}}|\text{y}_\text{i}-\hat{\text{y}}_\text{i}|,\quad \text{RMSE}=\sqrt{\frac{1}{\text{N}}\sum_{\text{i}=1}^{\text{N}}(\text{y}_\text{i}-\hat{\text{y}}_\text{i})^2} \tag{19}$$

where, MAE stands for mean absolute error; RMSE stands for root mean square error. N is the number of test samples. yi denotes the true comfort score of the ith sample; $\hat{y}_i$ denotes the predicted comfort score of the ith sample.

The fusion gain is used to measure the performance improvement of the full multimodal model over the baseline model and is calculated as follows.

$$\text{Gain}=\frac{\text{M}_{\text{ours}}-\text{M}_{\text{base}}}{\text{M}_{\text{base}}}\times 100\% \tag{20}$$

where, Gain represents the fusion gain; Mours represents the evaluation index value of the complete multimodal fusion model. Mbase represents the evaluation index value of the baseline model. A higher value in the classification index indicates a better performance, while a lower value in the error index indicates a better performance. When calculating the gain of the error class, the error decrease amplitude is used to convert.

The experimental configuration needs to simultaneously cover the dataset scale, the software and hardware environment, and the evaluation metrics; Table 5 organizes the main Settings.

Table 5: Configuration Table of Experimental Data Set, Operating Environment and Evaluation Indicators

Category	Project	Configuration Value
Dataset	Dataset Name	GarmentComfort-MM
	Total Samples	7200
	Subjects	60
	Garment Types	12
	Environment Temperature	18℃, 22℃, 26℃, 30℃
	Relative Humidity	40%, 60%, 80%
	Split Ratio	70% / 15% / 15%
Data Modality	Image Data	Front Image, Side Image, Wrinkle Detail Image
	Sensor Data	Skin Temperature, Skin Humidity, Microclimate Temperature and Humidity, Local Pressure
	Structured Data	Thickness, Weight, Thermal Resistance, Moisture Permeability, Elastic Recovery, Bending Stiffness
Hardware	CPU	Intel Xeon Silver 4314 × 2
	GPU	NVIDIA RTX 4090 24GB
	Memory	128 GB
Software	Operating System	Ubuntu 22.04
	Framework	PyTorch 2.2
Training	Batch Size	64
	Max Epoch	150
Evaluation	Metrics	Accuracy, Precision, Recall, F1-score, MAE, RMSE, Fusion Gain

The above setup enables the testing process to test both the rank classification ability and the continuous score prediction ability of the model. The image, sensing, structural parameters and environmental variables in the dataset were all bound to the same comfort label. The subsequent comparison experiment and ablation experiment were completed under the same division ratio, the same hardware environment and the same evaluation index to ensure the comparability of the results of different models.

4.2 Comparison model and ablation experimental design

In order to avoid the experimental verification only staying in the result display of a single model, 11 groups of experimental models are set up in the test phase, including 6 groups of contrast models, 4 groups of ablation models and 1 group of complete models. All models were partitioned in the same way as the GarmentComfort-MM dataset, with 5040, 1080 and 1080 groups for training, validation and test sets, respectively. The traditional machine learning model inputted 14-dimensional structured variables, CNN inputted a clothing image of size $3 \times 224 \times 224$, and BiGRU inputted a sensing sequence of size 120×8. The fusion class model received image, sensing and structured variables at the same time. The training rounds of the deep learning model are uniformly set to 150, the Batch Size is 64, and the initial learning rate is 0.001. The number of parameters of the complete model is controlled at 3.86M, Transformer Fusion is about 4.42M, and CNN+BiGRU Concat is about 3.21M.

The ablation experiment takes the full model as the baseline, and removes the cross-modal attention layer, image input branch, sensing input branch, and structured input branch, respectively. When the attention layer is removed, the three types of features are directly

concatenated into the prediction layer. When the image branches are removed, only the sensing sequence and structured variables are retained. When the sensing branch is removed, only the image and structured variables are retained. When the structured branches are removed, only the image and sensing sequence are retained. To quantify the performance change after module removal, the performance degradation rate is used to calculate:

$$\text{Drop}=\frac{M_{full}-M_{abl}}{M_{full}}\times 100\% \tag{21}$$

where, Drop represents the performance degradation rate of the ablation model relative to the complete model; Mfull represents the evaluation index value of the complete multimodal fusion model. Mabl represents the evaluation index value of the ablation model. For positive metrics such as Accuracy and F1-score, a decrease in value indicates a decrease in performance. For error metrics such as MAE and RMSE, the error increase is used to represent the performance degradation degree. Table 6 organizes the comparison models and ablation models participating in the experiment to illustrate the model setting and the verification purpose, respectively.

Table 6: Comparison Model and Ablation Experiment Settings Table

Model Name	Model Setting	Verification Purpose
SVM	14-dimensional structured variables, RBF kernel function	Test the basic prediction ability of the traditional classification model
Random Forest	14-dimensional structured variables, 100 decision trees	Test the modeling ability of the ensemble model for material parameters
CNN	Clothing images, input size 3 × 224 × 224	Verify the contribution of the image modality
BiGRU	Sensor sequence, input size 120 × 8	Verify the contribution of thermal-humidity pressure time-series information
CNN+BiGRU Concat	Direct concatenation of image features and sensor features	Test the effect of simple concatenation fusion
Transformer Fusion	Image, sensor, and structured variables are input into the Transformer fusion layer	Test the performance of general attention fusion
Ours-w/o-Attn	Remove the cross-modal attention layer	Verify the role of dynamic modal weight allocation
Ours-w/o-Img	Remove the image input branch	Verify the contribution of clothing appearance features
Ours-w/o-Sensor	Remove the sensor input branch	Verify the contribution of sensor dynamic information
Ours-w/o-Struct	Remove the structured input branch	Verify the contribution of fabric parameters and environmental variables
Ours	Complete multimodal fusion prediction model	Serve as the final comprehensive comparison model

All the models used the same test set and evaluation metrics. The subsequent performance differences mainly stemmed from the changes in the input modal combination and fusion

structure.

4.3 Prediction accuracy, fusion effect and error distribution analysis

The test set results show that the difference between different models in comfort level recognition is obvious. SVM and Random Forest only rely on structured variables, with the Accuracy of 78.4% and 81.3%, respectively, and the F1-score of 77.2% and 80.2%, respectively, indicating that fabric parameters and environmental variables can provide basic judgment information, but it is difficult to express the state of clothing fit and the change of heat and humidity pressure during wearing. The Accuracy of CNN model is 84.6%, and the F1-score is 83.5%. The Accuracy of the BiGRU model is improved to 86.2%, and the F1-score is 85.3%, indicating that the sensing sequence has a more direct role in discriminating comfort changes. The CNN+BiGRU Concat model achieves 88.7% Accuracy and 87.8% F1-score, and the Transformer Fusion model achieves 90.5% Accuracy and 89.7% F1-score. The complete model achieves 93.8% Accuracy, 93.4% Precision, 92.9% Recall and 93.1% F1-score, which is the highest classification performance.

In order to present the differences in the indicators of different models in the comfort level recognition task, Figure 6 compares the four classification indicators of Accuracy, Precision, Recall and F1-score.

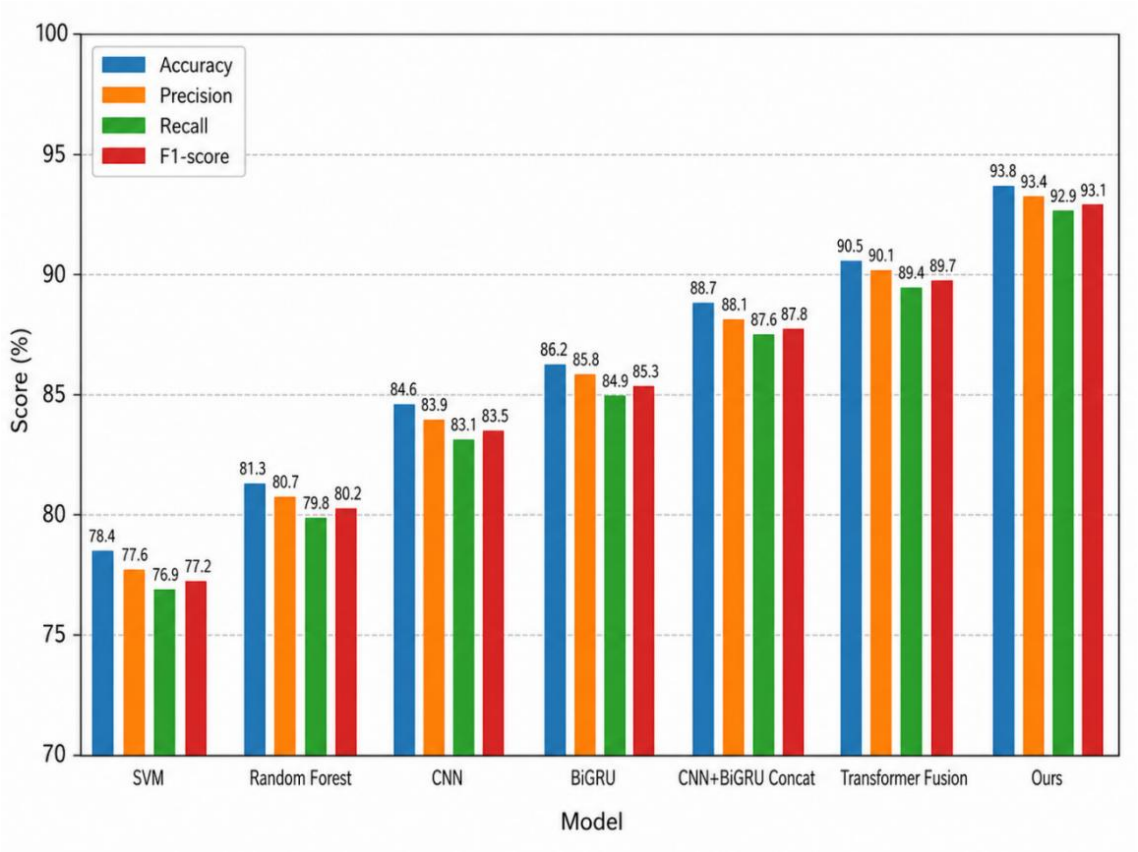

Figure 6: Comparison of comfort prediction classification performance of different models

From the classification results, compared with the CNN+BiGRU Concat model, the complete model improves the Accuracy, Precision, Recall and F1-score by 5.1, 5.3, 5.3 and 5.3 percentage points respectively. Compared with the Transformer Fusion model, it increases by 3.3, 3.3, 3.5 and 3.4 percentage points respectively. The results show that the cross-modal attention layer can adjust the modal contribution according to different sample states, so that

the high temperature and high humidity samples are more dependent on sensing information, the samples with loose or obvious body-fitting differences make more use of image features, and the samples with obvious material differences retain the structural parameter correction ability.

The results of ablation experiments further illustrate the role of each input branch and fusion structure. After removing the cross-modal attention layer, the Accuracy is reduced by 3.4 percentage points, the F1-score is reduced by 3.7 percentage points, the MAE is increased by 12.6%, and the RMSE is increased by 10.8%. After removing the image branches, the Accuracy decreases by 2.1 percentage points and the F1-score decreases by 2.4 percentage points. After removing the sensing branch, the Accuracy decreased by 4.6 percentage points, the F1-score decreased by 4.9 percentage points, the MAE increased by 15.4%, and the RMSE increased by 13.2%, which was the ablation group with the most obvious performance degradation. After removing the structured branches, the Accuracy decreased by 1.8 percentage points, and the F1-score decreased by 2.0 percentage points, indicating that material properties and environmental conditions still have a correction effect on the judgment of adjacent comfort levels.

To further compare the differences in the original index values between the full model and each ablation model, Figure 7 compares the Accuracy, F1-score, MAE and RMSE in parallel.

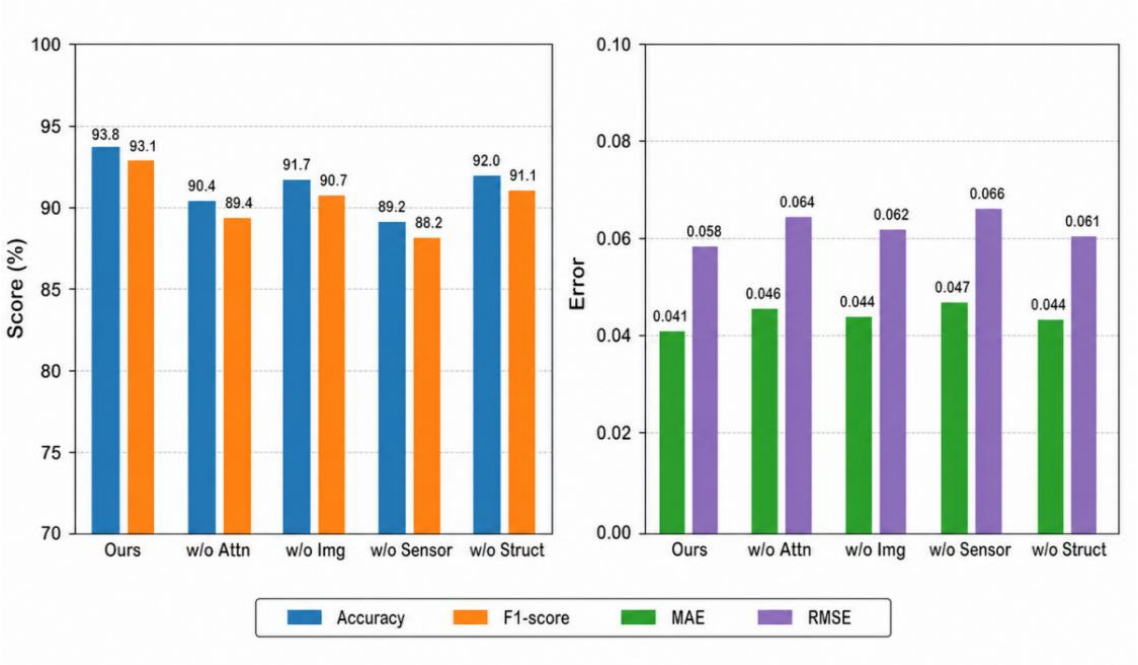

Figure 7: Comparison Chart of Key Indicators between the Complete Model and the Ablation Model

The complete model maintains the highest level in terms of Accuracy and F1-score, and the lowest error in MAE and RMSE. The performance degradation after removing the sensing branch is the most significant, with the largest decline in Accuracy and F1-score, and the most prominent increase in MAE and RMSE. The change in indicators after removing the attention layer is second only to the sensing branch, indicating that the dynamic thermal-humid pressure information and cross-modal weighting mechanism have a strong supporting effect on comfort prediction. After removing the image branch and the structured branch, although the model performance has declined, the overall degradation amplitude is relatively smaller, indicating that these two modalities mainly undertake auxiliary discrimination and error

correction functions. To compare the performance degradation amplitudes after removing different modules, Figure 8 organizes the decline in Accuracy, the decline in F1-score, the increase in MAE, and the increase in RMSE into an ablation result matrix.

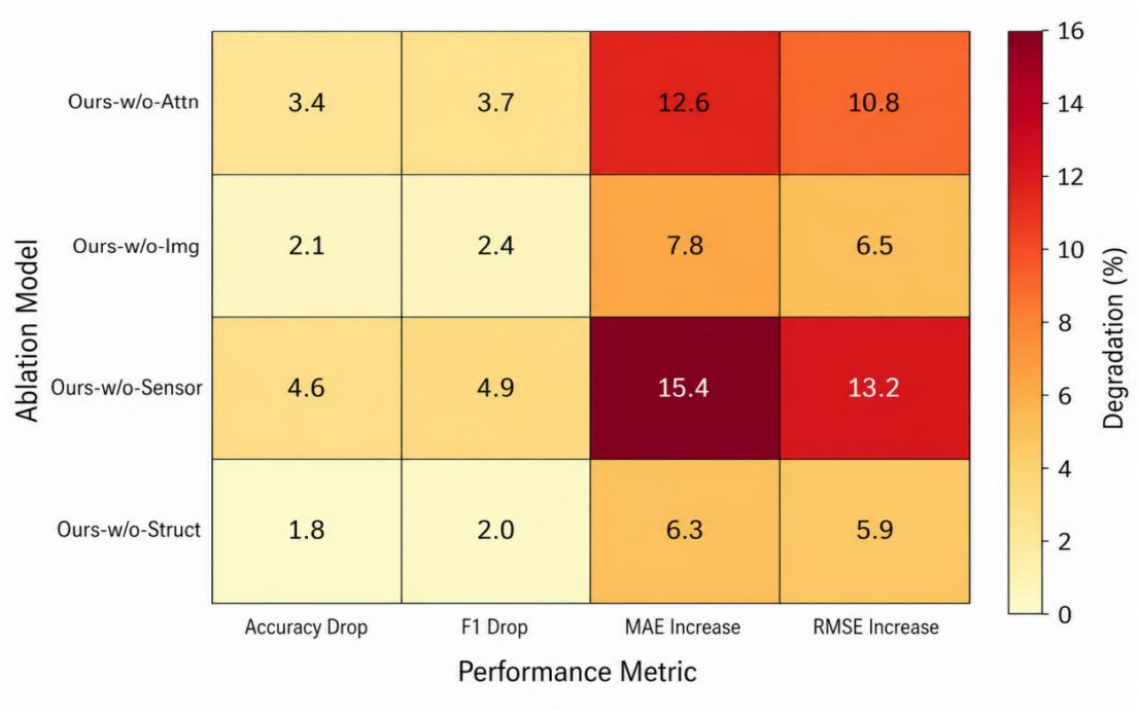

Figure 8: Heat map of ablation model prediction performance degradation

Among the ablation results, the degradation corresponding to the sensing branch is the largest, which indicates that body surface temperature, skin humidity, in-garment microclimate, and local pressure are key dynamic signals in comfort prediction. The degradation of the attention layer is second only to the sensing branch, indicating that the fusion structure itself has a significant impact on the performance of the model. The image branch mainly supplements the appearance and fit status of clothing, and the structured branch mainly supplements the material properties and environmental conditions. The two types of information together reduce the misjudgment between adjacent comfort levels.

In the continuous comfort score prediction, the MAE and RMSE of the full model were 0.041 and 0.058, respectively. The prediction error of most samples is concentrated between 0.02 and 0.07, and the samples with error more than 0.10 are mainly distributed around the "general" level. The mean absolute error is 0.046 for the "uncomfortable" sample, 0.052 for the "normal" sample, and 0.036 for the "comfortable" sample. The error of the middle grade is higher, which is mainly related to the lack of clear boundary of the subjective score. The label boundaries of the samples at the two ends of "comfortable" and "uncomfortable" are clearer, and the prediction errors are relatively more stable.

In order to present the fitting effect of the continuous comfort score and the error distribution of different grade samples, Figure 9 puts the true score, predicted score, grade error and typical sample results into the same effect display.

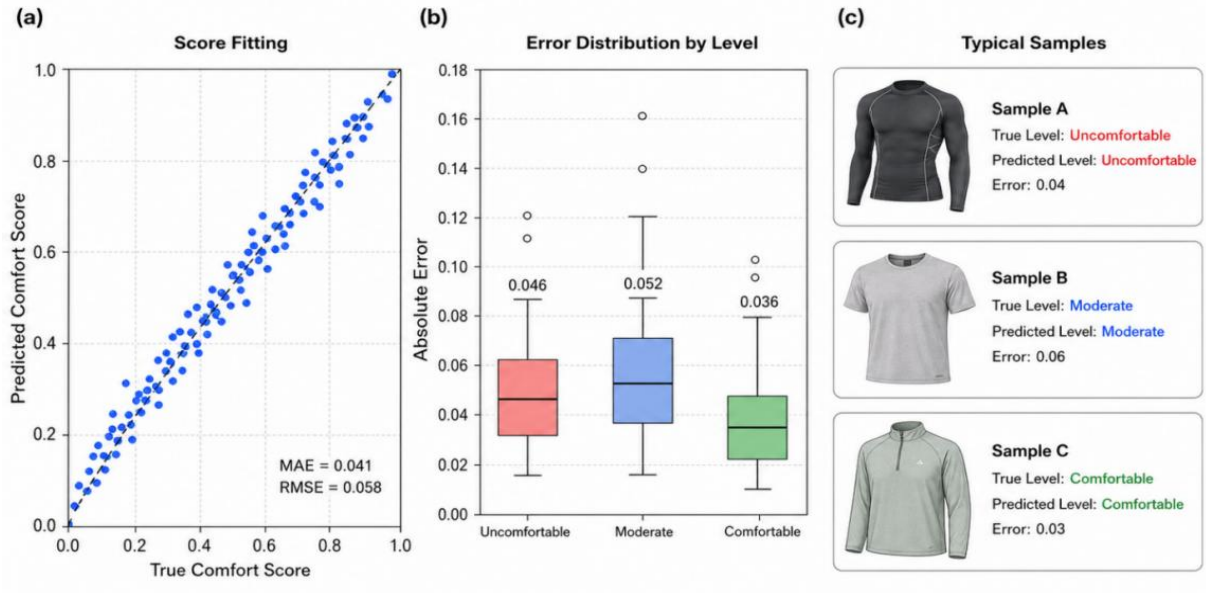

Figure 9: Garment comfort prediction error distribution and sample renderings

From the error distribution, the predicted score is close to the true score as a whole, and the point cloud is mainly distributed near the diagonal. Among the samples of different grades, the box range of the "general" category is wider, and the dispersion degree is higher than that of the "uncomfortable" and "comfortable" categories, which indicates that the adjacent grade boundaries are still difficult to predict. In the typical sample results, the true level of sample A is uncomfortable, the predicted level is uncomfortable, and the error is 0.04. For sample B, the true grade is general, the prediction grade is general, and the error is 0.06. For sample C, the true level is comfortable and the predicted level is comfortable with an error of 0.03. The overall results show that the model can simultaneously complete the comfort level recognition and continuous comfort score prediction.

4.4 Discussion

From the comparison results, the complete model is superior to the traditional machine learning model, the single-modal deep learning model and the simple stitching fusion model, the main reason is that the input information covers three key factors: clothing appearance, human dynamic feedback and material environmental conditions. SVM and Random Forest can only use structured variables, which are difficult to identify wrinkles, fit states, and changes in heat and humidity pressure during wearing. CNN can extract clothing image features, but it lacks continuous wearing feedback. BiGRU can model the sensing sequence, but it cannot use the fabric structure and appearance information. Although simple stitching introduces multi-modal input, the weight of each modality is fixed, which is difficult to adapt to different sample states. The cross-modal attention layer can dynamically adjust the contribution of image, sensing and structured variables according to the sample characteristics, so that the model can maintain a relatively stable prediction ability under different wearing conditions.

Among the ablation results, the performance degradation after removing the sensing branch is the most obvious, which indicates that body surface temperature, skin humidity, in-garment microclimate, and local pressure are the core dynamic information in comfort prediction. The performance also deteriorates significantly after removing the attention layer, indicating that the multi-modal features are not simple superposition, and the weight

allocation between modalities will directly affect the prediction results. The continuous score errors are mainly concentrated near the "general" level, which is in the transition interval between comfort and discomfort. Subjects have different sensitivities to the sense of wetness, pressure and activity obstruction, which is easy to cause fuzzy label boundaries. In contrast, the physiological feedback and subjective ratings of the "comfortable" and "uncomfortable" two end samples were more consistent, and the prediction error was relatively lower.

5 Conclusion

Focusing on the garment comfort prediction task, this paper completes the model construction from multimodal sample organization, feature joint coding, cross-modal attention fusion to application verification. Compared with the prediction methods that only rely on fabric parameters or single sensor data, the proposed model puts clothing appearance form, dynamic feedback of wearing process and material environmental conditions into the same computing link, which enhances the expression ability of comfort prediction for complex wearing states. The experimental part verifies the stability of the multimodal fusion structure in grade recognition and continuous score prediction through the comparison model, ablation model and error distribution analysis. The study illustrates that clothing comfort evaluation can shift from empirical judgment and static testing to data-driven modeling. In the future, long-term wearing samples, different exercise intensity scenes and more crowd data should be added to further improve the adaptability of the model in real clothing development and wearing evaluation.

Author's Profile

Yu Miao, female, holds a Doctorate in Design, is an associate professor, and is recognized as a Shanghai School Curriculum Ideological and Political Teaching Expert. She serves as the director of the Visual Communication Design Department at the Faculty of Art and Design of Shanghai Business School. Previously, she was an Assistant Professor at Tongmyong University in South Korea, a master's supervisor in Design at Dalian Polytechnic University, and a council member of the Dalian Fashion Designers Association. Currently, she is a member of the Shanghai Aesthetics Association and a researcher at the Tongmyong Institute of Sino-Korean Cultural and Artistic Exchange in South Korea. She has published over ten academic papers in Chinese, English, and Korean in core academic journals both domestically and internationally.

Funding

This work was supported by Shanghai Municipal Key Courses in Higher Education Institutions for the Year 2025,Shanghai Business School Course Construction AI-Enabled (AI+) Courses and Shanghai Business School 2025 annual curriculum ideological and political demonstration course construction project.

References

[1] ISLAM M R, GOLOVIN K, DOLEZ P I. Clothing thermophysiological comfort: A textile science perspective[J]. Textiles, 2023, 3(4): 353-407.

[2] CHENG P, ZENG X, BRUNIAUX P, et al. Design and research on multi-sensory comfort data acquiring of tight sportswear in motion[J]. Journal of Industrial Textiles, 2024, 54: 1-31.

[3] FENG M, FENG Y, CHENG J, et al. Clothing comfort sensing system based on triboelectric and tribological behavior of fabrics[J]. Nano Energy, 2024, 127: 109721.

[4] WEI Z, CALAUTIT J K, WEI S, et al. Real-time clothing insulation level classification based on model transfer learning and computer vision for PMV-based heating system optimization through piecewise linearization[J]. Building and Environment, 2024, 253: 111277.

[5] KANG L, GUO H, ZHOU X, et al. Deep learning and thermographic imaging method for thermal comfort prediction in different genders[J]. International Journal of Thermal Sciences, 2024, 197: 108804.

[6] YANG L, WANG F, ZHAO S, et al. Research on the local clothing thermal insulation prediction model with different dress state of indoor occupants[J]. Energy and Buildings, 2024, 324: 114861.

[7] TU Y F, KWAN M Y, YICK K L. A systematic review of AI-driven prediction of fabric properties and handfeel[J]. Materials, 2024, 17(20): 5009.

[8] RIBEIRO R, PILASTRI A, MOURA C, et al. A data-driven intelligent decision support system that combines predictive and prescriptive analytics for the design of new textile fabrics[J]. Neural Computing and Applications, 2023, 35: 17375-17395.

[9] RAMOS L, RIVAS-ECHEVERRÍA F, PÉREZ A G, et al. Artificial intelligence and sustainability in the fashion industry: A review from 2010 to 2022[J]. SN Applied Sciences, 2023, 5: 387.

[10] BAYOUDH K. A survey of multimodal hybrid deep learning for computer vision: Architectures, applications, trends, and challenges[J]. Information Fusion, 2024, 105: 102217.

[11] ZHAO F, ZHANG C, GENG B. Deep multimodal data fusion[J]. ACM Computing Surveys, 2024, 56(9): 1-36.

[12] LIU J, CAPURRO D, NGUYEN A, et al. Attention-based multimodal fusion with contrast for robust clinical prediction in the face of missing modalities[J]. Journal of Biomedical Informatics, 2023, 145: 104466.

[13] ZHANG H, WAN X, ZHENG R, et al. A new model for evaluating dynamic clothing thermal comfort[J]. International Journal of Thermal Sciences, 2025, 215: 109946.

[14] ZHANG H, HU Z, FAN J, et al. A new model for predicting the clothing overall and local thermal resistance under various body postures[J]. Building Simulation, 2025, 18(6): 1483-1498.

[15] CHEN L, ZHANG Y, WANG Y, et al. Meta-learning of personalized thermal comfort model and fast identification of the best personalized thermal environmental

conditions[J]. Building and Environment, 2023, 242: 110530.

[16] CUREAU R J, PIGLIAUTILE I, KOUSIS I, et al. Multi-domain human-oriented approach to evaluate human comfort in outdoor environments[J]. International Journal of Biometeorology, 2022, 66: 2033-2045.

[17] ZHANG X, HE L, ZOU H, et al. Evaluation and adjustment of clothing comfort based on fuzzy inference[J]. Smart Construction and Sustainable Cities, 2025, 3: 78.

[18] CAY G, YILMAZ G, KAYA T, et al. SolunumWear: A smart textile system for dynamic respiration monitoring across various postures[J]. iScience, 2024, 27(8): 110442.

[19] AZEEM M, WIENER J, PETRU M, et al. Design and development of textile-based wearable sensors for real-time biomedical monitoring[J]. The Journal of The Textile Institute, 2025, 116(1): 1-26.

[20] YANG K, ISAIAH A, HE Z, et al. E-textiles for sports and fitness sensing: Current state, challenges, and future opportunities[J]. Advanced Intelligent Systems, 2024, 6(3): 2300457.

[21] AKTER A, HOSSAIN M F, RAHMAN M M, et al. Recent studies on smart textile-based wearable sweat sensors[J]. Biosensors, 2024, 13(4): 40.

[22] ROLICH T, TOMASZEWSKI W, CHUDZIK M, et al. Advanced image analysis and machine learning models for textile porosity estimation[J]. Fibers, 2024, 12(5): 45.

[23] MA D, LI Y, ZHANG H, et al. Skin-core-fiber-based fabric integrated with pressure sensors and deep learning for real-time sitting posture recognition[J]. Nano Energy, 2024, 132: 110316.

[24] PENG D, LIU Y, ZHOU H, et al. Integration of attention mechanism and CNN-BiGRU for TDOA/FDOA collaborative mobile underwater multi-scene localization algorithm[J]. Complex & Intelligent Systems, 2024, 10: 9471-9492.

[25] GUO M H, XU T X, LIU J J, et al. Attention mechanisms in computer vision: A survey[J]. Computational Visual Media, 2022, 8(3): 331-368.