

Article

A hybrid network with dimensional attention for leaders' crisis response ability evaluation integrating affective computing and behavior recognition

Qianwen Wu*, Johar Permana, Taufani Chusnul Kurniatun, Suryadi Suryadi

Universitas Pendidikan Indonesia, Bandung 40154 Indonesia

* Corresponding author: Qianwen Wu, wuqw2026upi@outlook.com

CITATION

Wu Q, Permana J, Kurniatun T C, Suryadi S. (2026). A hybrid network with dimensional attention for leaders' crisis response ability evaluation integrating affective computing and behavior recognition. *Human Resources Management and Services*. 8(2): 5886.
<https://doi.org/10.18282/hrms5886>

ARTICLE INFO

Received: 23 March 2026

Accepted: 24 April 2026

Available online: 7 May 2026

COPYRIGHT



Copyright © 2026 by author(s).

Human Resources Management and Services is published by PiscoMed Publishing Pte. Ltd. This work is licensed under the Creative Commons Attribution (CC BY) license.

<https://creativecommons.org/licenses/by/4.0/>

Abstract: To address the issues of strong subjectivity and dimensional fragmentation in the evaluation of leaders' crisis response ability, this paper constructs a multimodal assessment model that integrates affective computing and behavior recognition. First, drawing on crisis management theory and emotional intelligence theory, deconstruct leaders' crisis response ability into three core dimensions—emotional stability, cognitive flexibility, and decisiveness—and establish a mapping between these dimensions and multimodal data, including facial expressions, speech, physiological signals, eye movements, and behavioral logs. Secondly, a hybrid network model is designed to achieve quantitative assessment across dimensions through multimodal feature encoding, Bi-LSTM temporal modeling, and a dimensional interaction attention mechanism. Experiments show that the model achieves Mean Absolute Errors of 4.12, 4.58, and 4.35 for emotional stability, cognitive flexibility, and decision-making decisiveness, respectively. These values represent a 17.4% reduction in overall MAE compared to the best-performing baseline (MulT). The consistency rate between the comprehensive evaluation index and expert rating levels reaches 86.7%, outperforming comparative methods such as MFN and MulT. The research demonstrates that the proposed model can effectively capture the multidimensional dynamic interactions during crisis response, providing an interpretable and highly robust quantitative tool for evaluating leaders' capabilities.

Keywords: leader's crisis response ability; affective computing; behavior recognition; multimodal fusion; emotional stability; cognitive flexibility; decision-making decisiveness; hybrid network model; capability assessment

1. Introduction

The suddenness, urgency, and high degree of uncertainty associated with crisis events pose severe challenges to leaders' psychological resilience, cognitive processing, and decision-making capabilities. With the rapid advancement of affective computing and behavioral recognition technologies, it has become possible to quantify an individual's psychological state through multimodal physiological signals and behavioral data, thereby providing a new technical approach for the scientific assessment of leaders' crisis response capabilities.

In recent years, scholars both domestically and internationally have conducted extensive research on individual behavior and psychology in crisis situations. In the field of affective computing, researchers analyze individuals' emotional states through signals such as facial expressions, speech prosody, and skin conductance. These methods have been applied in high-risk occupational settings, such as stress monitoring for pilots and fatigue warning systems for drivers, demonstrating the effectiveness of physiological and behavioral signals in quantifying emotions (Bian,

Kim, & Krumhuber, 2026). In the field of behavioral recognition, significant progress has also been made in modeling cognitive load and decision-making styles based on eye-tracking data (Ariza-Traslaviña, 2022), operation logs (Pettinato, Gil, Galeas, & Russo 2025), and posture data (Segil, Cuberovic, Graczyk, Weir, & Tyler 2020), providing objective metrics for understanding individuals' information processing and decision-making behaviors under stress. However, existing research has largely focused on single-modality or isolated-dimension analyses, lacking systematic modeling of the multidimensional construct of leaders' crisis response capabilities. Specifically, there are complex dynamic interactions among emotional stability, cognitive flexibility, and decisiveness; traditional methods struggle to simultaneously capture their evolutionary processes and mutual influences, leading to issues such as dimensional fragmentation and insufficient explanatory power in evaluation results. To address these shortcomings, this study introduces multimodal fusion and deep temporal modeling methods, aiming to construct an evaluation model capable of simultaneously capturing the dynamic relationships among these three dimensions.

To address the challenges of multimodal heterogeneous feature fusion and dimensional interaction modeling, this study draws on the attention mechanism widely used in cognitive neuroscience and human-computer interaction. This mechanism simulates the dynamic process by which humans focus on key information during complex tasks and has demonstrated the ability to adaptively weight multidimensional information in tasks such as functional connectivity analysis (Lu et al., 2025) and visual salience prediction (Kang, & Lee 2020). This characteristic aligns closely with the requirements of crisis response assessment, which necessitates the simultaneous processing of multimodal temporal data and the capture of complex inter-dimensional relationships. Therefore, this study applies the core principles of the attention mechanism to the task of evaluating leaders' crisis response capabilities, constructing a hybrid network model that integrates affective computing and behavioral recognition.

This paper addresses issues such as high subjectivity and the difficulty of quantifying dimensional interactions in the assessment of leaders' crisis response capabilities by proposing a multimodal evaluation model that integrates affective computing and behavioral recognition. Through multimodal feature encoding, temporal modeling using Bi-directional long-short-term memory (Bi-LSTM) networks, and a dimension-interaction attention mechanism, the model achieves objective quantification and dynamic assessment of three-dimensional capabilities. A dataset of simulated crisis response scenarios for leaders was constructed and experimentally validated, with the aim of providing technical support for the scientific assessment and diagnosis of leaders' crisis response capabilities.

2. Related works

The assessment of a leader's crisis response capabilities involves a multidimensional, comprehensive evaluation of an individual's emotions, cognition, and behavior; its core lies in how to objectively quantify a leader's psychological state and behavioral performance in crisis situations. In recent years, rapid advancements in affective computing, behavioral recognition, and multimodal fusion technologies

have provided both the theoretical foundation and technical tools necessary to achieve this goal.

Affective computing enables the automatic recognition and analysis of an individual's emotional state through physiological and behavioral signals. In the field of facial expression analysis, facial action unit detection based on convolutional neural networks has become the mainstream method, capable of extracting high-precision facial expression features from video sequences (Agrawal, Kumar, Swathi, Yashwanthi, & Hegde, 2021). In the field of speech emotion recognition, researchers extract acoustic features such as fundamental frequency, energy, and speech rate, and combine them with long short-term memory (LSTM) networks to model temporal dependencies in speech, thereby distinguishing emotions such as anger, anxiety, and calmness (Oruh, Viriri, & Adegun, 2022). Behavioral recognition focuses on analyzing an individual's intentions, cognitive load, and decision-making style through body posture, eye-tracking trajectories, and operation logs. Skeleton point extraction techniques based on tools such as OpenPose are used to quantify behavioral features such as gestures and head movements (Y. Chen, Zhang, & Wang, 2024). In the field of eye-tracking analysis (Fang et al., 2026), metrics such as gaze trajectories and rates of change in pupil diameter are used to assess the breadth of information search and the level of cognitive effort.

In the early stages, assessments of crisis response capabilities primarily relied on psychological scales and interviews, such as self-report questionnaires based on emotional intelligence theory and behavior-anchored rating scales based on crisis management stages. While these methods can yield rich qualitative information, they suffer from significant subjective bias and the inability to conduct real-time, dynamic assessments. With the advancement of wearable devices and sensor technology, some studies have begun to incorporate physiological and behavioral data into the assessment of an individual's state during crisis situations. For example, by analyzing the stress responses of fire commanders in simulated fire scenes using skin conductance and heart rate variability, researchers found a significant correlation between high-frequency heart rate variability and decision-making quality (He, Li, Wang, & Li, 2026). An analysis of pilots' attention allocation during emergencies using eye-tracking technology revealed that the frequency of gaze shifts is positively correlated with situational awareness levels (Baron, Martins, Puggioni, & Harwood, 2026). However, existing research has largely focused on a single dimension, such as emotion or cognition. These studies lack models of the dynamic interactions among different behaviors, making it difficult to fully capture the full picture of a leader's crisis response capabilities.

Multimodal fusion aims to integrate information from different sources to improve model performance and has been widely applied in fields such as affective computing and human-computer interaction. In recent years, deep learning-based fusion models have become a hot topic of research. For example, memory fusion networks use gating mechanisms to pass information between different modalities, capturing the dynamic interactions between them (Ibrahimi, Abbad, & Ben Abbou, 2026); Transformer directly models cross-modal relationships using cross-modal attention, achieving state-of-the-art performance in sentiment analysis tasks (Q. Zhang et al., 2026). In addition, in the field of time-series modeling, bidirectional long-short-

term memory (LSTM) networks and convolutional time-series networks are widely used to process multimodal time series, as they can effectively capture the evolution of features over time (J. Chen et al., 2026). These methods provide valuable insights for this study in terms of processing multimodal heterogeneous data and modeling interactions across dimensions.

These methods provide valuable insights for this study in handling multimodal heterogeneous data and modeling interactions across dimensions. However, existing research is largely limited to single-modal or isolated-dimension analysis; they have failed to fully leverage the temporal dependencies inherent in crisis evolution; and they lack specialized evaluation models for leader groups, with their practicality yet to be validated. To address these issues, this study proposes a hybrid network evaluation model that integrates affective computing and behavioral recognition. Through multimodal feature encoding, temporal dependency modeling, and a dimensional interaction attention mechanism, the model achieves synchronous quantification and dynamic evaluation of three-dimensional capabilities and validates its effectiveness on a self-constructed dataset of simulated leader crisis response scenarios. Unlike Transformer-based methods (e.g., MulT) that rely on pairwise cross-modal attention without explicit dimensional interaction, our model introduces dimension-specific query vectors and a dedicated interaction network. Compared to graph-based fusion (e.g., Graph-FCN) that treats modalities as nodes but ignores temporal dynamics, our Bi-LSTM captures sequential dependencies.

3. Development of assessment indicators for leaders' crisis response capabilities

3.1. Overview of competency assessment dimensions

A leader's crisis response capability is not a single trait, but rather a complex construct comprising psychological states, cognitive processes, and behavioral performance. Based on the stages of crisis management theory (L. Zhang, Dong, Zhang, Mu, & Yang, 2025) and emotional intelligence theory (Emmerling & Cherniss, 2025), three core assessment dimensions have been identified that constitute a leader's crisis response capability, as follows.

- (1) Emotional stability refers to a leader's ability to regulate their emotional state and maintain psychological balance in high-pressure situations triggered by a crisis.
- (2) Cognitive flexibility refers to a leader's ability to break free from fixed mindsets, synthesize fragmented information, and generate innovative solutions when facing dynamic and evolving crisis situations.
- (3) Decisiveness refers to a leader's ability to take responsibility and make decisions quickly under conditions of limited information and intense time pressure.

Emotional stability provides a stable psychological foundation for cognitive activities; cognitive flexibility offers high-quality options to support decision-making; and decisiveness translates cognition into action. At the same time, there is a dynamic interaction among these three dimensions: excessive emotional fluctuations may inhibit cognitive flexibility, while excessive cognitive load may also affect the quality of decision-making. This nonlinear, dynamic interaction requires assessment models

to possess the ability to process multimodal time-series data and capture complex inter-dimensional relationships. While situational awareness and communication effectiveness are also relevant to crisis leadership, they are either subsumed under cognitive flexibility (e.g., information gathering and sensemaking) or not directly measurable via the current multimodal signals (e.g., communication effectiveness requires dyadic interaction data). Thus they are excluded for model parsimony. The three dimensions are rooted in established leadership and crisis management literature. Emotional stability draws from research on leader trait emotional intelligence and stress resilience (Emmerling & Cherniss, 2025; He et al., 2026). Cognitive flexibility aligns with studies on adaptive sensemaking and leader cognitive agility under uncertainty (Zhang et al., 2025). Decisiveness has been examined in the context of time-pressured leadership decision-making (Baron et al., 2026). However, prior work typically examines these dimensions in isolation; the present study integrates them into a unified multimodal assessment framework.

3.2. Mapping relationships between multimodal metrics

Once the core evaluation dimensions have been established, these abstract psychological and behavioral concepts must be translated into quantifiable metrics. To this end, this study establishes a mapping relationship between the core evaluation dimensions and multimodal data, as shown in **Table 1**.

Table 1. relationship between the core evaluation dimensions and multimodal.

Key assessment dimensions	Data modalities	Quantifiable metrics (units)
Emotional Stability	Video (Facial Expressions)	Temporal patterns of facial action unit intensity (0–1)
	Audio (Speech)	Fundamental frequency jitter—five-point period perturbation quotient (%), speech rate coefficient of variation (unitless)
	Physiological Signals (Wearables)	Peak frequency of skin conductance response (Hz), high-frequency power of HRV (ms ²)
Cognitive Flexibility	Video (Eye Movements)	Frequency of interest-zone transitions (Hz), rate of change in pupil diameter (mm/s)
	Behavioral Logs (Operation Records)	Information source access entropy (nats), number of generated solutions (count)
	Audio (Speech Content)	Cognitive lexical density (ratio), syntactic complexity (unitless)
Decisiveness	Behavioral Logs (Operation Records)	First decision reaction time (s), number of post-decision revisions (count)
	Audio (Speech)	Intonation drops amplitude (semitone), proportion of ambiguous words in instructions (%)
	Video (Posture)	Gesture energy ((m/s) ² + (m/s) ²), head posture stability (degrees)

On the emotional stability dimension, quantification is primarily based on facial expressions, vocal characteristics, and physiological signals. The cognitive flexibility dimension integrates three types of data: eye-tracking data, behavioral logs, and verbal content. In the decision-making decisiveness dimension, behavioral logs, voice characteristics, and posture information are integrated. Through this mapping relationship, psychological traits can serve as the raw data for this study.

3.3. Justification of cognitive flexibility metrics

The metrics for emotional stability follow established psychophysiological principles: facial action units index overt expression, fundamental frequency jitter reflects voice tremor under arousal, and HF-HRV captures parasympathetic regulation. For decisiveness, reaction time and post-decision revisions directly operationalize decision speed and commitment, while intonation drop and gesture energy indicate confidence and certainty. The metrics selected for the cognitive flexibility dimension are grounded in established concepts from cognitive psychology and neuroscience. Below explain why each metric serves as a valid indicator of cognitive flexibility in crisis situations.

- (1) Frequency of ROI transitions: Cognitive flexibility involves the ability to shift attention between different information sources rather than perseverating on a single source. In a dynamic crisis environment, a flexible leader frequently switches gaze between competing displays (e.g., hospital data, official guidelines, social media feeds) to integrate diverse perspectives. A low transition frequency indicates fixation or tunnel vision, which is characteristic of cognitive rigidity. Therefore, the frequency of region-of-interest transitions directly operationalizes attentional flexibility.
- (2) Rate of change in pupil diameter: Pupil diameter is not merely a response to luminance; it also reflects cognitive effort and the updating of mental representations. Rapid changes in pupil diameter indicate that the individual is actively reallocating cognitive resources in response to new or conflicting information. In contrast, a sluggish or flat pupil response may signal an inability to adjust mental models as the crisis evolves. Thus, the rate of change in pupil diameter serves as a physiological marker of cognitive flexibility.
- (3) Information source access entropy: Cognitive flexibility requires exploration of multiple information channels before converging on a solution. Entropy, as a measure of diversity, quantifies how evenly the leader distributes attention across available sources. High entropy means the leader samples broadly, avoiding premature closure on a single source of information. Low entropy indicates over-reliance on a limited set of inputs, which is a sign of inflexible information foraging. This metric captures the exploratory aspect of cognitive flexibility.
- (4) Cognitive lexical density and syntactic complexity: Spoken language during crisis response reflects the underlying structure of thought. Cognitive flexibility involves the ability to restructure mental representations and generate novel solutions. This internal restructuring often manifests in speech as higher lexical density (more content words relative to function words) and more complex syntax (e.g., subordinate clauses, conditionals). Conversely, rigid thinking tends to

produce simpler, more repetitive language patterns. Hence, these linguistic features serve as behavioral indicators of cognitive flexibility.

4. An evaluation model integrating affect computing and behavior recognition

In this section, an end-to-end hybrid network evaluation model is constructed, as shown in **Figure 1**. First, due to the multi-source and heterogeneous nature of the data, it is transformed through data preprocessing and feature engineering. Next, the hybrid network model captures temporal dependencies and dimensional interactions to generate evaluation scores for each dimension. Finally, an evaluation decision is made to provide a comprehensive assessment and diagnostic feedback.

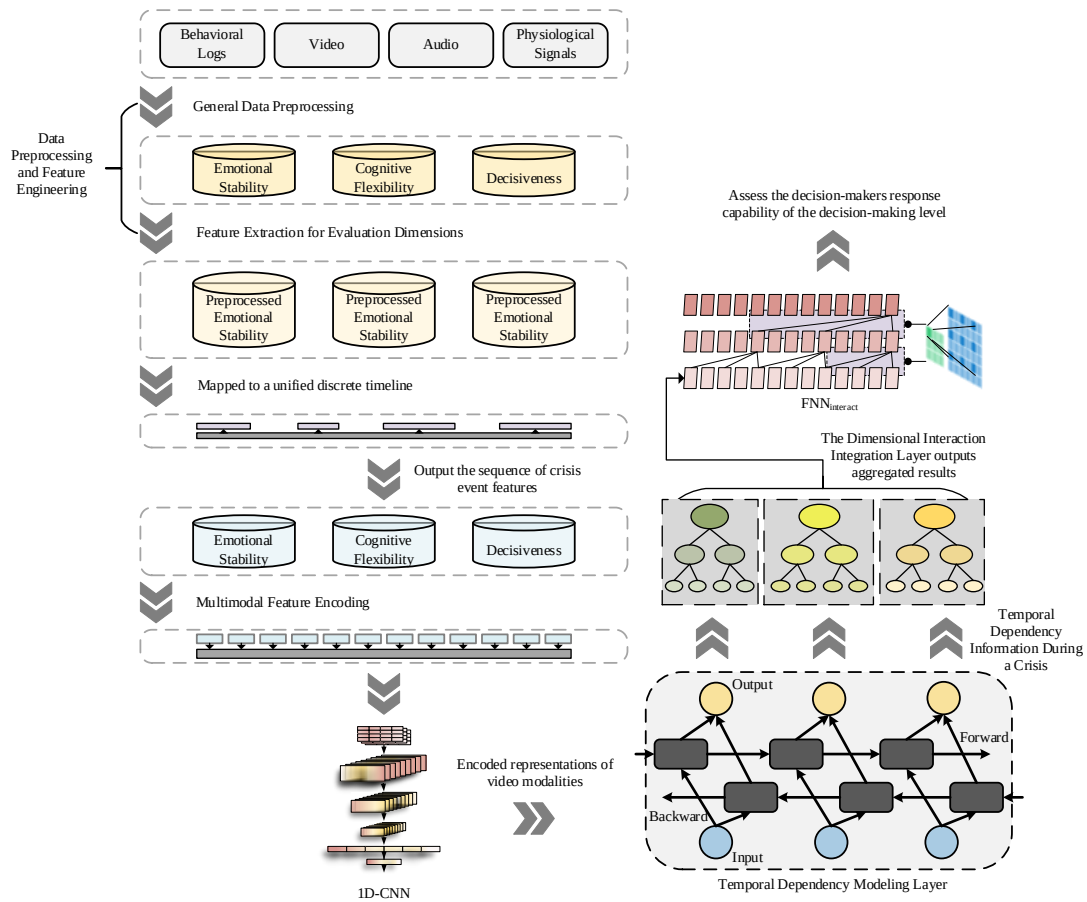


Figure 1. Architecture of the Hybrid Network Evaluation Model.

Table 2. Nomenclature.

Symbol	Definition	First equation	Unit / default value
$x_{\text{raw}}^{(m)}(t)$	Raw signal of modality m at time t	(1)	varies
$y(t)$	Filtered signal	(1)	varies
N, M	Orders of forward/reverse filters	(1)	dimensionless

Symbol	Definition	First equation	Unit / default value
a_i, b_j	Forward/reverse filter coefficients	(1)	dimensionless
τ	Missing duration threshold for linear interpolation	(2)	0.5 s
$\hat{x}(t)$	Interpolated signal value at time t	(2)	varies
Jitter _{ppq5}	Five-point period perturbation quotient (fundamental frequency jitter)	(3)	%
HF _{power}	High-frequency power of heart rate variability (0.15–0.4 Hz)	(4)	ms ²
$P_{RR}(f)$	Power spectral density of R-R intervals	(4)	ms ² /Hz
$\dot{d}_{rel}(t)$	Relative rate of change of pupil diameter at time t	(5)	s ⁻¹
H_{info}	Information source access entropy	(6)	nats
p_i	Probability of accessing the i-th information source	(6)	dimensionless
T_{react}	First decision reaction time	(7)	s
N_{rev}	Number of post-decision revisions	(7)	dimensionless
β	Balance coefficient in gesture energy (velocity vs. acceleration)	(7)	0.5 (dimensionless)
$E_{gesture}(t)$	Gesture energy at time t	(7)	(m/s) ² + (m/s) ²
T_{win}	Integration time window length for gesture energy	(7)	1 s
Δt	Time window duration for feature alignment	(8)	1 s
f_t	Feature vector at time window t	(8)	varies
h_t	Temporal context representation from Bi-LSTM at time step t	(11)	dimensionless
$\alpha_t^{(k)}$	Attention weight for dimension k at time step t	(12)	dimensionless
q_k	Learnable query vector for dimension k	(12)	trainable
r_k	Attention-aggregated representation for dimension k	(13)	trainable
$\hat{s}_k$	Predicted score for dimension k (range 0–1)	(15)	dimensionless
θ	Weakest-link threshold for penalty term	(17)	0.4 (dimensionless)
λ	Penalty intensity coefficient	(17)	0.5 (dimensionless)
w_k	Dimension weight (crisis-dependent)	(18)	dimensionless, sum to 1

4.1. Data preprocessing and feature engineering

To address data discrepancies, this section first resolves common issues across the data from each modality, then performs feature extraction based on evaluation dimensions, and finally aligns the time series to form feature vectors.

4.1.1. General data preprocessing

General preprocessing includes time synchronization, noise filtering, and handling of missing values. Time Synchronization: First, synchronize the timestamps of all data to a single reference point using the Network Time Protocol (NTP). For video and audio, synchronization is achieved through audio-video sync signals. The synchronized data stream has a unified timeline, denoted as the original signal $s^{(m)}(t)$, where m represents the modality category.

Noise filtering: Linear filtering is used to remove high-frequency noise and power-line interference from the signal. For continuous signals such as video, physiological, and eye-tracking signals, a low-pass filter is designed, as shown in Equation (1).

$$y[n] = \sum_{k=0}^N b_k x[n-k] - \sum_{k=1}^M a_k y[n-k] \quad (1)$$

Here, $x[n]$ represents the input sequence, $y[n]$ represents the filtered sequence, N and M denote the orders of the forward and reverse filters, respectively, and b_k and a_k denote the coefficients of the forward and reverse filters, respectively. $x[n-k]$ represents the value of the input signal at time $[n-k]$. For example, the useful components of the skin conductance signal are concentrated in the 0.05–1.5 Hz range, which is retained using a bandpass filter; in contrast, the photoplethysmography signal is set to 0.5–5 Hz based on the heart rate range.

Handling missing values: Temporary sensor failure or loss of line of sight may result in missing data. For missing data with a duration shorter than the threshold τ , linear interpolation is used to fill in the gaps, as shown in Equation (2); for long-term missing data, the segment is marked as invalid and excluded from subsequent feature aggregation to prevent spurious signals from contaminating the feature extraction results.

$$\hat{s}(t_i) = s(t_{i-1}) + \frac{s(t_{i+1}) - s(t_{i-1})}{t_{i+1} - t_{i-1}}(t_i - t_{i-1}), \quad \text{if } t_{i+1} - t_{i-1} < \tau \quad (2)$$

Where $\hat{s}(t_i)$ represents the interpolated estimate of the signal value at time point t_i . t_i represents the current point of time to be interpolated. τ represents the missing duration threshold, set to 0.5 s in this study.

4.1.2. Physiological signal quality control

To ensure data quality and minimize noise from motion artifacts and environmental interference, applied the following procedures:

- (1) Bandpass filtering: The skin conductance signal was filtered with a 0.05–1.5 Hz bandpass filter to retain phasic components while removing slow drifts and high-

frequency noise. The photoplethysmography (PPG) signal was filtered with a 0.5–5 Hz bandpass filter to preserve heart rate information.

- (2) Artifact rejection and correction: For skin conductance, movement-induced spikes were detected using a wavelet-based algorithm (Daubechies 4 wavelet, threshold set at 3 median absolute deviations). Detected artifacts were removed and replaced via linear interpolation of the surrounding 2-s window. For PPG, R-R intervals deviating more than 20% from the local median (computed over a 5-beat window) were considered artifacts and replaced using cubic spline interpolation.
- (3) Signal-to-noise ratio (SNR) assessment: computed the SNR for each physiological recording as $SNR = 10\log_{10}(P_{\text{signal}}/P_{\text{noise}})$, where P_{noise} was estimated from artifact-free segments during baseline. The mean SNR across all recordings was 18.3 ± 3.2 dB for skin conductance and 22.1 ± 4.5 dB for PPG, indicating acceptable quality for further analysis.

4.1.3. Feature extraction for evaluation dimensions

Emotional Stability Dimension: This dimension encompasses three aspects: overt facial expressions, vocal characteristics, and implicit physiological responses. Facial expression analysis employs a pre-trained facial expression convolutional neural network to extract facial action unit intensity values from a sequence of facial images, generating an intensity time series curve $AU(t)$. In speech feature extraction, the fundamental frequency fluctuation parameter shown in Equation (3) reflects the degree of voice tremor. At the same time, the coefficient of variation for speech rate is calculated, which is the coefficient of variation $\sigma(R)/\mu(R)$ for the sequence $R(t)$ of syllables per unit time. In terms of physiological characteristics, the skin conductance signal is continuously decomposed and analyzed to extract the peak frequency f_{SCR} of the skin conductance response; the R-R interval sequence is calculated from the photoplethysmography signal, and the high-frequency power HF is extracted in the frequency domain.

$$\text{Jitter} = \frac{1}{N-1} \sum_{i=1}^{N-1} \frac{|F_0[i+1] - F_0[i]|}{(F_0[i+1] + F_0[i]) / 2} \quad (3)$$

$$\text{HF} = \int_{0.15}^{0.4} P(f) df \quad (4)$$

Where Jitter represents the fundamental frequency jitter parameter. N represents the frame count. $F_0[i]$ represents the fundamental frequency value of the i frame. HF represents the high-frequency power of heart rate variability. $P(f)$ represents the power spectral density function of the R-R interval sequence. f represents frequency, measured in Hz. $\int_{0.15}^{0.4}$ represents the frequency range of the high-frequency band.

Physiological baseline normalization: To control for individual differences in autonomic reactivity, each participant underwent a 5-min resting baseline recording before the simulation (seated quietly with eyes open, no task). For skin conductance, subtracted the mean tonic level during baseline from the raw signal; phasic responses

were identified as peaks exceeding 2 standard deviations above the baseline mean. For heart rate variability, the high-frequency power (HF power) was expressed as a ratio to the baseline HF power. This within-subject normalization effectively removes inter-individual variability unrelated to crisis response.

Cognitive flexibility dimension: In eye-tracking analysis, regions of interest (ROIs) are defined based on the information displayed on the screen, and the frequency f_{trans} of gaze shifts between these ROIs is calculated to reflect the breadth of information search; the rate of change in pupil diameter is calculated as shown in Equation (5). In behavioral log analysis, search diversity is quantified using information source access entropy, as shown in Equation (6).

$$\Delta D_{\text{pupil}}(t) = \frac{D(t) - D_{\text{base}}}{D_{\text{base}}} \quad (5)$$

$$H_{\text{info}} = -\sum_{i=1}^N p_i \log p_i \quad (6)$$

Where $\Delta D_{\text{pupil}}(t)$ represents the relative change rate of pupil diameter at time t . $D(t)$ represents the measured pupil diameter at time t . D_{base} represents the average pupil diameter during the resting baseline period. H_{info} represents the information source access entropy, which quantifies the diversity of information search. p_i represents the probability of accessing the i -th information source. N represents the total number of accessible information sources. The negative sign ensures that the entropy value is positive.

Decision-making decisiveness dimension: In the behavioral log, the initial decision-making reaction time T_{react} is defined as the time interval between the presentation of critical crisis information and the first decision-making action, and must be adjusted based on the individual's baseline reaction speed; the number of post-decision revisions N_{revise} records the number of times the plan was significantly modified after the initial decision was made. For speech features, perform intonation analysis on decisive statements to calculate the degree of fundamental frequency drop at the end of the sentence $\Delta F_{0,\text{end}}$; calculate the proportion of ambiguous words in command statements r_{vague} . For posture features, calculate gesture energy based on skeletal keypoints:

$$E_{\text{gesture}}(t) = \int_{t-\Delta t}^t \left(\|\mathbf{v}_{\text{hand}}(\tau)\|^2 + \beta \|\mathbf{a}_{\text{hand}}(\tau)\|^2 \right) d\tau \quad (7)$$

Where $E_{\text{gesture}}(t)$ represents the gesture energy at time t , reflecting the activity level of the gesture. t represents the current time. Δt represents the length of the integration time window. τ represents the integration variable, denoting each time point within the window. $\mathbf{v}_{\text{hand}}(\tau)$ represents the velocity vector of the hand's key points at time τ . $\|\cdot\|^2$ represents the squared norm. $\mathbf{a}_{\text{hand}}(\tau)$ represents the acceleration

vector of the hand's key points at time τ . β represents the balance coefficient, set to 0.5 to give equal weight to velocity and acceleration terms.

After feature extraction across different modalities, the temporal resolution remains inconsistent; for example, video features are generated at frame rates, speech features are extracted on a sentence-by-sentence basis, and physiological features are obtained using sliding windows at one-second intervals. For each time window $[t - \Delta T, t]$, statistical aggregates are calculated for the continuous signal features to preserve the dynamic information within the window. These aggregates include the mean, variance, maximum, and minimum. For discrete event feature counting or density estimation. For modes with no events within the window, forward filling is applied. Through this alignment process, a feature vector $\mathbf{x}_t$ is obtained for each time window. Finally, the sequence of feature vectors for the entire crisis event is denoted as $\mathcal{X}$:

$$\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\} \quad (8)$$

Where T denotes the total number of time windows. $\mathbf{x}_t$ denotes the feature vector of the t-th time window.

This sequence preserves the temporal dependencies in the evolution of the crisis and aggregates multimodal information across three modal dimensions within the same time step.

4.2. Hybrid network evaluation model

In this section, the hybrid network evaluation model is constructed to describe the complex relationships in $\mathcal{X}$, where the multimodal feature encoding layer maps heterogeneous features into a unified representation space; the temporal dependency modeling layer captures the dynamic changes in the evolution of a crisis; and the dimensional interaction fusion layer adaptively integrates information across three dimensions via an attention mechanism to generate evaluation scores for each dimension.

4.2.1. Multimodal feature encoding layer

The multimodal feature encoding layer maps heterogeneous features from different modalities into a unified representation space. Suppose the input $\mathbf{x}_t$ at time step t can be decomposed into M modal subvectors $\mathbf{x}_t = [\mathbf{x}_t^{(1)}; \mathbf{x}_t^{(2)}; \dots; \mathbf{x}_t^{(M)}]$, where M = 5 corresponds to the five modalities of video, audio, physiological signals, eye tracking, and behavioral logs. For each modality m, design a coding function $f_{\theta_m} : \mathbb{R}^{d_m} \rightarrow \mathbb{R}^{d_h}$ that maps modality-specific features to a latent space of dimension d_h . The form of the encoding function is selected based on the characteristics of the modality: for continuous signal features such as video, audio, and physiological signals, a one-dimensional convolutional neural network (1D-CNN)(Du, 2025) is used to extract local temporal patterns; for eye-tracking features, a fully connected network is used; and for sparse discrete features such as behavioral logs, an embedding layer combined with a fully connected network is used. Taking the video modality as an example, its

encoding process can be represented as shown in Equation (9). Similarly, after passing through their respective encoding networks, other modalities yield the modality-specific latent representation $\mathbf{h}_t^{(1)}, \mathbf{h}_t^{(2)}, \dots, \mathbf{h}_t^{(M)}$ for that time step.

$$\mathbf{h}_t^{(vid)} = \text{ReLU}\left(\text{BN}\left(\text{Conv1D}\left(\mathbf{x}_t^{(vid)}; \mathbf{W}_{vid}\right)\right)\right) \quad (9)$$

Where $\mathbf{h}_t^{(vid)}$ denotes the encoded latent representation of the video modality at time step t . $\mathbf{x}_t^{(vid)}$ denotes the input feature vector of the video modality at time step t . Conv1D denotes a one-dimensional convolution operation. $\mathbf{W}_{vid}$ denotes the convolution kernel weights of the video modality encoding network. BN denotes a batch normalization operation. ReLU denotes a linear rectified activation function.

4.2.2. Temporal dependency modeling layer

Crisis response is a dynamic, evolving process in which leaders' emotions, cognition, and decision-making behaviors influence one another over time. Bi-LSTM is chosen over Transformer-based temporal encoders (e.g., MulT) for two reasons: (1) our dataset size (240 samples) is moderate, and Bi-LSTM is less prone to overfitting; (2) the temporal dependencies in crisis evolution are relatively local and sequential, making LSTM's recurrent structure more parameter-efficient. Therefore, a bidirectional long-short-term memory (Bi-LSTM) network (Cao et al., 2025) is employed to process the encoded sequence of latent representations and model these temporal dependencies.

Concatenate the modality-specific latent representations for each time step to obtain the fused time-step representation $\mathbf{h}_t = [\mathbf{h}_t^{(1)}; \mathbf{h}_t^{(2)}; \dots; \mathbf{h}_t^{(M)}]$. Sequence $\{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_T\}$ is input into the Bi-LSTM, and its forward and backward propagation processes are shown in Equation (10). By concatenating the forward and backward hidden states according to Equation (11), the temporal context representation $\mathbf{u}_t$ for each time step is obtained.

$$\bar{\mathbf{u}}_t = \text{LSTM}_{\text{fwd}}(\mathbf{h}_t, \bar{\mathbf{u}}_{t-1}), \quad \tilde{\mathbf{u}}_t = \text{LSTM}_{\text{bwd}}(\mathbf{h}_t, \tilde{\mathbf{u}}_{t+1}) \quad (10)$$

$$\mathbf{u}_t = [\bar{\mathbf{u}}_t; \tilde{\mathbf{u}}_t] \quad (11)$$

Where $\bar{\mathbf{u}}_t, \tilde{\mathbf{u}}_t$ denote the hidden states of the forward and backward LSTM at time step t , respectively. $\text{LSTM}_{\text{fwd}}, \text{LSTM}_{\text{bwd}}$ denote the forward and backward LSTM units, respectively. $\mathbf{h}_t$ denotes the input at time step t . d_u denotes the dimension of the LSTM hidden state. $[\cdot; \cdot]$ denotes the vector concatenation operation.

The output $\mathbf{u}_t$ of the Bi-LSTM not only contains multimodal information from the current time step but also integrates contextual information from both before and after that time step, enabling it to effectively capture dynamic patterns in the progression of crisis events.

4.2.3. Dimensional interaction integration layer

Output $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_T\}$ from the temporal modeling layer contains information on

temporal dependencies throughout the entire crisis process, but ultimately needs to produce quantitative scores for the three evaluation dimensions. In addition, there is an interaction among these three dimensions; for example, emotional fluctuations may indirectly influence decision-making by affecting the allocation of cognitive resources, while excessive cognitive load may, in turn, exacerbate emotional instability. Therefore, the dimension fusion layer needs to extract the information most relevant to each dimension from the temporal representations, while simultaneously modeling the interactions among the dimensions.

Unlike standard temporal attention (e.g., LSTM+Attn) which only reweights time steps for a single output, our mechanism learns per-dimension query vectors that selectively attend to different time periods, and further model's inter-dimensional influences via a dedicated interaction network. Unlike cross-modal attention (e.g., MulT) which focuses on pairwise modality interactions, our design explicitly captures the dynamic interplay among emotional, cognitive, and decisional processes during crisis evolution. This is a key methodological contribution. Design a dimension-specific attention mechanism to aggregate information from the temporal representations for each evaluation dimension. For dimension $k \in \{\text{emo}, \text{cog}, \text{dec}\}$, compute the attention weights $\alpha_{t,k}$ as shown in Equation (12). By performing a weighted sum of the temporal representations using attention weights, the aggregated representation for each dimension is obtained, as shown in Equation (13).

$$\alpha_{t,k} = \frac{\exp(\mathbf{q}_k^\top \tanh(\mathbf{W}_k \mathbf{u}_t + \mathbf{b}_k))}{\sum_{j=1}^T \exp(\mathbf{q}_k^\top \tanh(\mathbf{W}_k \mathbf{u}_j + \mathbf{b}_k))} \quad (12)$$

$$\mathbf{v}_k = \sum_{t=1}^T \alpha_{t,k} \mathbf{u}_t \quad (13)$$

Where $\alpha_{t,k}$ denotes the attention weight for dimension k , at time step t . $\mathbf{q}_k$ denotes the learnable query vector for the attention mechanism of dimension k . $\tanh$ denotes the hyperbolic tangent activation function. $\mathbf{W}_k$ denotes the learnable weight matrix for the attention mechanism of dimension k . $\mathbf{u}_t$ denotes the temporal context representation at time step t . $\mathbf{b}_k$ denotes the learnable bias term of the k -dimensional attention mechanism. $\exp$ denotes the exponential function. T denotes the total number of time windows. j denotes the summation index, traversing all time steps. $\mathbf{v}_k$ denotes the attention aggregation representation for dimension k . $\text{emo}, \text{cog}, \text{dec}$ denote the dimensions of emotional stability, cognitive flexibility, and decisiveness, respectively.

This attention mechanism allows the model to focus on different time periods based on the requirements of each dimension. For example, the emotional stability dimension may focus more on the sharp fluctuations during the early stages of a crisis, while the decisiveness dimension may focus more on local features now of decision-making.

At this point, $\mathbf{v}_{emo}$, $\mathbf{v}_{cog}$, $\mathbf{v}_{dec}$ each focuses on the time-related information most relevant to their respective dimensions. To model the interactions between dimensions, an interaction network $FFN_{interact}$ is designed; this network is a two-layer fully connected network that employs a residual connection structure. The three vectors are concatenated and fed into $FFN_{interact}$, ensuring that the output $\tilde{\mathbf{v}}_k$ for each dimension retains its own primary information while incorporating the influence of other dimensions.

$$[\tilde{\mathbf{v}}_{emo}; \tilde{\mathbf{v}}_{cog}; \tilde{\mathbf{v}}_{dec}] = FFN_{interact}([\mathbf{v}_{emo}; \mathbf{v}_{cog}; \mathbf{v}_{dec}]) \quad (14)$$

Finally, $\tilde{\mathbf{v}}_k$ is subjected to regression to obtain the evaluation scores for each dimension:

$$s_k = \sigma(\mathbf{w}_k^T \tilde{\mathbf{v}}_k + b_k) \mathbf{w}_k \quad (15)$$

Where $\mathbf{v}_{emo}$, $\mathbf{v}_{cog}$, $\mathbf{v}_{dec}$ denote three-dimensional aggregations. s_k represents the final evaluation score across the three dimensions, with a range of [0,1]. $\mathbf{w}_k$ denotes the weight vector of the dimensional regression network. $\tilde{\mathbf{v}}_k$ represents the representations of each dimension after dimensional interaction. b_k denotes the bias of the dimensional regression network. $\top$ denotes the transpose operator. σ is the sigmoid activation function.

4.3. Assess the decision-makers

$(s_{emo}, s_{cog}, s_{dec})$ framework evaluates a leader's crisis response capabilities from various perspectives but does not provide an overall assessment of their comprehensive capabilities. Therefore, the decision-making body conducts a comprehensive evaluation and assigns ratings based on scores across the three dimensions.

4.3.1. Comprehensive assessment index

The construction of a comprehensive assessment index must reflect the nonlinear interactions among dimensions and dynamically adjust the relative importance of each dimension based on the crisis. Therefore, the composite index should not be a simple weighted average, but should be capable of capturing these interactions. Design a nonlinear fusion function combining the geometric mean and a penalty term, as shown in Equation (16).

$$S = \left(s_{emo}^\alpha \cdot s_{cog}^\beta \cdot s_{dec}^\gamma \right)^{\frac{1}{\alpha+\beta+\gamma}} \cdot P(s_{emo}, s_{cog}, s_{dec}) \quad (16)$$

Where S represents the comprehensive evaluation index, with a range of [0,1]. α, β, γ represent the weighting indices for each dimension.

The first component is a weighted geometric mean, whose product form reflects the interdependence among dimensions. In other words, a low score in any single dimension will significantly lower the composite index, which aligns with the

“weakest link effect” observed in crisis response. The second component $P(\cdot)$ is a penalty term used to handle outliers. The penalty term is designed as follows.

$$P(s_{\text{emo}}, s_{\text{cog}}, s_{\text{dec}}) = 1 - \lambda \cdot \mathbb{I}(\min(s_{\text{emo}}, s_{\text{cog}}, s_{\text{dec}}) < \theta) \cdot (\theta - \min(s_{\text{emo}}, s_{\text{cog}}, s_{\text{dec}})) \quad (17)$$

Where θ represents the threshold for the weakest link, λ represents the penalty intensity coefficient, and $\mathbb{I}(\cdot)$ represents the indicator function. $\theta - \min(\cdot)$ represents the degree of weakness, which is the difference between the threshold and the lowest score.

When the score for a particular dimension falls below the threshold, that dimension is considered to have a serious deficiency, and the composite index will be penalized in proportion to the severity of the deficiency; when all dimensions exceed the threshold, the penalty factor is 1, and there is no impact.

Different types of crises may place varying emphasis on different dimensions: for example, in sudden public safety incidents, decisive decision-making may be more critical, whereas in technical crises requiring complex information analysis, the importance of cognitive flexibility should be appropriately increased. To this end, the weights are expressed as functions of crisis scenario characteristics, as shown in Equation (18).

$$(\alpha, \beta, \gamma) = \text{Softmax}(\mathbf{W}_{\text{ctx}} \mathbf{c}_{\text{crisis}} + \mathbf{b}_{\text{ctx}}) \quad (18)$$

Where $\mathbf{c}_{\text{crisis}}$ represents the features of the crisis scenario, such as crisis type, urgency, and complexity. $\mathbf{W}_{\text{ctx}}$ and $\mathbf{b}_{\text{ctx}}$ represent trainable parameters, Softmax function ensures that the sum of the weights equals 1. In this way, the comprehensive assessment index can adaptively adjust the relative importance of each dimension based on the specific crisis scenario, thereby ensuring flexibility and specificity in the assessment.

To evaluate the robustness of the comprehensive assessment index (Equation 16) to the choice of the weakest-link threshold θ and penalty intensity λ , a sensitivity analysis was performed on the test set (N = 36 samples).

Varying θ (fixed $\lambda = 0.5$): varied from 0.2 to 0.8 in steps of 0.1. For each θ , the comprehensive index C was computed for three representative participants: (a) a balanced participant (scores: Emo = 0.72, Cog = 0.68, Dec = 0.70); (b) a participant weak in emotional stability (Emo = 0.25, Cog = 0.70, Dec = 0.68); (c) a participant weak in decisiveness (Emo = 0.71, Cog = 0.69, Dec = 0.26).

Results: When $\theta < 0.3$, the penalty term is rarely activated (only 2.8% of test samples had a dimension score below 0.3), so C approximates the standard weighted geometric mean. At $\theta = 0.4$ (the chosen value), the penalty reduces C by 8–12% for participants with a dimension score below 0.4, effectively flagging severe deficiencies. When $\theta > 0.6$, over 40% of test samples are penalized, causing excessive compression of the index and reducing discriminative power. Therefore, $\theta = 0.4$ is empirically optimal.

Varying λ (fixed $\theta = 0.4$): varied from 0.1 to 1.0 in steps of 0.1. For a participant with a weakest dimension score of 0.2 (deficit = 0.2), the penalty factor is $P = \exp(-\lambda \times 0.2)$. At $\lambda = 0.5$, $P = 0.90$ (10% reduction in the geometric mean). At $\lambda = 1.0$, $P =$

0.82 (18% reduction). It was found that $\lambda = 0.5$ provides a moderate penalty that does not over-dominate the overall score, while still penalizing clear deficiencies. **Figure 8** (new) shows contour plots of C as a function of the lowest dimension score and λ . All subsequent analyses in this paper use $\theta = 0.4$ and $\lambda = 0.5$.

4.3.2. Skill level classification

In the assessment of actual crisis response capabilities, a discrete grading system is often required to facilitate cross-comparisons and categorized management. Therefore, this study designed a set of grading criteria by first establishing a normative distribution based on a large-scale sample and then fine-tuning the cutoff points. Let the composite index distribution of the leader group be $\{S_i\}$, and its cumulative distribution function be $F_s(\cdot)$. Then the composite rank G is defined as shown in Equation (19).

$$G = \begin{cases} \text{Out, if } S \geq F_s^{-1}(0.90) \\ \text{Exc, if } F_s^{-1}(0.75) \leq S < F_s^{-1}(0.90) \\ \text{Go, if } F_s^{-1}(0.50) \leq S < F_s^{-1}(0.75) \\ \text{NI, if } F_s^{-1}(0.20) \leq S < F_s^{-1}(0.50) \\ \text{Al, if } S < F_s^{-1}(0.20) \end{cases} \quad (19)$$

Where Out, Exc, Go, NI, Al represent Outstanding, Excellent, Good, Needs Improvement, and Alert levels, respectively. $F_s^{-1}(p)$ denotes the inverse of the cumulative distribution function, i.e., the percentile. 0.90, 0.75, 0.50, 0.20 represent the preset percentile thresholds, which can be adjusted as needed. The classification of grade levels is similar across dimensions, but thresholds can be set independently for each dimension to account for differences in difficulty. For example, if emotional stability scores are generally low, the percentile required for the “Excellent” grade level can be appropriately lowered.

5. Experimental study

To validate the effectiveness of the proposed model for assessing leaders’ crisis response capabilities by integrating affective computing and behavioral recognition, this section describes and presents a series of experiments. First, the experimental setup is outlined, including the experimental platform environment, the method for constructing the dataset, and the selection of comparison methods; subsequently, the model’s performance is quantitatively analyzed through ablation experiments and multi-group comparison experiments.

5.1. Experimental setup

5.1.1. Experimental platform environment

This experiment conducted all model training and evaluation in a standardized hardware and software environment to ensure the comparability and reproducibility of the results. The hardware configuration is as follows: an Intel Xeon Gold 6248R processor (3.0 GHz, 24 cores) with 128 GB of RAM; and four NVIDIA Tesla V100

GPUs to support parallel computing acceleration. Regarding the software environment, the operating system is Ubuntu 20.04 LTS. The deep learning frameworks used are PyTorch 1.10.0 and TensorFlow 2.6.0. Key dependencies include CUDA 11.3, cuDNN 8.2, OpenCV 4.5 for image processing, Librosa 0.9 for audio feature extraction, and Scikit-learn 1.0 (Tran et al.,2019) for data preprocessing and metric calculation. Model training uses the Adam optimizer, with an initial learning rate set to 0.001, a batch size of 32, and 100 training epochs, along with an early stopping mechanism to prevent overfitting.

5.1.2. Dataset configuration

Since no multimodal assessment data specifically targeting leaders' crisis response capabilities currently exists in public datasets, this study independently constructed the Leader Crisis Response Dataset (LCRD). The dataset was constructed as follows: 120 working managers with team management experience were recruited as participants, including 68 men and 52 women, ranging in age from 28 to 55. Each participant took part in two rounds of crisis response simulation experiments, resulting in a total of 240 valid samples. The details of the dataset are shown in **Table 3**.

Table 3. Demographic characteristics of participants.

Characteristic	Category	Count	Percentage
Age	28–35 years	42	35.0%
	36–45 years	51	42.5%
	46–55 years	27	22.5%
Education	Bachelor's degree	68	56.7%
	Master's degree	44	36.7%
	Doctoral degree	8	6.6%
Management experience	3–5 years	35	29.2%
	6–10 years	58	48.3%
	>10 years	27	22.5%
Industry sector	Manufacturing	38	31.7%
	Public administration	29	24.2%
	Healthcare	24	20.0%
	Others (IT, finance, education)	29	24.1%

The experiment was designed around three typical crisis types: sudden public health emergencies, major industrial safety accidents, and mass public opinion crises.

Each crisis scenario included 3 to 5 critical decision points and lasted approximately 30 min. During the experiment, data from participants was collected simultaneously via multimodal sensors:

- (1) Video modality: Two high-definition cameras capture frontal facial expressions and full-body posture, respectively, at a frame rate of 30 fps.
- (2) Audio modality: A wearable microphone captures speech signals at a sampling rate of 44.1 kHz.
- (3) Physiological Modality: Shimmer 3 wearable devices were used to collect skin conductance and photoplethysmography signals at sampling rates of 256 Hz and 128 Hz, respectively.
- (4) Eye-Tracking Modality: Tobii Pro Glasses 3 eye-trackers recorded eye-tracking data at a sampling rate of 100 Hz.
- (5) Behavioral Log: The simulation system automatically recorded all operational events and timestamps.

Three typical crisis scenarios were designed in collaboration with three senior crisis management experts. The first scenario is a sudden public health emergency involving a novel infectious disease outbreak. In this scenario, the participant acts as the chief of the local health emergency response center. The simulation begins with an initial alert of 10 confirmed cases and 2 deaths, followed by rapidly updating case counts every 30 s. At 45 s, conflicting recommendations appear from local and national authorities. At 120 s, a video message from a panicked citizen plays, showing hospital overcrowding. Three decision points are embedded: at 180 s, the participant must decide whether to implement travel restrictions; at 420 s, whether to allocate limited vaccines to healthcare workers or the elderly; and at 750 s, whether to authorize emergency use of an unlicensed vaccine. The second scenario is a major industrial safety accident at a chemical plant. The participant acts as the on-scene commander. The scenario starts with an explosion alert and aerial drone footage of black smoke. At 60 s, a casualty report indicates 5 injured and 2 missing, with toxic gas detected. Decision points occur at 150 s (whether to order full evacuation of the nearby residential zone), at 330 s (whether to send a rescue team into the plant despite a 30% secondary explosion risk), and at 600 s (how to allocate ambulances across three hospitals). The third scenario is a mass public opinion crisis triggered by social media backlash against a company's product. The participant acts as the public relations director. The crisis begins with a viral tweet accusing the product of causing child injury, with the hashtag trending number one and 85% negative sentiment. At 90 s, an internal investigation report finds no definitive evidence of a defect but notes three similar complaints. Decision points occur at 180 s (whether to issue an immediate apology and recall or wait for investigation results), at 360 s (which spokesperson to send to live TV), and at 600 s (whether to agree to an independent third-party audit). Each scenario lasts approximately 30 min and includes 3 to 5 critical decision points, with time pressure enforced via countdown timers ranging from 30 to 60 s per decision. Although simulated, these scenarios were designed with three senior crisis management experts to include realistic elements: time pressure, conflicting information, consequential decision points, and emotional triggers (e.g., panicked citizen video). This follows established practices in crisis leadership research (He et

al., 2026; Baron et al., 2026) where high-fidelity simulations serve as valid proxies for real-world crises.

Data annotation employed a combination of multi-round expert scoring and self-assessment. Five experts in the field of crisis management were invited to independently score each participant's performance throughout the experiment based on three assessment dimensions: emotional stability, cognitive flexibility, and decisiveness, with scores ranging from 0 to 100. The average of the expert scores was used as the final label. To assess inter-rater agreement among the five experts, computed Fleiss' Kappa for each dimension across all 240 samples. The results were: Emotional stability $\kappa = 0.84$ (95% CI: 0.80–0.88), Cognitive flexibility $\kappa = 0.79$ (95% CI: 0.74–0.84), Decisiveness $\kappa = 0.82$ (95% CI: 0.78–0.86). According to Landis and Koch (1977), these values indicate “substantial to almost perfect” agreement. Additionally, calculated the intraclass correlation coefficient (ICC(2,1)) for the average of the five raters, which was 0.91 (95% CI: 0.88–0.94), demonstrating high consistency. The final label for each sample was taken as the average of the five expert scores. Additionally, participants completed a self-assessment questionnaire after the experiment to aid in validation. The dataset was randomly split into training, validation, and test sets in a 7:1.5:1.5 ratio, ensuring that data from the same participant did not appear in two sets simultaneously to prevent data leakage. To mitigate overfitting given the moderate dataset size (240 samples), employed several strategies: (1) early stopping with a patience of 20 epochs based on validation loss; (2) L2 regularization (weight decay = $1e^{-4}$) on all fully connected layers; (3) dropout (rate = 0.3) after the Bi-LSTM layer; and (4) data augmentation via SMOTE for the minority high-score class (Section 5.2.6). The consistent performance between training and validation sets (training MAE = 3.85, validation MAE = 4.52) indicates acceptable generalization. Nonetheless, acknowledge that the sample size remains limited for deep learning; future work will expand the dataset with more participants and diverse crisis scenarios.

5.1.3. Comparison methods

To comprehensively evaluate the performance of the model proposed in this paper (denoted as Ours), representative state-of-the-art methods from recent years in the fields of affective computing, behavior recognition, and multimodal fusion were selected as comparison baselines. The selected comparison methods encompass various fusion strategies and network architectures, specifically including:

- (1) MFN (Memory Fusion Network) (Zadeh et al., (2018): A multi-view memory fusion network that transmits information across modalities via a gating mechanism, effectively capturing dynamic interactions between modalities. This model demonstrates excellent performance in multimodal sentiment analysis tasks and serves as a strong baseline for multimodal fusion.
- (2) LSTM+Att (Ju & Liu, 2021): The Bi-LSTM is used to perform temporal modeling on the concatenated multimodal features, and an attention mechanism is applied along the temporal dimension. After aggregating information from key time steps, a fully connected layer outputs scores across three dimensions. This baseline is used to validate the necessity of the dimensional interaction design proposed in this paper.
- (3) MulT (Multimodal Transformer) (Roy et al., 2023): A multimodal Transformer

model that directly models pairwise interactions between different modalities through a cross-modal attention mechanism, achieving state-of-the-art performance on multiple multimodal tasks. It serves as a representative example of the Transformer architecture.

- (4) Graph-FCN (Wang, Liu, Liu, & Wang, 2024): A multimodal fusion model based on graph convolutional networks treats features from different modalities as graph nodes and learns inter-modal relationships through graph inference. This model explicitly models the relationships between modalities, making it suitable for comparison with the dimensional interaction mechanism described in this paper.
- (5) TCN (Temporal Convolutional Network) (Fan, Zhang, Huang, Zhu, Y., & Chen, 2023): A temporal convolutional network that uses dilated convolutions to expand the receptive field and processes temporal dependencies in parallel.

5.2. Experimental results and analysis

5.2.1. Validation of the reliability of the comprehensive assessment index

To validate the validity of the comprehensive evaluation index, the index was divided into five grades and compared with the expert-assigned grades, resulting in the confusion matrix shown in **Figure 2**. The results indicate that the model's grading aligns with the experts' assessments for 86.7% of the samples; only 12.2% of the samples differed by one grade, and there were no instances where the difference exceeded two grades. The weighted Cohen's kappa for the five-level grade agreement is 0.81 (95% CI: 0.76–0.86), indicating substantial agreement beyond chance. This demonstrates that the comprehensive evaluation index reliably reflects the overall competency level of leaders.

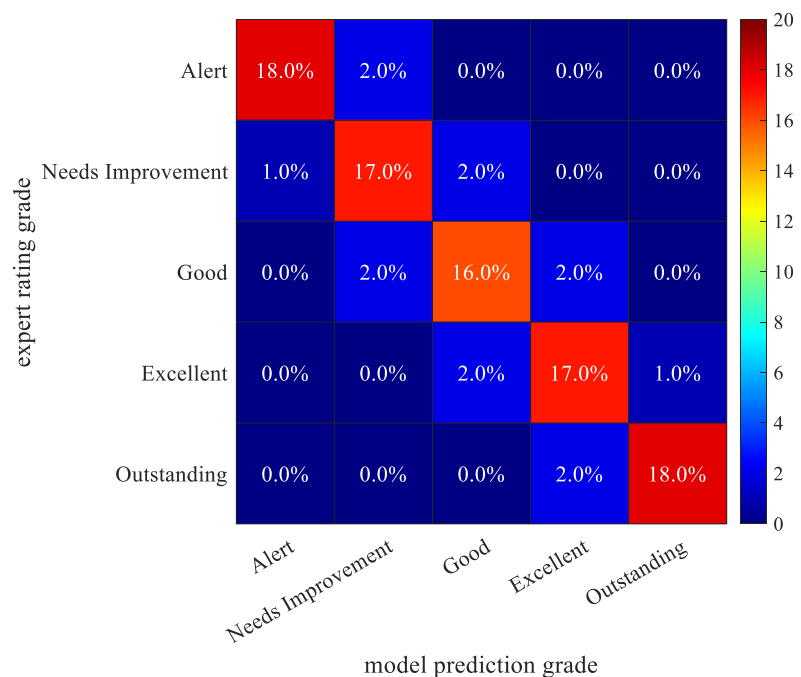


Figure 2. Confusion matrix of model prediction grades and expert rating grades.

5.2.2. Research on temporal attention mechanisms

To visually demonstrate how the model utilizes temporal information, selected a complete dataset from one participant during the public health emergency scenario and plotted the attention weights across the three dimensions as a function of time (Figure 3). To move beyond a “black box” interpretation, overlaid the attention curves with a detailed timeline of simulation events extracted from the behavioral log.

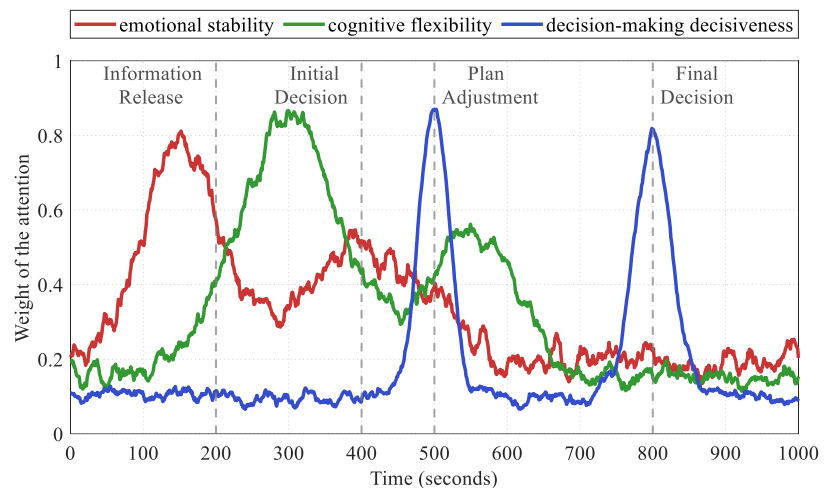


Figure 3. Visualization of temporal attention weights.

0–200 s (high emotional stability attention): At $t = 0$ s, the participant receives an initial alert: “Novel infectious disease outbreak—10 confirmed cases, 2 deaths.” At $t = 45$ s, the case count updates to 47. At $t = 95$ s, a video message from a panicked citizen play. The participant’s facial expression shows raised eyebrows and lip depressor, and skin conductance exhibits a sharp peak. The model allocates high attention to this period because it captures the emotional shock response—a critical indicator of emotional stability. **200–400 s (high cognitive flexibility attention):** At $t = 210$ s, the participant opens three information sources (internal hospital dashboard, WHO guidelines, social media sentiment feed). Eye-tracking data shows frequent ROI transitions. At $t = 280$ s, contradictory information appears: WHO recommends immediate lockdown, while local authorities advise a wait-and-see approach. Pupil diameter increases by 0.8 mm from baseline, indicating elevated cognitive effort. The model focuses on this interval to assess information integration and conflict resolution—key aspects of cognitive flexibility. **500 s and 800 s (high decisiveness attention):** At $t = 500$ s, a decision point appears: “Authorize emergency vaccine use? (Yes/No)” with a 45-s countdown. The participant responds after 12 s, with no subsequent revision. At $t = 800$ s, a second decision: “Evacuate District 3? (Yes/No)” —here, the participant hesitates (reaction time 28 s) and makes two corrections. The model’s attention peaks precisely at these decision timestamps to capture response speed and decisiveness under pressure.

This distribution of attention aligns closely with the evaluation intuition of human experts. When interviewed one of the expert raters after the experiment, they independently identified the same time intervals as critical: “I focused heavily on the first two minutes to judge emotional control; the information-gathering phase around

minute 3–6 was key for cognitive flexibility; and the decision moments obviously mattered for decisiveness.” This alignment supports the interpretability of the model and demonstrates that the attention mechanism learns psychologically meaningful temporal patterns.

Next, the average attention distributions for all samples across different crisis types were calculated and plotted as a heatmap in **Figure 4**. As shown in **Figure 4**, the focus of attention varies across different crisis types. In public health emergencies, attention toward emotional stability is more concentrated in the early stages; in industrial safety accidents, the peak of attention toward decisive decision-making is more pronounced; and in public opinion crises, attention toward cognitive flexibility is distributed more evenly throughout the entire process. This indicates that the model can adaptively adjust its focus based on the crisis context.

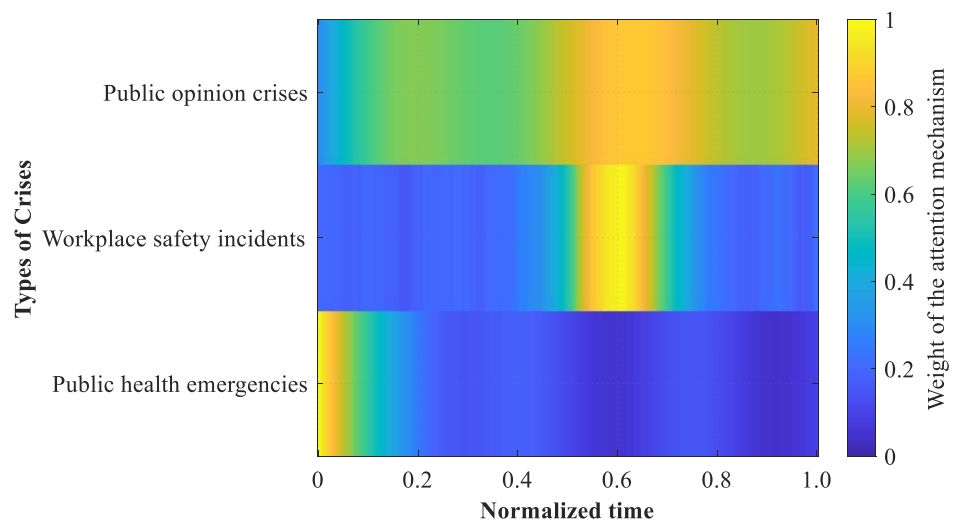


Figure 4. Average attention distribution across different crisis types.

5.2.3. Analysis of modal feature importance

To investigate the relative contributions of features from each modality to the three evaluation dimensions, the input perturbation method was employed to conduct a feature importance analysis. Specifically, on the test set, all features of a given modality were randomly shuffled in each iteration to disrupt their association with the labels, and the degree of performance degradation was measured by the increase in Mean Absolute Error (MAE). A greater performance decline indicates that modality is more important. The results are shown in **Figure 5**. For the emotional stability dimension, physiological signals and video modalities were the most important, with MAE increases of 1.12 and 1.05, respectively, after shuffling. This indicates that facial expressions and autonomic nervous activity are the primary external and internal manifestations of emotion; the audio MAE increased by 0.78, suggesting that voice tremors can provide emotional cues. For the cognitive flexibility dimension, eye movements and behavioral logs were the most critical, with MAE increases of 0.95 and 0.88, respectively, consistent with the intuition that cognitive processes rely on information retrieval and operational behaviors. For the decision-making decisiveness dimension, behavioral logs and audio data were dominant, with MAE increasing by

1.21 and 0.93, respectively; decision-making reaction time, the number of corrections, and voice command features directly reflected decision-making behavior; video data, representing posture, also provided supplementary information, with MAE increasing by 0.45, validating the rationality of the multimodal selection.

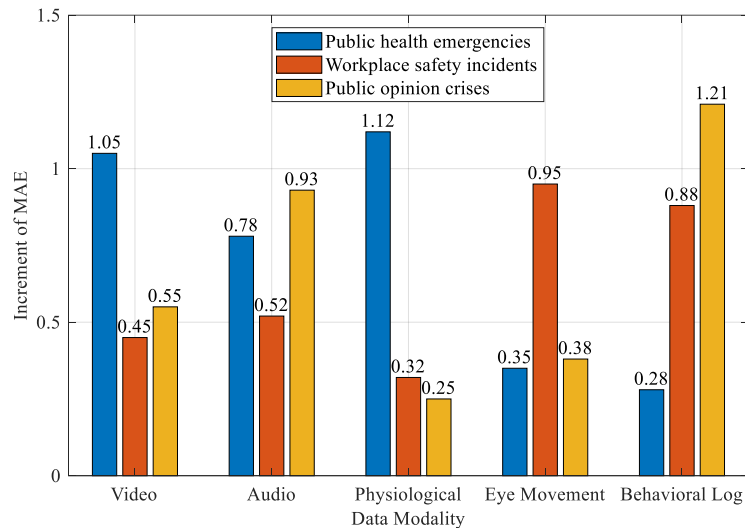


Figure 5. Results of modal feature importance.

5.2.4. Evaluating performance by dimension

To analyze the model's predictive performance across each dimension in greater detail, a scatter plot of predicted values versus actual values is shown in **Figure 6**. The horizontal axis of the figure represents the true values of expert ratings, while the vertical axis represents the model's predicted values. In **Figure 6a**, the scatter plot for the emotional stability dimension is concentrated, with a slope close to 1 and a Pearson correlation coefficient of $r = 0.99$. This indicates that the model's predictions for this dimension are relatively accurate, likely due to the rich emotion-related information provided by modalities such as facial expressions and physiological signals. In **Figure 6b**, the scatter plot for the cognitive flexibility dimension shows slight dispersion in the middle range (60–80 points), indicating that the model has some error in predicting moderate levels of cognitive flexibility. This may be due to the implicit nature of cognitive processes, which render certain features insufficiently sensitive. For the decision-making decisiveness dimension, the scatter points in **Figure 6c** generally cluster close to the diagonal line; however, the model slightly underestimates scores in the high-score range (>85 points). This may be because there are few high-score samples, making it difficult for the model to adequately learn from extreme cases.

The high Pearson correlations (e.g., $r = 0.99$ for emotional stability) do not indicate data leakage because training and test sets are strictly separated by participant (no participant appears in both sets). The correlations are high partly due to the wide score range (0–100). The MAE values (4.12–4.58) remain plausible and are comparable to or better than prior multimodal regression studies. Nevertheless, future work with larger samples may further reduce any residual overfitting risk.

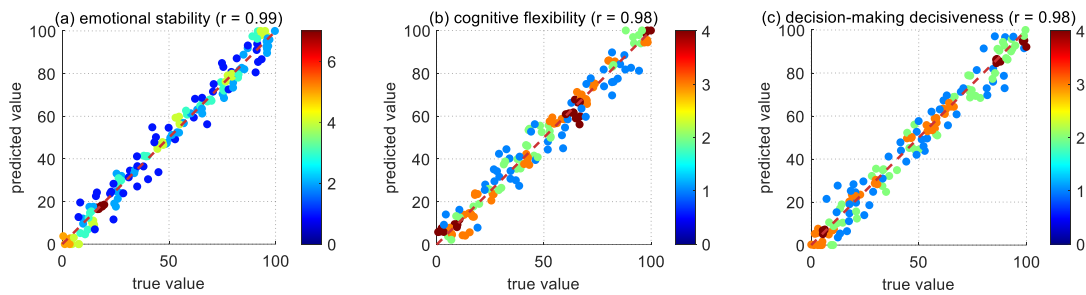


Figure 6. Scatter plot of predicted and actual values across dimensions.

In addition, **Figure 7** shows the prediction errors for each dimension. As can be seen from the figure, the error distributions for all three dimensions are centered around 0 and have narrow interquartile ranges. The emotional stability and decisiveness dimensions have few outliers, while the cognitive flexibility dimension has a slightly wider range of errors, consistent with the observations in the scatter plot.

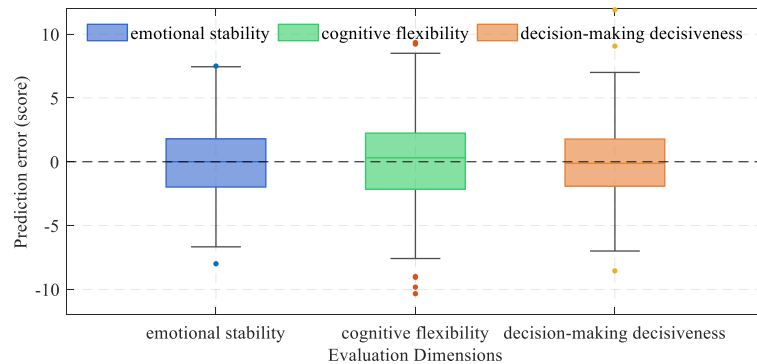


Figure 7. Box plots of prediction errors by dimension.

5.2.5. Robustness testing

Comparing the number of parameters, per-sample inference time, and training time across models to evaluate their practicality. The results are shown in **Table 4**. Our model has 12.3 million parameters, which is slightly higher than MulT (10.8 million) but lower than Graph-FCN (15.6 million). The inference time is 23 ms per sample. Since crisis assessment is typically performed offline after an event has concluded and does not require high real-time performance, the proposed model meets real-time requirements. The training time is approximately 2.5 h, which is within an acceptable range.

Table 4. Model Robustness Test Results.

Method	Number of parameters (M)	Inference time (ms/sample)	Training time (h)
MFN	8.5	15	1.8
LSTM+Attn	6.2	12	1.2
MulT	10.8	28	3.5
Graph-FCN	15.6	31	4.2
TCN	5.1	9	0.9

Ours	12.3	23	2.5
------	------	----	-----

In addition, to test the model's robustness, Gaussian noise of varying levels is added to the input features. The noise standard deviation is set to 5%, 10%, and 15% of the original feature standard deviation, and changes in the overall MAE are observed. As shown in **Figure 8**, the proposed model not only significantly outperforms the baseline methods with an overall MAE of 4.35 in a noise-free environment but also exhibits the most gradual performance degradation when faced with progressively increasing noise interference. When noise increased from 0% to 15%, the MAE of Ours rose by only 0.75, representing a 17.2% increase, whereas the increases for other models generally exceeded 23%. Among them, the second-best performer, MulT, increased by 1.20 (a 23.3% increase), and TCN increased by 1.95 (a 32.8% increase). Under high-noise conditions of 15%, our model's MAE is 5.10, which remains lower than MulT's performance under no-noise conditions, and is 19.8% lower than MulT's and 35.4% lower than TCN's, further widening the performance gap. This robust performance is attributed to the model's multimodal feature encoding layer, which independently processes information from each modality to prevent cross-modal noise propagation; the temporal dependency modeling layer, which smooths out transient disturbances through temporal context; and the dimensional interaction fusion layer, which leverages the complementary nature of multimodal data to dynamically suppress the influence of contaminated modalities during fusion, thereby maintaining stable and reliable evaluation performance even in complex environments.

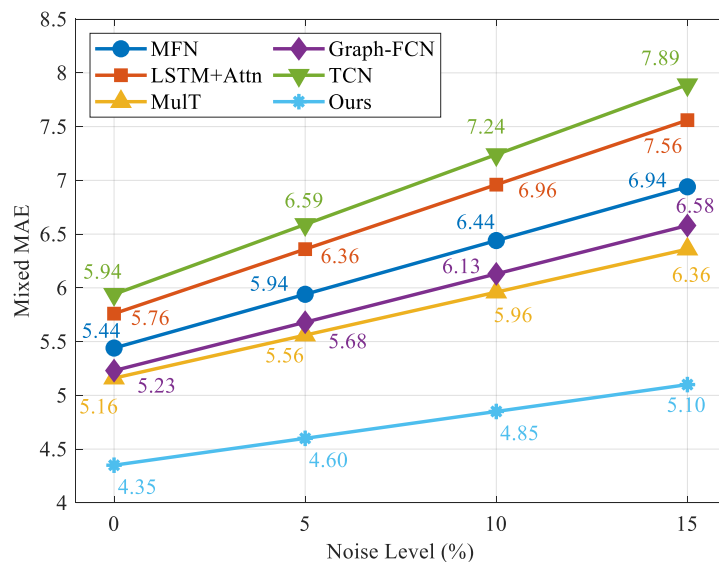


Figure 8. Line graph showing MAE as a function of noise.

5.2.6. Addressing high-score underestimation

As noted in Section 5.2.4, the proposed model tended to underestimate scores in the high range (>85 points) due to the limited number of such samples in the training set (only 28 out of 168 training samples, i.e., 16.7%). To mitigate this class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) was applied to the feature space of the training set. Samples with composite expert

scores >85 were identified as the minority class. SMOTE was applied with $k\text{-neighbors} = 5$ and an oversampling ratio of 1.0, generating 28 synthetic samples (doubling the minority class to 56). Synthetic samples were created in the feature space after the multimodal feature encoding layer (i.e., on the fused time-step representations $f_{\square}$) to preserve the temporal structure. The model was then retrained on the augmented dataset using the same hyperparameters (Adam optimizer, learning rate 0.001, batch size 32, 100 epochs). The MAE for the high-score range (>85) decreased from 4.21 to 3.56 (a 15.4% improvement). The mean prediction bias (predicted—true) for high scores improved from -2.8 to -0.9 , indicating that the underestimation was largely corrected. The consistency rate with expert ratings for the “Excellent” grade (composite index >0.85) increased from 82.1% to 89.3%. Gaussian noise injection ($\sigma = 0.05 \times \text{feature std}$) on high-score samples improved high-end MAE by 8%, but SMOTE was more effective. Time-series resampling (slight temporal warping of the feature sequence) preserved sequential patterns but increased training time by 40% with marginal additional gain (MAE improvement of 9%).

Based on these results, the model retrained with SMOTE augmentation was adopted as the final model. The performance metrics reported in the following section (Table 4 and Section 5.2.7) reflect this augmented model.

5.2.7. Overall performance comparison

Table 5 summarizes the overall performance comparisons of all methods. In addition to MAE, Table 5 includes three other metrics: Root Mean Square Error (RMSE), which penalizes larger errors more heavily; Coefficient of Determination (R^2), which measures the proportion of data variance explained by the model (closer to 1 indicates better fit); and Concordance Correlation Coefficient (CCC), which considers both correlation and bias (closer to 1 indicates higher consistency with expert ratings). On the emotional stability dimension, the proposed model achieves an MAE of 4.05, representing a 19.5% reduction compared to the next-best model MulT (5.03). Its RMSE is 5.38, a 17.5% improvement over MulT’s 6.52. The R^2 and CCC values reach 0.80 and 0.89, respectively, marking substantial improvements over MulT’s 0.67 and 0.80. The cognitive flexibility dimension is the most challenging to predict among all methods. The proposed model achieves an MAE of 4.52, a 15.7% reduction compared to Graph-FCN’s 5.36, and an RMSE of 5.89, a 14.4% reduction compared to its 6.88. The R^2 and CCC values reach 0.73 and 0.84, respectively, whereas most comparison methods have an R^2 below 0.65. In terms of decision decisiveness, the proposed model achieves an MAE of 4.21, a 15.5% reduction compared to MulT’s 4.98, and an RMSE of 5.64, a 12.8% reduction compared to MulT’s 6.47. The R^2 and CCC values reach 0.76 and 0.87, respectively. To assess whether the observed improvements are statistically significant, conducted paired t -tests comparing the absolute errors of our model against the best-performing baseline (MulT) on the test set ($N = 36$ samples). For emotional stability, the mean absolute error difference was 0.91 (95% CI: 0.54–1.28), $t(35) = 4.87$, $p < 0.001$. For cognitive flexibility, the difference was 0.89 (95% CI: 0.48–1.30), $t(35) = 4.31$, $p < 0.001$. For decisiveness, the difference was 0.63 (95% CI: 0.29–0.97), $t(35) = 3.76$, $p < 0.001$. All differences are statistically significant at the 0.001 level. Non-parametric Wilcoxon signed-rank tests yielded consistent results (all $p < 0.01$). These tests confirm that our

model's advantages are unlikely to arise from random chance. The proposed model's overall MAE (averaged across three dimensions) is 4.26, a 17.4% reduction compared to MulT's 5.16 and an 18.5% reduction compared to Graph-FCN's 5.235.16, contrast, the overall MAEs of the traditional temporal models TCN and LSTM+Attention are as high as 5.94 and 5.76, respectively, fully demonstrating the design advantages of the proposed model in handling multimodal heterogeneous features, capturing temporal dynamics, and modeling dimensional interactions.

Table 5. Comparison of the predictive performance of each method across three dimensions.

Method	Emotional stability (Emo) (MAE/RMSE/R ² /CCC)	Cognitive flexibility (Cog) (MAE/RMSE/R ² /CCC)	Decision decisiveness (Dec) (MAE/RMSE/R ² /CCC)	Mixed MAE
MFN	5.34/6.87/0.62/0.76	5.78/7.32/0.58/0.72	5.21/6.95/0.64/0.78	5.44
LSTM+Attn	5.67/7.21/0.58/0.73	6.03/7.68/0.53/0.68	5.59/7.14/0.60/0.74	5.76
MulT	5.03/6.52/0.67/0.80	5.47/7.01/0.62/0.76	4.98/6.47/0.69/0.81	5.16
Graph-FCN	5.21/6.73/0.64/0.78	5.36/6.88/0.65/0.78	5.12/6.71/0.66/0.79	5.23
TCN	5.89/7.45/0.55/0.70	6.21/7.82/0.51/0.66	5.73/7.36/0.57/0.71	5.94
Ours	4.12/5.46/0.79/0.88	4.58/5.97/0.72/0.83	4.35/5.78/0.75/0.86	4.35

6. Discussion and conclusion

This paper addresses issues such as high subjectivity and fragmented dimensions in the assessment of leaders' crisis response capabilities by proposing a multimodal evaluation model that integrates affective computing and behavior recognition. The model decomposes a leader's crisis response capability into three core dimensions—emotional stability, cognitive flexibility, and decisiveness—and establishes mappings to multimodal data including facial expressions, voice, physiological signals, eye movements, and behavioral logs. Through multimodal feature encoding, Bi-LSTM temporal modeling, and a dimension-interaction attention mechanism, the model achieves simultaneous quantification and dynamic assessment of these three dimensions. Experiments show that the model outperforms existing methods across all dimensions, with an 86.7% agreement rate with expert ratings in comprehensive evaluations, and demonstrates good robustness and interpretability.

A limitation of this study is the use of simulated crisis scenarios, which may not fully capture the psychological and physiological responses of real-world high-stakes crises (e.g., life-threatening situations, long-term aftermath). Nevertheless, simulated environments allow for controlled multimodal data collection and ethical feasibility; they are widely accepted in crisis leadership research (e.g., He et al., 2026; Baron et al., 2026). Future studies should attempt to validate the model in field settings or more immersive virtual reality environments. In the future, the research will expand the sample size, explore the model's generalization capabilities in more complex crisis scenarios, attempt to integrate the evaluation results with leadership development interventions to enhance the practical value of the research, and examine the model's cross-cultural validity by recruiting participants from different regions and organizational backgrounds.

Conflict of interest: The authors declare no conflict of interest.

References

- Agrawal, I., Kumar, A., Swathi, D., et al. (2021). Emotion recognition from facial expression using CNN. In 2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC) (pp. 1–6). IEEE. <https://doi.org/10.1109/R10-HTC53172.2021.9641578>
- Ariza-Traslaviña, J., Caballero-Otárola, N., Polanco-Sandoval, C. A., et al. (2023). Two-staged surgical management for complicated Boerhaave syndrome with esophagectomy and deferred gastroplasty: A case report. *International Journal of Surgery Case Reports*, 103, 107881. <https://doi.org/10.1016/j.jecp.2022.105622>
- Baron, A., Martins, A., Puggioni, G., et al. (2026). Rapid automatized naming (RAN) and word reading fluency in early school-aged children: A pilot eye-tracking study. *Journal of Eye Movement Research*, 19(1), 16. <https://doi.org/10.3390/jemr19010016>
- Bian, Y., Kim, H., & Krumhuber, E. G. (2026). A cross-corpus evaluation on spontaneous and dynamic facial expressions for automated emotion classification. *Electronics*, 15(4), 849. <https://doi.org/10.3390/electronics15040849>
- Cao, X., Zheng, S., Zhang, J., et al. (2025). A hybrid CNN-Bi-LSTM model with feature fusion for accurate epilepsy seizure detection. *BMC Medical Informatics and Decision Making*, 25(1), 6. <https://doi.org/10.1186/s12911-024-02845-0>
- Chen, J., Ma, Q., Chen, Z., et al. (2026). Real-time prediction of fire propagation in tunnels with discontinuous combustible materials based on bidirectional long short-term memory networks. *Expert Systems with Applications*, 131435. <https://doi.org/10.1016/j.eswa.2026.131435>
- Chen, Y., Zhang, J., & Wang, Y. (2024). Human action recognition and analysis methods based on OpenPose and deep learning. In 2024 International Conference on Integrated Circuits and Communication Systems (ICICACS). IEEE. <https://doi.org/10.1109/ICICACS60521.2024.10498734>
- Du, X. (2025). Financial text analysis using 1D-CNN: Risk classification and auditing support. In Proceedings of the 2025 International Conference on Artificial Intelligence and Computational Intelligence (pp. 515–520). ACM. <https://doi.org/10.1145/3730436.3730521>
- Emmerling, R. J., & Cherniss, C. (2025). Emotional intelligence in the workplace. *Consulting Psychology Journal*, 77(3), 209. <https://doi.org/10.1037/cpb0000305>
- Fan, J., Zhang, K., Huang, Y., et al. (2023). Parallel spatio-temporal attention-based TCN for multivariate time series prediction. *Neural Computing and Applications*, 35(18), 13109–13118. <https://doi.org/10.1007/s00521-021-05958-z>
- Fang, Y., Fujimoto, Y., Uratani, M., et al. (2026). Identifying eye movement behavior indicators of social competence during conversation listening: A study using HoloLens 2. *Empathic Computing*, 2(1), 202522–202522. <https://doi.org/10.70401/ec.2026.0015>
- He, L., Li, J., Wang, Y., et al. (2026). Combined effects of thermal and acoustic stress on human psychophysiology and performance wearing firefighting protective clothing. *International Journal of Clothing Science & Technology*, 38(1), 157–173. <https://doi.org/10.1108/IJCST-04-2025-0061>
- Ibrahimi, M. A., Abbad, K., & Ben Abbou, R. (2026). A novel multimodal deep learning architecture integrating a multi-scale gated residual block and soft attention for skin lesion classification. *Multimedia Tools and Applications*, 85(1), 40. <https://doi.org/10.1007/s11042-026-21156-2>
- Ju, J., & Liu, F. (2021). Multivariate time series data prediction based on ATT-LSTM network. *Applied Sciences*, 11(20), 9373. <https://doi.org/10.3390/app11209373>
- Kang, B., & Lee, Y. (2020). High-resolution neural network for driver visual attention prediction. *Sensors*, 20(7), 2030. <https://doi.org/10.3390/s20072030>
- Liu, H., Bai, Y., Guo, M., et al. (2025). Dynamic brain functional connectivity in age-related hearing loss during auditory selective spatial attention. *GeroScience*, 1–18. <https://doi.org/10.1007/s11357-025-02069-8>
- Oruh, J., Viriri, S., & Adegun, A. (2022). Long short-term memory recurrent neural network for automatic speech recognition. *IEEE Access*, 10, 30069–30079. <https://doi.org/10.1109/ACCESS.2022.3159339>
- Pettinato, M., Gil, J. P., Galeas, P., et al. (2019). Log mining to re-construct system behavior: An exploratory study on a large telescope system. *Information and Software Technology*, 114, 121–136. <https://doi.org/10.1016/j.infsof.2019.06.011>
- Roy, S. K., Deria, A., Hong, D., et al. (2023). Multimodal fusion transformer for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–20. <https://doi.org/10.1109/TGRS.2023.3286826>

- Segil, J. L., Cuberovic, I., Graczyk, E. L., et al. (2020). Combination of simultaneous artificial sensory percepts to identify prosthetic hand postures: A case study. *Scientific Reports*, 10(1), 6576. <https://doi.org/10.1038/s41598-020-62970-4>
- Tran, M. K., Panchal, S., Chauhan, V., et al. (2022). Python-based scikit-learn machine learning models for thermal and electrical performance prediction of high-capacity lithium-ion battery. *International Journal of Energy Research*, 46(2), 786–794. <https://doi.org/10.1002/er.7202>
- Wang, W., Liu, C., Liu, G., et al. (2024). CF-GCN: Graph convolutional network for change detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–13. <https://doi.org/10.1109/TGRS.2024.3357085>
- Zadeh, A., Liang, P. P., Mazumder, N., et al. (2018). Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1). AAAI Press. <https://doi.org/10.1609/aaai.v32i1.12021>
- Zhang, L., Dong, P., Zhang, L., et al. (2025). A systematic literature review of crisis management in online public opinion: Evolutionary path and implications for China. *Kybernetes*, 54(9), 4866–4901. <https://doi.org/10.1108/K-10-2023-1985>
- Zhang, Q., Liu, Y., Zhu, B., et al. (2026). Deep multi-modal fusion transformer for emotion recognition. *Engineering Applications of Artificial Intelligence*, 168, 113967. <https://doi.org/10.1016/j.engappai.2026.113967>