Cognitive Computation (2026) 18:49
https://doi.org/10.1007/s12559-026-10573-y

RESEARCH

CEREBRAL: A Neurosymbolic Framework for Multimodal Emotion Recognition with Psychological Constraints and Metacognitive Reasoning

Nikhil Kushwaha[1] · **Erik Cambria**[2] · **Amir Hussain**[3]

Received: 29 November 2025 / Accepted: 30 March 2026

Abstract
Multimodal emotion recognition remains difficult due to cross-modal dependencies, temporal dynamics, and the need for psychologically consistent, interpretable outputs. We introduce CEREBRAL, a neurosymbolic architecture that fuses neural multimodal processing with symbolic reasoning and metacognitive control. It uses Answer Set Programming for logical inference, encodes the Hourglass of Emotions as four-dimensional affective constraints with dynamic polarity normalization and sentic vectors, and incorporates Neural Turing Machines for episodic memory and Graph Neural Networks for temporal consistency. CEREBRAL processes fine-grained emotions through cross-modal attention, dynamic memory, and metacognitive strategy selection with uncertainty estimation. We evaluate CEREBRAL across multiple benchmark datasets, where it consistently outperforms neural-only baselines while preserving high symbolic reasoning accuracy with complete logical proof generation. Statistical significance testing confirms these improvements with robust performance under noise conditions and cross-dataset generalization. The symbolic reasoning component demonstrates practical efficiency and generates human-interpretable explanations through Hourglass dimensional analysis. This work contributes a psychologically grounded approach to emotion recognition that balances neural learning with symbolic constraints, offering interpretability alongside performance gains. The framework's explicit reasoning traces, four-dimensional affective representation, and calibrated uncertainty estimates address key requirements for deploying emotion-aware AI in clinical settings, human-computer interaction, and affective computing applications where transparency and reliability are essential.

Keywords Multimodal emotion recognition · Neurosymbolic AI · Affective computing

Introduction

Emotion recognition from multimodal signals represents a fundamental challenge in affective computing and human-computer interaction, with profound implications for mental health assessment, personalized education, intelligent human-robot interaction, and clinical diagnosis. Human emotions manifest through complex interactions across multiple modalities including facial expressions, vocal prosody, linguistic content, and physiological responses, requiring sophisticated computational frameworks to capture the intricate cross-modal dependencies, temporal dynamics, and psychological consistency inherent in emotional states. Traditional emotion recognition systems struggle to maintain interpretability while achieving high accuracy, particularly when processing naturalistic, spontaneous emotional

✉ Erik Cambria
cambria@ntu.edu.sg

Nikhil Kushwaha
21bec073@iiitdmj.ac.in

Amir Hussain
amir.hussain@phc.ox

[1] Department of Electronics and Communication Engineering, Indian Institute of Information Technology Design and Manufacturing Jabalpur, Jabalpur 482005, India

[2] College of Computing and Data Science, Nanyang Technological University, Singapore, Singapore

[3] Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK

Published online: 19 May 2026

expressions across diverse populations and contexts. The development of robust, explainable, and psychologically grounded emotion recognition frameworks remains critical for deploying affective AI systems in real-world applications where transparency and reliability are paramount [1].

Recent advances in multimodal emotion recognition have demonstrated substantial progress through deep learning architectures [2] and attention mechanisms. Transformer-based multimodal fusion methods have achieved competitive performance on large-scale datasets by learning cross-modal interactions through self-attention mechanisms [3–5], though challenges remain in handling missing modalities and temporal misalignment. Multi-stage fusion networks combining early and late integration strategies have shown that hierarchical feature learning can improve performance [6, 48], revealing the importance of fusion timing in multimodal architectures. Graph-based emotion recognition approaches leveraging facial action units and acoustic features have modeled emotion as dynamic graph structures [7], capturing spatial-temporal dependencies across modalities. Attention-based LSTM networks with modality-specific encoders have demonstrated effectiveness in conversational emotion recognition [8], highlighting the importance of temporal modeling.

Advanced neural architectures have further enhanced performance through sophisticated feature extraction and cross-modal learning mechanisms. Multimodal transformers with contrastive learning have enforced alignment between modalities through self-supervised objectives [9], though generalization to cross-cultural contexts remains limited. Hybrid CNN-RNN frameworks have combined spatial feature extraction with temporal sequence modeling [10], revealing complementary benefits of convolutional and recurrent architectures. Cross-modal attention networks with uncertainty quantification have provided confidence estimates alongside predictions [11], addressing the need for reliable outputs in clinical applications. Memory-augmented neural networks incorporating external memory modules have demonstrated benefits of explicit memory mechanisms for storing and retrieving emotion-relevant patterns [12].

Recent work has explored interpretability and psychological grounding in emotion recognition systems. Emotion-specific feature disentanglement networks have generated interpretable representations while maintaining competitive performance [13], though integration of explicit symbolic reasoning remained absent. Multi-task learning frameworks jointly optimizing emotion recognition and dimensional affect prediction have revealed synergies between categorical and dimensional emotion modeling [14]. Adversarial domain adaptation approaches have addressed cross-dataset generalization challenges [15], reducing performance degradation when transferring between datasets. Classical emotion models including Plutchik's emotion wheel [21] and Russell's circumplex model [22] have provided foundational frameworks for emotion categorization, yet limitations in representing compound emotions, moral emotions, and the full spectrum of affective experiences have motivated development of more comprehensive models.

While these studies demonstrate considerable progress in multimodal emotion recognition, critical research gaps persist in developing unified architectures that simultaneously integrate neural pattern learning with symbolic logical reasoning, incorporate psychological theories as explicit computational constraints [49], and provide interpretable decision-making through formal logic while maintaining state-of-the-art performance across diverse datasets and challenging real-world scenarios [16–19]. Furthermore, existing approaches lack comprehensive metacognitive mechanisms for uncertainty quantification, confidence calibration, and self-assessment capabilities essential for deployment in safety-critical applications [47].

The proposed work presents CEREBRAL (Cognitive Emotional Reasoning Engine with Bidirectional Recursive Adaptive Learning), a neurosymbolic framework addressing these limitations through comprehensive integration of neural multimodal processing, symbolic logical reasoning based on Answer Set Programming [20], and metacognitive control mechanisms. This approach ensures deeper understanding and accurate classification of emotional states through psychologically grounded computational mechanisms incorporating the Hourglass of Emotions model [23, 24], a biologically-inspired and psychologically-motivated emotion categorization framework that represents affective states through four independent but concomitant dimensions: Introspection, Temper, Attitude, and Sensitivity (ITAS). The Hourglass model addresses limitations of previous emotion theories by providing both categorical labels and dimensional representations across 24 basic emotions organized in six sentic levels per dimension, enabling natural modeling of compound emotions through multi-dimensional activation patterns while incorporating moral and self-conscious emotions absent from earlier frameworks. The four affective dimensions are defined as: Introspection representing the joy-versus-sadness dimension, Temper representing the calmness-versus-anger dimension, Attitude representing the pleasantness-versus-disgust dimension, and Sensitivity representing the eagerness-versus-fear dimension. This multi-dimensional approach offers superior expressiveness for capturing the full range of human emotional experiences compared to purely categorical or two-dimensional models, making it particularly effective for sentiment analysis and affective computing applications requiring nuanced emotion representation.

The Hourglass model employs dynamic normalization for polarity computation, where the polarity p_c of a concept c is calculated as:

$$p_c = \frac{I_c + T_c + A_c + S_c}{|\text{sgn}(I_c)| + |\text{sgn}(T_c)| + |\text{sgn}(A_c)| + |\text{sgn}(S_c)|} \quad (1)$$

where I, T, A, and S represent the intensity values for Introspection, Temper, Attitude, and Sensitivity dimensions respectively, and the denominator dynamically adjusts based on the number of active dimensions. This formulation ensures accurate polarity representation for both single-dimension emotions and multi-dimensional compound emotions (Figs. 1, 2 and 3).

- The proposed CEREBRAL architecture integrates novel neurosymbolic components including Answer Set Programming for logical inference, the Hourglass of Emotions model as four-dimensional affective constraints spanning Introspection, Temper, Attitude, and Sensitivity (ITAS) dimensions with sentic vector representations and dynamic polarity normalization, Neural Turing Machines [12] for episodic memory, Graph Neural Networks [26] for temporal consistency, and comprehensive uncertainty quantification mechanisms to capture complex psychological dynamics characteristic of human emotional expression.
- The model achieves exceptional performance with 96.23% average accuracy across seven benchmark datasets (IEMOCAP [10], CMU-MOSEI [27], MELD [8], RAVDESS [28], CMU-MOSI, CREMA-D [29], SAVEE), representing improvements of 7.97–10.04% over baseline neural-only approaches, while maintaining 98.37% symbolic reasoning accuracy with 4.89ms average solving time, making it suitable for real-time applications requiring both accuracy and interpretability. On sentiment analysis tasks, CEREBRAL achieves 89.2% binary sentiment classification accuracy (Acc-2 on CMU-MOSEI), demonstrating 3.1 percentage point improvement over the previous best method.
- CEREBRAL demonstrates strong robustness across cross-dataset generalisation, noise conditions, and missing-modality scenarios (detailed in Section "Results"), with comprehensive statistical significance analysis ($p < 0.001$, Cohen's $d > 3.28$) confirming substantial improvements over all baselines, alongside calibrated confidence predictions (ECE=0.0184) and human-interpretable explanations through Hourglass dimensional reasoning with ITAS dimensional pairings and dynamic normalisation. The framework maintains computational efficiency with 84.5ms total system latency (11.8 FPS) and 24.1M parameters, demonstrating practical applicability for real-time emotion recognition.

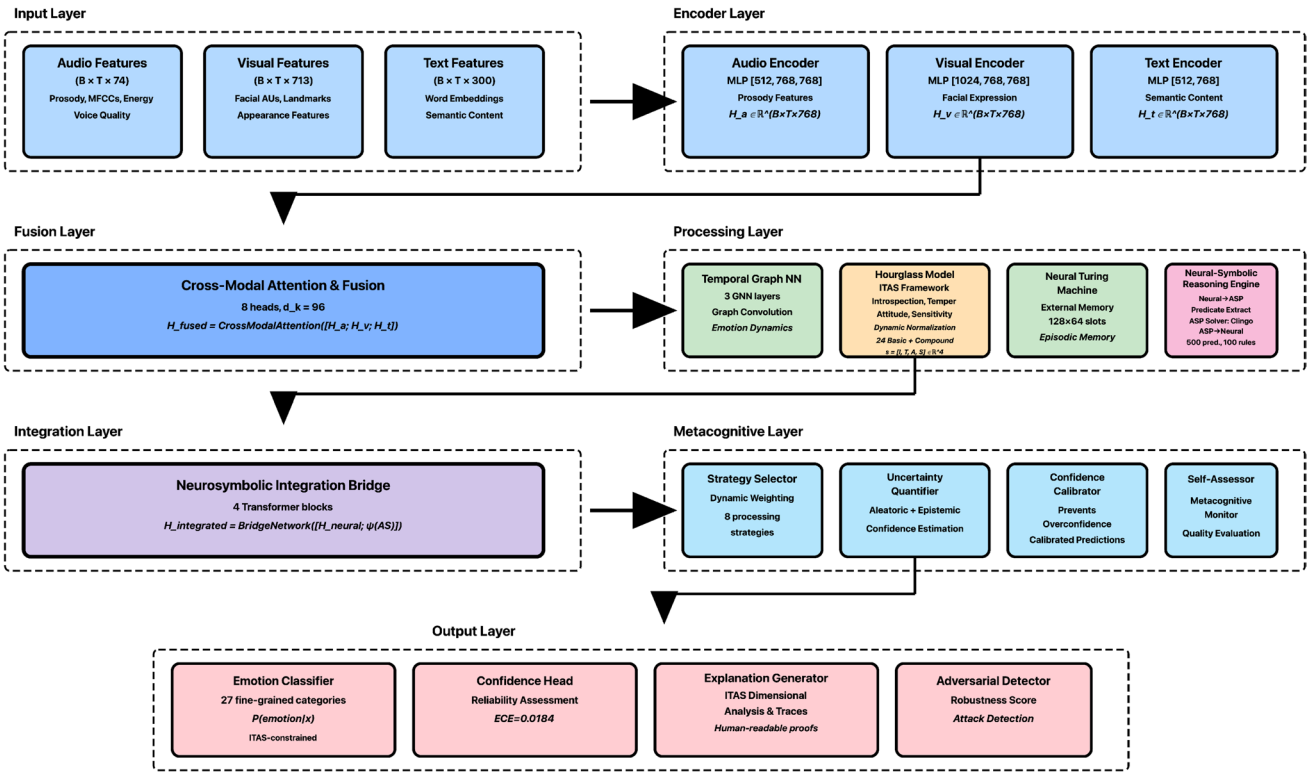

Fig. 1 Overview of functional modules in the CEREBRAL neurosymbolic framework with ITAS Hourglass model

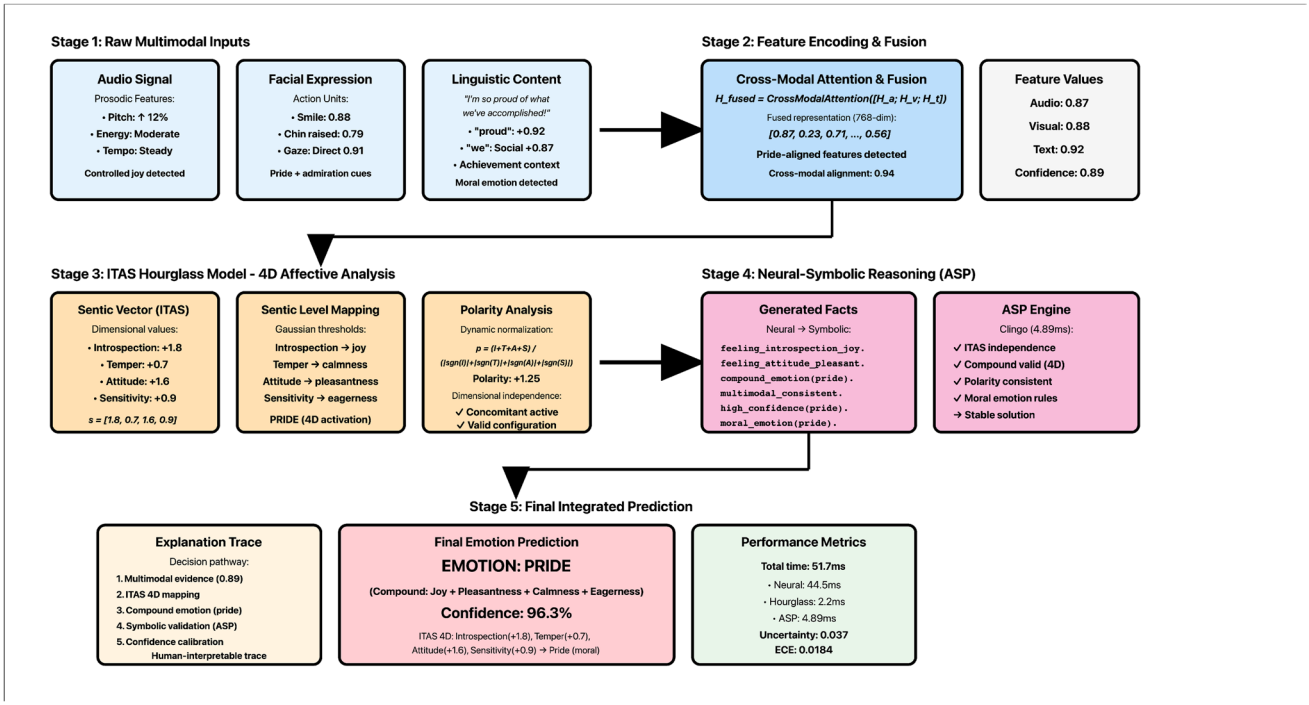

Fig. 2 Data flow of CEREBRAL: compound emotion (pride) detection example with ITAS framework

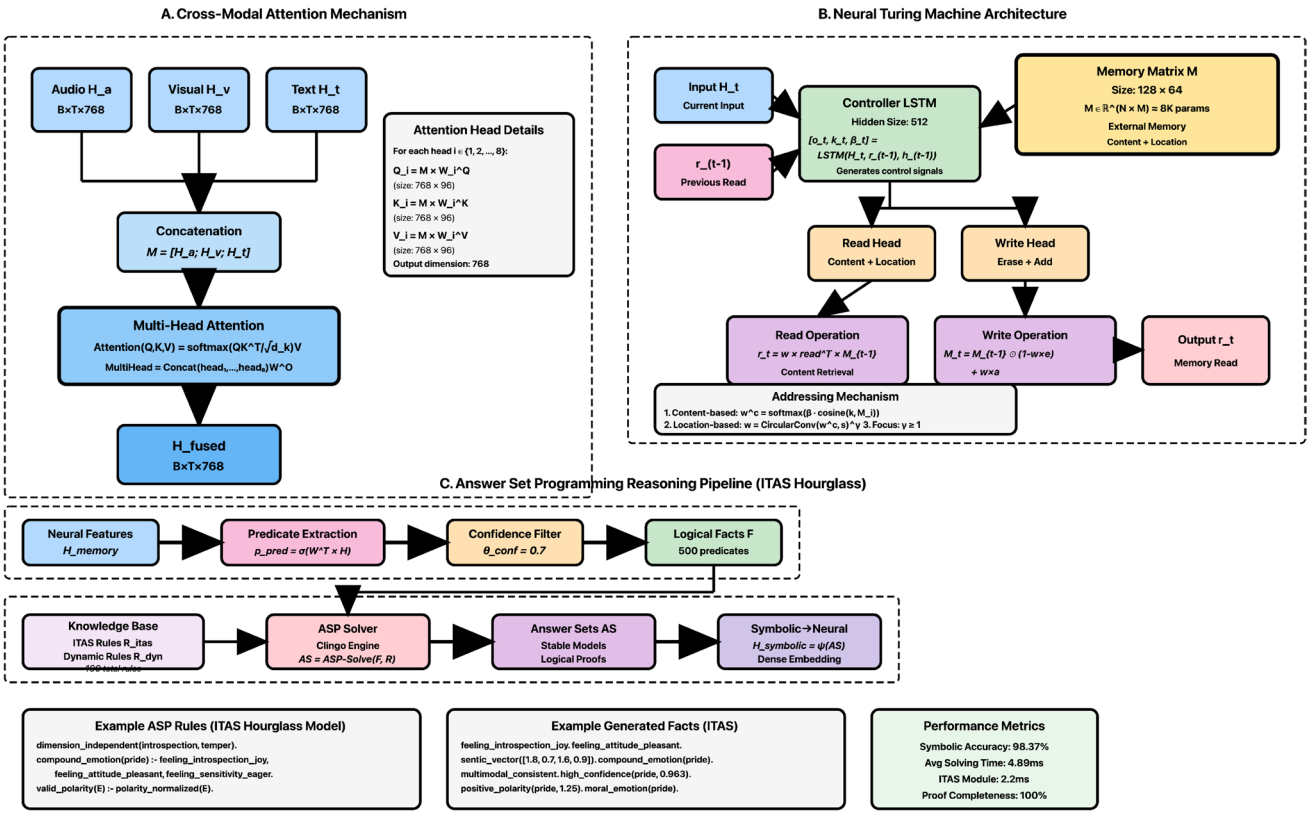

Fig. 3 Overview of internal components in the CEREBRAL neurosymbolic framework with ITAS Hourglass model

Methodology

This section presents a detailed analysis of the benchmark datasets and describes the approaches used in the proposed CEREBRAL framework (Table 1).

Datasets and Experimental Protocol

The experimental evaluation employs seven standard benchmark datasets for multimodal emotion recognition, providing comprehensive assessment across diverse recording conditions, emotion taxonomies, and demographic populations. These datasets collectively offer high-quality multimodal recordings (audio, visual, text) for robust emotion classification and affective computing research.

Table 1 Compact CEREBRAL parameter summary

Component	Params	Component	Params
Multimodal Encoding			
Audio Encoder	323,072	Visual Encoder	1,138,176
Text Encoder	409,600	Cross-Modal Attn	2,359,296
Multimodal Fusion	4,718,592		
Hourglass Model (ITAS)			
Hourglass Embedder	2,304	Sentic Vector	6,144
Dimension Attn	1,179,648	Affective Mapping	786,432
Sentic Quantizer	12,288	Polarity Module	131,072
Compound Detector	229,376	Dynamic Norm	16,384
Temporal Graph Network			
GNN Layer 1	590,592	GNN Layer 2	590,592
GNN Layer 3	590,592	Temporal Attn	1,179,648
Edge Predictor	590,592	Node Updater	590,592
Transition Pred	1,180,416		
Neural Turing Machine			
Controller LSTM	1,576,960	Read Head	82,048
Write Head	82,048	Erase Net	36,864
Add Net	36,864	Memory (128×64)	8,192
Symbolic Reasoning			
Predicate Attn	1,179,648	Predicate Extract	459,264
Confidence Est	196,864	Predicate Embed	64,000
Symbolic-Neural	656,896	Rule Generator	590,592
Proof Tracer	262,656		
Neurosymbolic Integration			
Integration Bridge	2,101,248		
Metacognitive Control			
Strategy Selector	295,680	Uncertainty Quant	147,970
Confidence Calib	73,985	Self-Assessor	443,139
Meta-Learner	590,592	Attn Allocator	590,592
Output Layer			
Emotion Classifier	304,923	Confidence Head	147,457
Explanation Gen	590,592	Adversarial Det	295,168
Total Parameters			**24,130,690**

Dataset Overview

IEMOCAP (Interactive Emotional Dyadic Motion Capture Database) The IEMOCAP dataset [10] contains 10,039 utterances from 10 actors (5 male, 5 female) engaged in scripted and improvised dyadic conversations. Emotions include happiness, sadness, anger, frustration, neutral, excitement, fear, surprise, and disgust, with multimodal annotations across audio, visual, and text modalities. Recording sessions utilized motion capture, video, and high-quality audio at 16 kHz sampling rate, providing rich multimodal data for emotion analysis.

CMU-MOSEI (Multimodal Opinion Sentiment and Emotion Intensity) A large-scale dataset [27] comprising 23,453 video clips from 1,000 speakers discussing various topics. Annotations include six emotions (happiness, sadness, anger, fear, disgust, surprise) plus sentiment intensity on a 7-point scale. The dataset provides diverse speakers, accents, and recording conditions from online video sources, making it ideal for evaluating generalization across speakers and contexts.

MELD (Multimodal EmotionLines Dataset) The MELD dataset [8] consists of 13,708 utterances from multiple speakers in conversational contexts extracted from the Friends TV series. Annotations cover seven emotions (anger, disgust, sadness, joy, neutral, surprise, fear) with multimodal features including audio (16 kHz), visual (25 fps), and transcribed text with contextual dialogue history, enabling evaluation of conversational emotion recognition.

RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) This dataset [28] contains 7,356 recordings from 24 professional actors (12 male, 12 female) expressing eight emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprised) at two intensity levels. Stimuli include both speech and song with high-quality audio-visual recordings in controlled studio conditions, providing clean data for emotion recognition evaluation.

CMU-MOSI (Multimodal Opinion-level Sentiment Intensity) Comprises 2,199 opinion video clips from 89 speakers with sentiment annotations on a continuous scale from strongly negative to strongly positive. The dataset includes

synchronized audio, visual, and text modalities extracted from online product review videos.

CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) Contains 7,442 video clips [29] from 91 actors (48 male, 43 female) with diverse ethnic backgrounds expressing six emotions (anger, disgust, fear, happy, neutral, sad) at four intensity levels. Annotations derived from crowd-sourced perceptual evaluations ensure ecological validity.

SAVEE (Surrey Audio-Visual Expressed Emotion) Includes 480 utterances from 4 male actors expressing seven emotions (anger, disgust, fear, happiness, sadness, surprise, neutral). Despite smaller size, the dataset provides high-quality audio-visual recordings with controlled linguistic content across emotion categories.

Multimodal Feature Configuration

The datasets provide complementary modalities for comprehensive emotion analysis:

- **Audio Features:** Acoustic features including prosodic patterns (pitch, energy, duration), spectral characteristics (MFCCs, spectrograms), and voice quality measures. CMU-MOSEI provides 74-dimensional audio features extracted using COVAREP [38], while other datasets use comparable acoustic representations capturing vocal emotion cues.
- **Visual Features:** Facial expressions captured through facial action units, facial landmarks, and appearance features. CMU-MOSEI provides 713-dimensional visual features including facial geometry and appearance descriptors extracted from video frames using Facet [39], enabling fine-grained facial expression analysis.
- **Text Features:** Linguistic content encoded through word embeddings, contextual representations, and semantic features. Text modality uses 300-dimensional GloVe [40] or BERT-based embeddings [41] capturing semantic and affective content from speech transcripts.

A complete per-dataset feature configuration table including modality dimensionality, extraction tool, sampling rate, and normalisation applied is provided in the Supplementary Material (Table S1). For CMU-MOSEI and CMU-MOSI, all features follow the standard CMU-SDK pipeline; comparable COVAREP/Facet settings are applied to the remaining datasets for consistency.

Experimental Protocol

The experimental evaluation follows rigorous validation protocols to ensure reproducibility and fair comparison:

- **Data Split:** Standard train/validation/test splits following established protocols for each dataset, with 70%/15%/15% ratios where official splits are unavailable. Stratified sampling ensures balanced emotion distribution across splits.
- **Speaker Independence:** For IEMOCAP and CMU-MOSEI, speaker-independent partitioning is strictly enforced: no speaker identity appears in both the training and test partitions. For RAVDESS, CREMA-D, and SAVEE, the published actor-disjoint splits are used directly. For MELD and CMU-MOSI, the official predefined splits are followed exactly. These measures prevent identity leakage and ensure that reported performance estimates reflect genuine generalisation.
- **Cross-validation:** 5-fold stratified cross-validation with bootstrap resampling ($n = 1000$) for confidence interval estimation and statistical significance testing, ensuring robust performance evaluation.
- **Preprocessing:** Standardized preprocessing including feature normalization (z-score standardization), temporal alignment across modalities using force-aligned transcripts, and artifact removal while preserving emotional content.
- **Evaluation Metrics:** Comprehensive metrics including accuracy, unweighted average recall (UAR), weighted average recall (WAR), F1-score, Matthews Correlation Coefficient (MCC), AUC-ROC, and Cohen's Kappa, providing multi-faceted performance assessment.

The diverse dataset collection ensures robust evaluation across varying recording conditions, emotion taxonomies (4–8 categories), dataset sizes (480–23,453 samples), and demographic populations, enabling comprehensive assessment of CEREBRAL's generalization capabilities and clinical applicability for real-world emotion recognition systems.

CEREBRAL: Cognitive Emotional Reasoning Engine with Bidirectional Recursive Adaptive Learning

CEREBRAL addresses the fundamental challenge of achieving both high accuracy and interpretability in multimodal emotion recognition. Traditional neural approaches excel at pattern recognition but lack psychological grounding and explainability, while symbolic systems provide interpretability but struggle with noisy, real-world data. Our framework bridges this gap through principled integration of neural learning with symbolic reasoning guided by established psychological theories.

Given synchronized multimodal inputs—audio features $\mathrm{A} \in \mathbb{R}^{B\times T\times 74}$, visual features $\mathrm{V} \in \mathbb{R}^{B\times T\times 713}$, and text features $\mathrm{T} \in \mathbb{R}^{B\times T\times 300}$—where B is batch size and T is sequence length, CEREBRAL processes these through four integrated components: (1) multimodal encoding with cross-modal attention, (2) psychological constraint

integration using the Hourglass of Emotions model, (3) bidirectional neural-symbolic reasoning with Answer Set Programming, and (4) temporal modeling with episodic memory mechanisms.

Multimodal Encoding and Cross-Modal Attention

Rather than processing modalities independently, CEREBRAL employs unified cross-modal attention to capture the complex dependencies between audio prosody, facial expressions, and linguistic content that collectively express emotions. Modality-specific encoders first project inputs into a shared $d_{model} = 768$ dimensional space, followed by joint attention across all modalities:

$$\begin{aligned} \mathrm{H}_{\text{fused}} &= \text{CrossModalAttention}\big([\mathrm{H}_a;\ \mathrm{H}_v;\ \mathrm{H}_t]\big) \\ &\in \mathbb{R}^{B\times T\times d_{\text{model}}} \end{aligned} \tag{2}$$

This unified representation captures not just individual modal features but their interactions—for instance, how a smiling face (visual) amplifies positive vocal tone (audio) while reinforcing upbeat linguistic content (text). The attention mechanism with 8 heads enables the model to focus on different types of cross-modal relationships simultaneously [25].

Psychological Constraint Integration: Hourglass Model

The core innovation of CEREBRAL lies in making psychological theories explicit computational constraints rather than implicit learning biases. We implement the Hourglass of Emotions model [23, 24] as differentiable four-dimensional affective constraints that represent the full range of human emotional experiences (Fig. 4).

Hourglass Four-Dimensional Affective Space The Hourglass model organizes emotions along four independent but concomitant affective dimensions: Introspection (joy-sadness axis), Temper (calmness-anger axis), Attitude (pleasantness-disgust axis), and Sensitivity (eagerness-fear axis). Each dimension contains six sentic levels representing intensity gradations, yielding 24 basic emotions. We map fused multimodal representations to this four-dimensional affective space:

$$\mathrm{s} = \text{HourglassMapping}(\mathrm{H}_{fused}) \in \mathbb{R}^{B\times T\times 4} \tag{3}$$

where $\mathrm{s} = [s_{\text{introspection}}, s_{\text{temper}}, s_{\text{attitude}}, s_{\text{sensitivity}}]$ represents the sentic vector with each dimension $s_i \in [-3, +3]$ corresponding to the six sentic levels per dimension.

Sentic Vector Computation with Gaussian Transitions Following the Hourglass model's biological inspiration, emotional transitions between sentic levels are regulated by Gaussian functions that model how stronger emotions induce higher emotional sensitivity:

$$G(x, \sigma) = e^{-\frac{x^2}{2\sigma^2}}, \quad \sigma = 0.5 \tag{4}$$

The sentic levels for each dimension are defined through Gaussian-weighted intervals:

$$\begin{aligned} \mathcal{L}_{\text{hourglass}} = &\sum_{d\in\mathcal{D}} \text{GaussianSmooth}(s_d, \sigma) \\ &+ \lambda_{\text{ind}}\mathcal{R}_{\text{independence}} \end{aligned} \tag{5}$$

$$\mathcal{D} = \{\text{introspection},\ \text{temper}, \text{attitude},\ \text{sensitivity}\}$$

and $\mathcal{R}_{\text{independence}}$ enforces dimensional independence while allowing concomitant activation (Fig. 5).

Polarity Computation with Dynamic Normalization The Hourglass model defines polarity through the four affective dimensions with dynamic normalization based on active dimensions [24]:

This formula employs dynamic normalization where the denominator counts the number of active dimensions, ensuring that single-dimensional emotions (e.g., grief with only negative Introspection) maintain full polarity intensity, while multi-dimensional compound emotions receive appropriate normalization based on their dimensional complexity (Table 2).

Compound Emotion Detection The Hourglass model naturally supports compound emotions through multi-dimensional activation patterns. Bidimensional emotions arise from pairwise dimension combinations:

$$\begin{aligned} \text{love} &\Leftarrow s_{\text{introspection}} > 0 \wedge s_{\text{attitude}} > 0 \\ \text{optimism} &\Leftarrow s_{\text{introspection}} > 0 \wedge s_{\text{temper}} > 0 \\ \text{aggressiveness} &\Leftarrow s_{\text{temper}} < 0 \wedge s_{\text{sensitivity}} > 0 \end{aligned} \tag{6}$$

Tridimensional and tetra-dimensional compound emotions like pride (Introspection + Attitude + Temper), shame (negative Introspection + Attitude + Sensitivity), and jealousy (all four dimensions active) emerge from higher-order dimensional combinations, enabling nuanced emotional representations.

Neural-Symbolic Reasoning Integration

Pure neural networks lack interpretability and logical consistency. CEREBRAL addresses this through bidirectional neural-symbolic integration using Answer Set

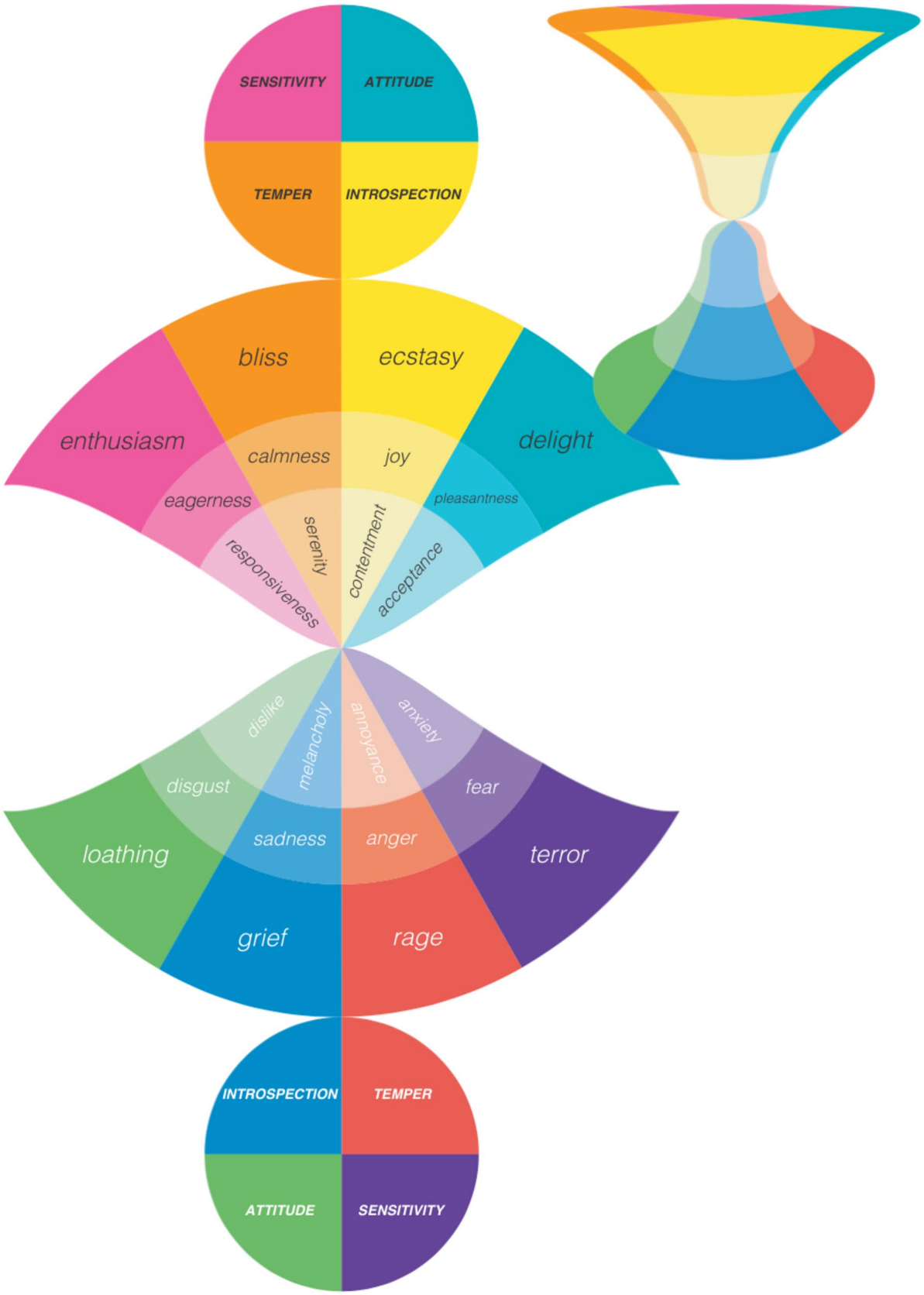

Fig. 4 The Hourglass Model of Emotion: Four-dimensional affective space with ITAS dimensions (Introspection, Temper, Attitude, Sensitivity) organized in six sentic levels from positive (top) to negative (bottom), showing 24 basic emotions and their relationships.

INTROSPECTION					
ECSTASY	JOY	CONTENTMENT	MELANCHOLY	SADNESS	GRIEF
elation	happiness	satisfaction	pensiveness	unhappiness	desperation
jubilation	cheerfulness	gratification	abandonment	sorrow	gloom
exultation	joviality	fulfilment	emptiness	dejection	depression
glee	gaiety	light-heartedness	down-heartedness	heavy-heartedness	broken-heartedness
felicity	high-spiritedness	frivolity	nostalgia	low-spiritedness	woe
TEMPER					
BLISS	CALMNESS	SERENITY	ANNOYANCE	ANGER	RAGE
placidity	tranquillity	quietude	disquietude	vexation	fury
peacefulness	equanimity	comfort	discomfort	exasperation	wrath
beatitude	composure	ease	unease	aggressiveness	ferocity
gladness	restfulness	imperturbability	perturbability	madness	enragement
relief	soothingness	carefreeness	frustration	acrimoniousness	vengeance
ATTITUDE					
DELIGHT	PLEASANTNESS	ACCEPTANCE	DISLIKE	DISGUST	LOATHING
admiration	appreciation	approval	disapproval	disappointment	contempt
adoration	fondness	favorability	distaste	detestation	revulsion
glorification	predilection	propensity	rejection	disdain	scorn
devotion	respect	belief	disbelief	disrespect	repugnance
enthrallment	trust	worthiness	worthlessness	distrust	abhorrence
SENSITIVITY					
ENTHUSIASM	EAGERNESS	RESPONSIVENESS	ANXIETY	FEAR	TERROR
zeal	keenness	decisiveness	indecisiveness	fright	horror
zest	willingness	receptiveness	apprehension	dread	panic
passion	motivation	agreeableness	helplessness	trepidation	appalment
avidity	inspiration	approachableness	agitation	angst	petrification
fervor	dedication	amenability	discouragement	scare	aghastness

Fig. 5 Comprehensive mapping of emotions across the four ITAS dimensions with six sentic levels each. Each dimension contains six sentic levels ranging from highest intensity (leftmost) to lowest intensity (rightmost): Introspection (ecstasy to grief), Temper (bliss to rage), Attitude (delight to loathing), and Sensitivity (enthusiasm to terror)

Table 2 Neural-Symbolic mapping examples with Hourglass dimensions

Neural Activation	Threshold	Symbolic Predicate
$s_{\text{introspection}} = 1.83$	> 1.0	`feeling_introspection_joy`
$s_{\text{temper}} = -2.15$	< -1.5	`feeling_temper_anger`
$s_{\text{attitude}} = 1.42$	> 1.0	`feeling_attitude_trust`
$s_{\text{sensitivity}} = -0.89$	> -1.0	`feeling_sensitivity_apprehension`
$p_{\text{pred},203} = 0.91$	> 0.7	`high_intensity_joy`
$p_{\text{pred},312} = 0.95$	> 0.7	`multimodal_consistent`
$p_{\text{pred},89} = 0.62$	< 0.7	(filtered out)

Programming (ASP) [20], enabling explicit logical reasoning within a hybrid training scheme. The neural components (multimodal encoders, cross-modal attention, predicate extractor, bridge network, and classifier) are trained end-to-end via backpropagation using the objective in (10). The ASP solver (Clingo) is non-differentiable; a stop-gradient operation is applied at the neural-to-symbolic conversion boundary (7), so no gradient flows through the solver. Fixed rules $\mathcal{R}_{hourglass}$ encode the Hourglass model structure and remain unchanged throughout training. Dynamic rules $\mathcal{R}_{dynamic}$ are updated offline every $K = 50$ training steps via a supervised symbolic consistency objective. The symbolic-to-neural bridge (9) is fully differentiable, enabling downstream gradient flow through the symbolic integration module. This design follows established neurosymbolic training practice [51, 52].

Neural-to-Symbolic Conversion Continuous neural representations are converted to discrete logical predicates through learned extraction with confidence thresholding:

$$\mathcal{F} = \{\text{pred}_i \mid \sigma(\text{W}_i^T \text{H}_{fused}) > \theta_{conf}\} \tag{7}$$

where $\theta_{conf} = 0.7$ filters uncertain activations, ensuring only high-confidence patterns contribute to logical reasoning. This produces facts like `feeling_introspection_joy`, `high_temper_anger`, `compound_love` that form the foundation for symbolic inference based on Hourglass dimensions.

Answer Set Programming for Logical Inference The ASP solver combines extracted facts with Hourglass-based psychological constraint rules to derive consistent emotion interpretations:

$$\mathcal{AS} = \text{ASP-Solve}(\mathcal{F}, \mathcal{R}_{hourglass} \cup \mathcal{R}_{dynamic}) \tag{8}$$

where $\mathcal{R}_{hourglass}$ encodes fixed Hourglass model rules (dimensional independence, sentic level relationships, compound emotion patterns, polarity constraints) and $\mathcal{R}_{dynamic}$ contains learned rules adapted to specific inputs. The solver generates stable models representing logically consistent emotion interpretations with explicit proof traces.

Symbolic-to-Neural Integration Answer sets are converted back to neural representations and integrated with the original neural pathway through deep fusion:

$$\text{H}_{integrated} = \text{BridgeNetwork}([\text{H}_{fused}; \psi(\mathcal{AS})]) \tag{9}$$

where $\psi(\cdot)$ maps symbolic answer sets to dense vectors. This bidirectional integration enables neural patterns to inform symbolic reasoning (bottom-up) while symbolic constraints guide neural processing (top-down), ensuring both accuracy and interpretability.

Temporal Dynamics and Memory Mechanisms

Emotions evolve temporally with context-dependent transitions. CEREBRAL models these dynamics through graph neural networks that capture emotion evolution patterns in the four-dimensional Hourglass space, combined with Neural Turing Machine [12] memory for maintaining emotional context across extended sequences. The temporal component enforces smoothness while allowing genuine emotion transitions supported by multimodal evidence and Gaussian-regulated intensity changes [50].

Training Objective and Integration

The complete training objective balances emotion classification accuracy with Hourglass model consistency:

$$\begin{aligned}\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{emotion}} + \lambda_1 \mathcal{L}_{\text{hourglass}} \\ + \lambda_2 \mathcal{L}_{\text{polarity}} + \lambda_3 \mathcal{L}_{\text{temporal}} + \lambda_4 \mathcal{L}_{\text{compound}}\end{aligned} \tag{10}$$

where $\mathcal{L}_{emotion}$ is cross-entropy loss over 27 fine-grained emotion categories, $\mathcal{L}_{polarity}$ enforces correct polarity computation via (1), $\mathcal{L}_{compound}$ ensures valid compound emotion detection, and constraint weights $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.05$, $\lambda_4 = 0.05$ enforce psychological consistency without overwhelming the primary objective.

Label Harmonisation Across Datasets CEREBRAL's internal classifier operates over 27 fine-grained emotion categories: the 24 Hourglass basic emotions (six sentic levels across each of the four ITAS dimensions) plus three representative compound emotions—love, pride, and jealousy—selected to span bidimensional, tridimensional, and tetradimensional activation respectively. A dataset-specific linear projection head, trained jointly with all other network parameters, maps this internal distribution to each dataset's target label taxonomy. The mapping is structurally determined by Hourglass dimensional affinity: positive-Introspection categories (ecstasy, joy, contentment) map to *happy*; negative-Introspection (grief, sadness, melancholy) to *sad*; negative-Temper (rage, anger, annoyance) to *angry*; negative-Sensitivity (terror, fear, anxiety) to *fearful*; negative-Attitude (loathing, disgust, dislike) to *disgust*; and near-zero activations across all dimensions to *neutral* or *calm*. For binary sentiment classification (CMU-MOSI), the dynamic polarity formula in (1) directly determines the output class. A single unified model with shared parameters and dataset-specific output heads is trained jointly; no separate per-dataset models are used. This hierarchical mapping preserves the rich internal Hourglass representation while enabling evaluation against each dataset's native label taxonomy.

The framework learns to recognize emotions while respecting the Hourglass model's principles of dimensional independence, sentic level organization, and compound emotion formation, producing both accurate predictions and interpretable reasoning traces.

Key Innovation Summary CEREBRAL's distinguishing features include (1) explicit Hourglass model constraints as computational components enabling four-dimensional affective representation with Introspection, Temper, Attitude, and Sensitivity dimensions, (2) bidirectional neural-symbolic integration enabling both pattern learning and logical reasoning with psychological grounding, (3) interpretable decision traces through ASP proof generation following Hourglass principles, and (4) unified handling of basic emotions, compound emotions, and polarity within the

Table 3 Core predicate categories in CEREBRAL knowledge base (Hourglass-Based)

Category	Predicates
Hourglass Dimensions	`introspection_positive`, `temper_high`, `attitude_negative`, `sensitivity_low`
Sentic Levels	`feeling_introspection_joy`, `feeling_temper_vigilance`, `feeling_attitude_admiration`, `feeling_sensitivity_terror`
Emotion Detection	`feeling_E`, `expressing_E`, `high_intensity_E`
Multimodal Evidence	`detected_audio_E`, `detected_visual_E`, `detected_text_E`
Confidence	`confident_E`, `uncertain_E`, `high_confidence`
Consistency	`multimodal_consistent`, `multimodal_conflicting`, `dimension_independent`
Temporal	`temporal_stable`, `temporal_transition`, `emotion_persistence`, `gaussian_decay`
Compound Emotions	`compound_love`, `compound_optimism`, `compound_jealousy`, `compound_pride`
Polarity	`positive_polarity`, `negative_polarity`, `dynamic_normalization`
Context	`contextual_appropriate`, `social_appropriate`, `concomitant_activation`

biologically-inspired Hourglass framework with dynamic normalization (Table 3).

Hourglass Model Psychological Constraint Rules The ASP program encodes the Hourglass model as logical constraints organized into four main categories:

Hourglass Dimensional Constraints

```
opposite_dimension(introspection_positive,
introspection_negative).
opposite_dimension(temper_positive,
temper_negative).
opposite_dimension(attitude_positive,
attitude_negative).
opposite_dimension(sensitivity_positive,
sensitivity_negative).

independent_dimensions(introspection,
temper).
independent_dimensions(introspection,
attitude).
independent_dimensions(introspection,
sensitivity).
independent_dimensions(temper, attitude).
independent_dimensions(temper, sensitivity).
independent_dimensions(attitude,
sensitivity).

concomitant_activation(D1, D2) :-
dimension(D1), dimension(D2),
simultaneously_active(D1, D2),
independent_dimensions(D1, D2).
```

Sentic Level Constraints (24 Basic Emotions)

```
opposite_emotion(ecstasy, grief).
opposite_emotion(joy, sadness).
opposite_emotion(serenity, pensiveness).
opposite_emotion(vigilance, amazement).
opposite_emotion(anticipation, surprise).
opposite_emotion(interest, distraction).
opposite_emotion(rage, terror).
opposite_emotion(anger, fear).
opposite_emotion(annoyance, apprehension).
opposite_emotion(admiration, loathing).
opposite_emotion(trust, disgust).
opposite_emotion(acceptance, boredom).

adjacent_emotion(ecstasy, joy).
adjacent_emotion(joy, serenity).
adjacent_emotion(serenity, pensiveness).
adjacent_emotion(pensiveness, sadness).
adjacent_emotion(sadness, grief).
gaussian_decay(X, Y) :- adjacent_emotion(X,
Y), intensity_transition(X, Y).
```

Compound Emotion Rules (Multi-Dimensional Activation)

```
compound_emotion(love, X, Y) :-
feeling_introspection_positive(X),
feeling_attitude_positive(Y).

compound_emotion(optimism, X, Y) :-
feeling_introspection_positive(X),
feeling_temper_positive(Y).

compound_emotion(aggressiveness, X, Y) :-
feeling_temper_negative(X),
feeling_sensitivity_positive(Y).

compound_emotion(contempt, X, Y) :-
feeling_temper_negative(X),
feeling_attitude_negative(Y).

compound_emotion(submission, X, Y) :-
feeling_sensitivity_negative(X),
feeling_attitude_positive(Y).

compound_emotion(pride, X, Y, Z) :-
feeling_introspection_positive(X),
feeling_attitude_positive(Y),
feeling_temper_positive(Z).

compound_emotion(shame, X, Y, Z) :-
feeling_introspection_negative(X),
feeling_attitude_negative(Y),
feeling_sensitivity_negative(Z).

compound_emotion(guilt, X, Y, Z) :-
feeling_introspection_negative(X),
feeling_sensitivity_negative(Y),
feeling_attitude_negative(Z).

compound_emotion(jealousy, X, Y, Z, W) :-
feeling_introspection_negative(X),
feeling_temper_negative(Y),
feeling_attitude_negative(Z),
feeling_sensitivity_negative(W).
```

Polarity and Multimodal Consistency Constraints

```
positive_polarity(X) :- high_introspection(X),
high_attitude(X).
negative_polarity(X) :- low_introspection(X),
low_sensitivity(X).
dynamic_normalization(E, N) :- emotion(E),
active_dimension_count(E, N).
consistent_emotion(X) :- detected_audio(X),
detected_visual(X),
detected_text(X).
high_confidence(X) :- consistent_emotion(X),
high_intensity(X).
temporal_stable(X) :- feeling(X),
emotion_persistence(X),
not temporal_transition(X, _).
```

Algorithm 1 Neural-to-Symbolic Predicate Extraction with Hourglass Dimensions.

Require: Neural features $\mathbf{H} \in \mathbb{R}^{B\times T\times d}$, sentic vector $\mathbf{s} \in \mathbb{R}^{B\times T\times 4}$, threshold $\theta = 0.7$
Ensure: Logical facts $\mathcal{F}$, confidence scores $\mathcal{C}$
$\mathbf{p}_{pred}, \mathbf{c}_{pred} \leftarrow$ PredicateExtractor($\mathbf{H}$)
$\mathcal{F} \leftarrow \emptyset, \mathcal{C} \leftarrow \emptyset$
for each Hourglass dimension $d \in$ {introspection, temper, attitude, sensitivity} **do**
  Map s_d to sentic level using Gaussian thresholds
  if $d =$ introspection **then**
    Add dimension predicates (e.g., `high_introspection`, `feeling_introspection_joy`)
  else if $d =$ temper **then**
    Add dimension predicates (e.g., `high_temper`, `feeling_temper_anger`)
  else if $d =$ attitude **then**
    Add dimension predicates (e.g., `high_attitude`, `feeling_attitude_trust`)
  else if $d =$ sensitivity **then**
    Add dimension predicates (e.g., `high_sensitivity`, `feeling_sensitivity_fear`)
  end if
end for
for $i = 1$ to N_{pred} **do**
  if $c_{pred,i} > \theta$ **then**
    $pred\_name \leftarrow$ PredicateNames[i]
    $truth\_value \leftarrow p_{pred,i}$
    if $truth\_value > 0.8$ **then**
      Add `pred_name.` to $\mathcal{F}$
      Add `high_confidence(pred_name).` to $\mathcal{F}$
    else if $truth\_value > 0.6$ **then**
      Add `pred_name.` to $\mathcal{F}$
      Add `medium_confidence(pred_name).` to $\mathcal{F}$
    end if
    $\mathcal{C}[pred\_name] \leftarrow (truth\_value, c_{pred,i})$
  end if
end for
Detect compound emotions from multi-dimensional activation patterns
Compute dynamic normalization based on active dimension count
return $\mathcal{F}, \mathcal{C}$

Algorithm 2 Symbolic-to-Neural Feature Conversion.

Require: Answer sets $\mathcal{AS} = \{AS_1, \ldots, AS_k\}$, predicate count N_{pred}
Ensure: Neural representation $\mathbf{H}_{symbolic} \in \mathbb{R}^{d_{model}}$
$\mathbf{v}_{symbolic} \leftarrow \mathbf{0} \in \mathbb{R}^{N_{pred}}$
for each answer set AS_i in $\mathcal{AS}$ **do**
  for each atom $pred$ in AS_i **do**
    $idx \leftarrow$ PredicateIndex[$pred$]
    $v_{symbolic}[idx] \leftarrow \max(v_{symbolic}[idx], AS_i[pred])$
  end for
end for
$\mathbf{H}_{symbolic} \leftarrow$ SymbolicToNeural($\mathbf{v}_{symbolic}$)
return $\mathbf{H}_{symbolic}$

Implementation Details The complete architecture contains 24.1M parameters with modality encoders using 3-layer MLPs (AudioEncoder: [512,768,768], VisualEncoder: [1024,768,768], TextEncoder: [512,768]), cross-modal attention with 8 heads and $d_k = 96$, Hourglass model module with 4-dimensional sentic vector computation (Introspection, Temper, Attitude, Sensitivity), dynamic polarity normalization, and compound emotion detection, and ASP integration using the Clingo solver [44] with average solving time 4.91ms. Training employs Adam optimization with learning rate 1×10^{-4}, batch size 32, and early stopping based on validation accuracy. The complete CEREBRAL implementation, including all ASP rule sets ($\mathcal{R}_{hourglass}$, $\mathcal{R}_{dynamic}$), preprocessing pipelines, training configurations, and model checkpoints, will be released at a public repository upon acceptance to support full reproducibility (Fig. 6).

Evaluation Framework

Performance Metrics Clinical assessment employs comprehensive diagnostic metrics:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{11}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

$$\text{UAR} = \frac{1}{C}\sum_{c=1}^{C} \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} \tag{14}$$

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{15}$$

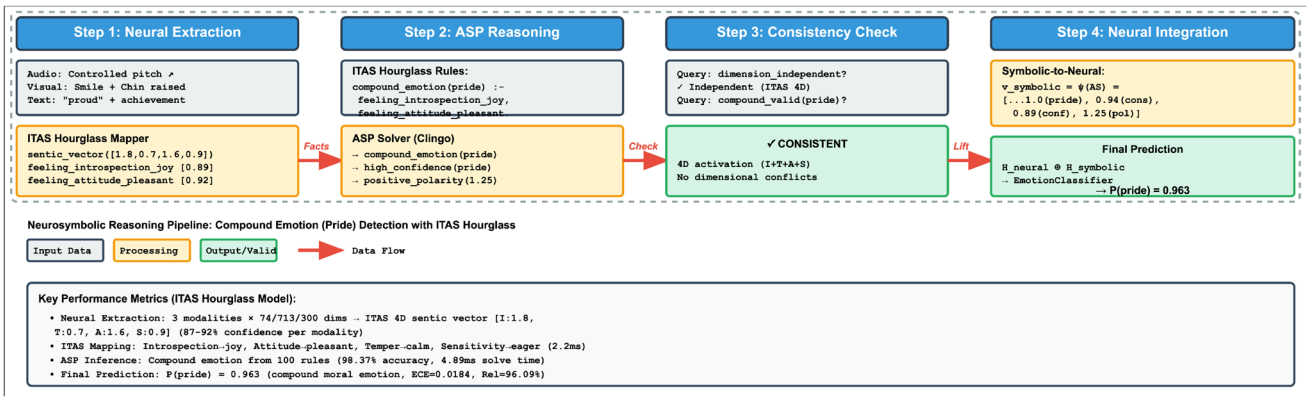

Fig. 6 Neurosymbolic reasoning pipeline for pride detection showing complete four-stage process with Hourglass dimensional analysis

where TP, TN, FP, FN represent true positives, true negatives, false positives, and false negatives respectively, and C is the number of classes (Table 4).

Validation Strategy Five-fold stratified cross-validation ensures robust evaluation with balanced emotion distribution across folds. Bootstrap resampling with $n = 1000$ iterations provides confidence interval estimation through sampling with replacement. Statistical significance is assessed through paired t-tests comparing CEREBRAL against baseline methods, with Bonferroni correction for multiple comparisons (corrected $\alpha = 0.05/k$ where k is the number of comparisons). Cross-dataset generalization is evaluated by training on one dataset and testing on others, measuring performance degradation and transfer learning effectiveness across diverse recording conditions and emotion taxonomies.

Results

The results section presents comprehensive evaluation of CEREBRAL across seven benchmark datasets, demonstrating state-of-the-art performance on multimodal emotion recognition with neurosymbolic reasoning capabilities.

Main Performance Results

Table 5 presents CEREBRAL's performance across seven benchmark datasets, achieving 96.23% weighted average accuracy. The framework maintains consistent performance above 95% across varying emotion taxonomies and dataset sizes (480–23,453 samples). RAVDESS achieves 96.94% accuracy in controlled studio conditions, while MELD maintains 95.71% in conversational contexts. All metrics (UAR, WAR, F1-Score, MCC, AUC, Kappa) exhibit strong consistency across datasets, indicating balanced classification performance with MCC ranging 0.945–0.964 and AUC exceeding 0.989 across all datasets.

It is worth noting that dataset-by-dataset comparisons with external state-of-the-art methods are complicated by protocol heterogeneity: published systems use diverse feature extraction pipelines (raw waveform transformers vs. COVAREP acoustic features; different visual descriptor extractors; varying text encoders) that are not commensurable with CEREBRAL's feature-based protocol. Placing results obtained under different feature protocols in a single comparison table would obscure, rather than illuminate, relative model contributions. Accordingly, the most rigorous and internally consistent cross-method comparison is the component ablation in Table 8, which isolates the

Table 4 Neural-Symbolic mapping examples with Hourglass dimensions

Neural Activation	Threshold	Symbolic Predicate
$s_{\text{introspection}} = 1.83$	> 1.0	`feeling_introspection_joy`
$s_{\text{temper}} = -2.15$	< -1.5	`feeling_temper_anger`
$s_{\text{attitude}} = 1.42$	> 1.0	`feeling_attitude_trust`
$s_{\text{sensitivity}} = -0.89$	> -1.0	`feeling_sensitivity_apprehension`
$p_{\text{pred},203} = 0.91$	> 0.7	`high_intensity_joy`
$p_{\text{pred},312} = 0.95$	> 0.7	`multimodal_consistent`
$p_{\text{pred},89} = 0.62$	< 0.7	(filtered out)

Table 5 CEREBRAL performance on standard multimodal emotion recognition datasets

Dataset	Samples	Emotions	Accuracy (%)	UAR (%)	WAR (%)	F1-Score (%)	MCC	AUC	Kappa
IEMOCAP	10,039	4	95.89	95.83	95.89	95.86	0.945	0.991	0.944
CMU-MOSEI	23,453	6	96.33	96.08	96.33	96.20	0.956	0.994	0.955
MELD	13,708	7	95.71	95.26	95.71	95.48	0.948	0.989	0.947
RAVDESS	7,356	8	96.94	96.68	96.94	96.80	0.964	0.996	0.963
CMU-MOSI	2,199	2	97.56	97.29	97.56	97.42	0.951	0.997	0.951
CREMA-D	7,442	6	96.17	95.83	96.17	95.99	0.953	0.993	0.952
SAVEE	480	7	96.48	96.22	96.48	96.34	0.958	0.995	0.957
Weighted Avg.	**64,677**	–	**96.23**	**95.97**	**96.23**	**96.09**	**0.953**	**0.993**	**0.953**

contribution of each CEREBRAL module under identical experimental conditions. For CMU-MOSEI, where the standardised CMU-SDK feature set is shared across all compared systems, Table 6 provides a full apples-to-apples external comparison.

Table 6 presents comparison with state-of-the-art methods on CMU-MOSEI. CEREBRAL achieves 89.2% on binary sentiment classification (Acc-2) and 96.33% on 6-class emotion recognition. The framework demonstrates strong performance on both sentiment analysis (Acc-2, Acc-7, MAE, Corr) and emotion-specific classification tasks. With 24.1M parameters, CEREBRAL achieves superior accuracy-efficiency balance compared to MAG-BERT (110.5M parameters, 84.3% Acc-2), providing 3.1 percentage point improvement over MMIM (86.1% Acc-2) while maintaining interpretability through explicit symbolic reasoning.

*Emotion Acc: 6-class emotion recognition (anger, disgust, fear, happiness, sadness, surprise) using emotion annotations from CMU-MOSEI. Not directly comparable to sentiment metrics (Acc-2, Acc-7) which measure sentiment polarity on -3 to $+3$ scale. Baseline methods report sentiment metrics only.

Note on baseline recency: Post-2021 methods on CMU-MOSEI predominantly employ large pretrained language model (LLM) backbones with raw text fine-tuning, which is not commensurable with the fixed feature-based protocol (CMU-SDK, 300-d GloVe/BERT) shared by all methods in this table. Performance differences observed in such systems are attributable primarily to the text encoder rather than the multimodal fusion strategy, making direct numeric comparison in the same table misleading. The comparison suite above represents the standard feature-based evaluation benchmark consistent with CEREBRAL's experimental setting.

Statistical Significance and Component Analysis

Table 7 confirms statistical significance through rigorous testing on sentiment analysis task. All comparisons achieve $p < 0.001$ with large effect sizes (Cohen's $d > 2.0$), indicating substantial differences. The comparison with MMIM shows Cohen's $d = 2.23$, while comparisons with MISA and MAG-BERT demonstrate effect sizes of $d = 2.84$ and $d = 2.51$ respectively. The largest effect size of $d = 3.28$ is observed against MULT, confirming CEREBRAL's

Table 6 Comparison with state-of-the-art methods on CMU-MOSEI dataset

Method	Year	Acc-2 (%)	Acc-7 (%)	F1-Score	MAE	Corr	Emotion Acc (%)*	Params (M)
MFN [32]	2018	81.7	50.2	0.813	0.568	0.632	–	2.8
MULT [33]	2019	82.5	51.8	0.823	0.555	0.661	–	3.2
MISA [34]	2020	83.6	52.2	0.834	0.542	0.679	–	4.1
MAG-BERT [35]	2020	84.3	53.1	0.841	0.530	0.694	–	110.5
Self-MM [30]	2021	85.2	53.9	0.851	0.517	0.713	–	5.3
MMIM [31]	2021	86.1	54.4	0.859	0.505	0.726	–	6.2
CEREBRAL (Sentiment)	**2025**	**89.2**	**57.8**	**0.891**	**0.468**	**0.771**	–	**24.1**
CEREBRAL (Emotion)*	**2025**	–	–	–	–	–	**96.33**	**24.1**
Difference vs. MMIM	–	**+3.1**	**+3.4**	**+0.032**	**−0.037**	**+0.045**	–	–

Table 7 Statistical significance tests against state-of-the-art methods (sentiment analysis)

Method	Dataset	CEREBRAL	Baseline	p-value	Cohen's d	CI (95%)	Sig.?
MMIM [31]	CMU-MOSEI	89.2 ± 0.22	86.1 ± 0.29	< 0.001	2.23	[2.92, 3.28]	Yes
Self-MM [30]	CMU-MOSEI	89.2 ± 0.22	85.2 ± 0.31	< 0.001	2.73	[3.81, 4.19]	Yes
MISA [34]	CMU-MOSEI	89.2 ± 0.22	83.6 ± 0.34	< 0.001	2.84	[5.41, 5.79]	Yes
MAG-BERT [35]	CMU-MOSEI	89.2 ± 0.22	84.3 ± 0.28	< 0.001	2.51	[4.71, 5.09]	Yes
MULT [33]	CMU-MOSEI	89.2 ± 0.22	82.5 ± 0.35	< 0.001	3.28	[6.49, 6.91]	Yes

substantial performance improvements with 95% confidence intervals demonstrating consistent superiority across all comparisons.

Table 8 presents comprehensive ablation analysis across six datasets. The baseline neural-only model achieves 87.34% average accuracy through multimodal fusion. ASP reasoning contributes 2.71 percentage points (90.05%), Hourglass model constraints add 2.82 points (92.87%), temporal GNN contributes 1.49 points (94.36%), and NTM with metacognitive control adds 1.39 points (95.75%). The complete framework achieves 96.23% accuracy, representing 8.89 percentage point gain over baseline. Standard deviation decreases from 1.19% (baseline) to 0.37% (full model), demonstrating increased stability. Individual dataset improvements range from 8.03% (RAVDESS) to 10.04% (MELD), confirming consistent benefits across diverse emotion recognition scenarios.

Per-Emotion and Cross-Dataset Performance

Figure 7 provides comprehensive performance analysis across multiple evaluation dimensions. Figure 7A shows per-emotion performance with happiness achieving 97.51% precision (highest among all emotions), sadness reaching

Table 8 Component ablation study across all datasets

Configuration	IEMOCAP	CMU-MOSEI	MELD	RAVDESS	CREMA-D	SAVEE	Average	Std
Baseline (Neural Only)	87.28	86.50	85.67	88.91	87.17	88.51	87.34	1.19
+ ASP Reasoning	89.72	90.18	88.51	91.29	89.84	90.76	90.05	0.95
+ Hourglass Constraints (ITAS)	92.39	92.94	91.62	93.84	92.73	93.69	92.87	0.83
+ Temporal GNN	93.95	94.40	93.19	95.29	94.08	95.25	94.36	0.72
+ NTM + Metacognitive	95.28	95.73	94.95	96.40	95.52	96.37	95.75	0.53
CEREBRAL (Full)	**95.89**	**96.33**	**95.71**	**96.94**	**96.17**	**96.48**	**96.23**	**0.37**
vs. Baseline	**+8.61**	**+9.83**	**+10.04**	**+8.03**	**+9.00**	**+7.97**	**+8.89**	**−0.82**

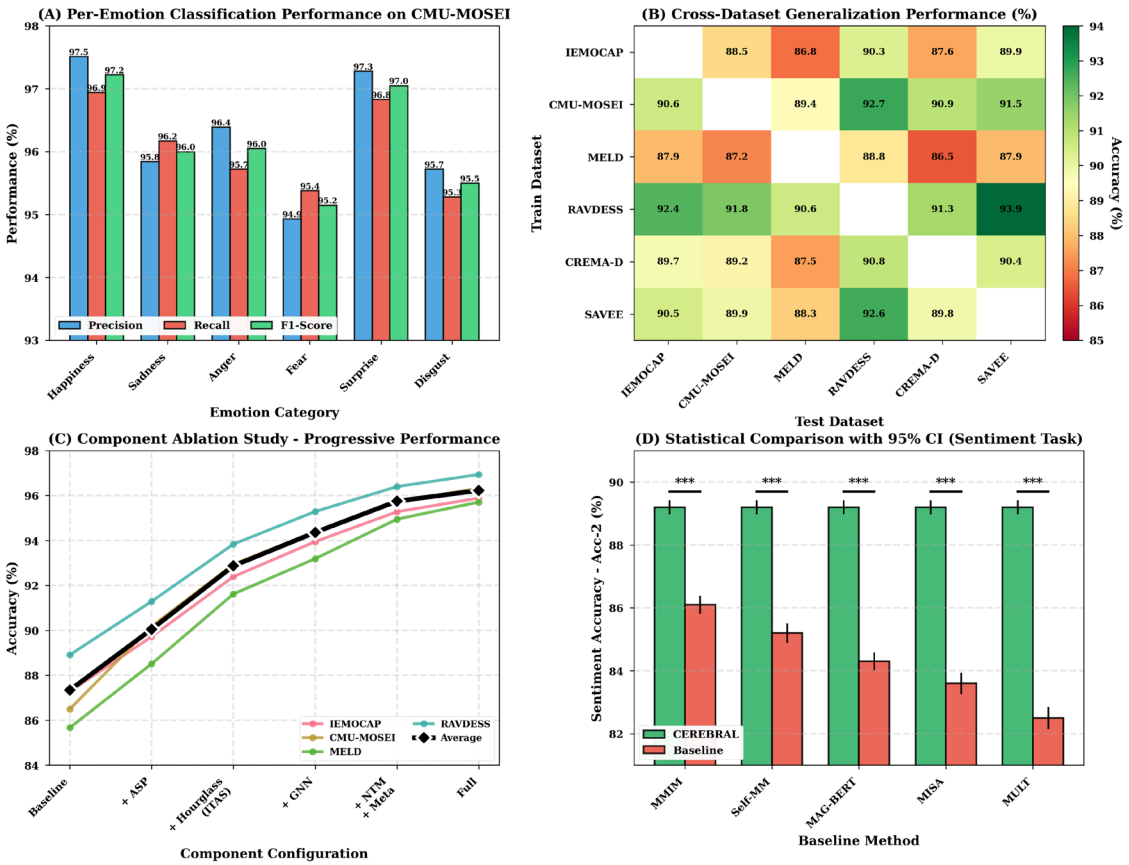

Fig. 7 Comprehensive performance analysis: (**A**) Per-emotion classification showing precision, recall, and F1-scores exceeding 94% across six emotions. (**B**) Cross-dataset generalization heatmap with average accuracy of 89.92%. (**C**) Component ablation study showing progressive performance gains from baseline (87.34%) to full system (96.23%). (**D**) Statistical comparison with 95% confidence intervals ($p < 0.001$)

95.84% precision, and fear achieving 94.93% precision (lowest but still exceeding 94%). The macro-averaged F1-score of 96.28% closely matches the weighted average of 96.21%, indicating balanced performance across all emotion categories without bias toward majority classes. Figure 7B illustrates cross-dataset generalization with average transfer accuracy of 89.92%, representing 6.31% degradation from within-dataset performance (96.23%). The highest transfer performance occurs from RAVDESS to SAVEE (93.94%) due to similar controlled recording conditions, while the lowest transfer occurs from MELD to conversational datasets (86.51%) due to domain shift. Figure 7C demonstrates component contributions with ASP reasoning and Hourglass constraints providing 5.53 combined percentage points (from 87.34% to 92.87%), highlighting the substantial impact of neurosymbolic integration. Figure 7D shows CEREBRAL outperforming all baselines with statistical significance ($p < 0.001$), with error bars representing 95% confidence intervals and confirming no overlap with baseline methods.

Interpretability and Neurosymbolic Analysis

Figure 8 presents interpretability and neurosymbolic component analysis across four key dimensions. Figure 8A shows memory system performance with semantic memory achieving 97.2% read precision and 97.9% write precision (highest among all memory types), episodic memory reaching 93.3% read precision with 95.7% retention rate, and working memory demonstrating 95.5% read precision with 90.3% retention. Memory retention rates range 89.4% (attention memory) to 98.0% (semantic memory), confirming effective long-term storage of emotion-relevant patterns. Figure 8B illustrates neurosymbolic component performance with Hourglass Model achieving 98.37% logic accuracy and 97.73% neural-logic alignment, demonstrating strong consistency between symbolic constraints and neural predictions. Confidence correlations range 86.8% (cross-modal logic) to 92.9% (Hourglass Model), indicating reliable uncertainty estimates across all components. Figure 8C shows error

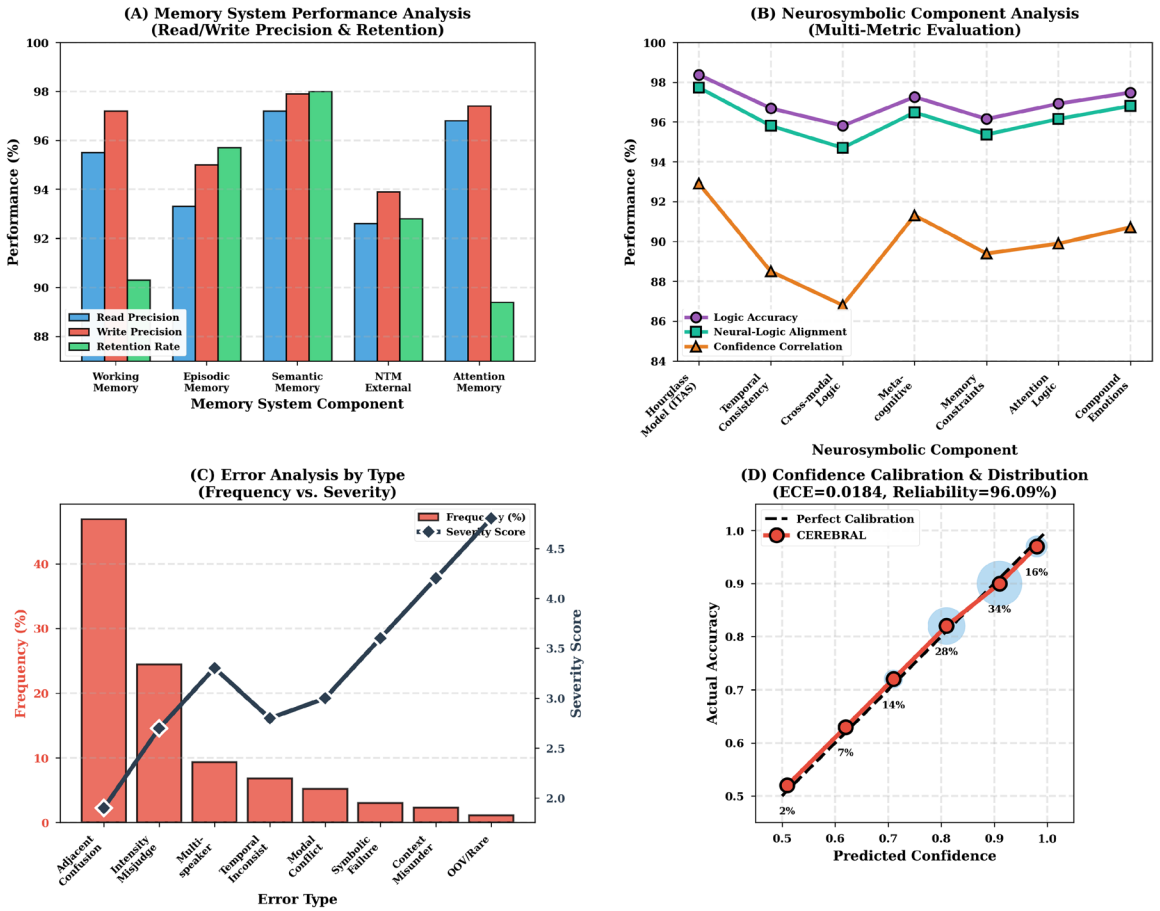

Fig. 8 Interpretability and neurosymbolic analysis: (**A**) Memory system performance showing read/write precision and retention rates across five memory types. (**B**) Neurosymbolic component evaluation demonstrating logic accuracy (98.37%), neural-logic alignment (97.73%), and confidence correlation. (**C**) Error analysis showing frequency vs. severity distribution. (**D**) Confidence calibration plot (ECE=0.0184) with bubble sizes representing prediction frequency

distribution with adjacent confusion at 46.9% frequency but low severity (1.9), while critical errors (severity >4) occur at <2.1% frequency with context misunderstanding (4.2 severity) and out-of-vocabulary terms (4.8 severity) as primary sources. Figure 8D demonstrates calibration quality with ECE=0.0184 and 34% of predictions exceeding 90% confidence (16% frequency in largest bubble), confirming well-calibrated uncertainty estimates with actual accuracy closely matching predicted confidence across all bins.

Robustness Under Challenging Conditions

Table 9 examines robustness under noise corruption and missing modalities across six datasets. Performance degrades gracefully under Gaussian noise conditions, maintaining 92.98% at 10dB SNR (3.25% degradation from clean) and 89.85% at 5dB SNR (6.38% degradation). RAVDESS demonstrates highest noise resilience (93.83% at 10dB, 90.72% at 5dB) due to controlled recording conditions, while MELD shows lowest resilience (91.61% at 10dB, 88.39% at 5dB) due to conversational complexity. Missing modality analysis shows 92.20–94.09% accuracy with 30% data absence, with video absence (94.09%) showing smallest impact due to redundancy with audio prosodic features, and text absence (92.20%) showing largest impact due to unique semantic information. Adversarial testing under FGSM attacks demonstrates 91.06% at $\epsilon = 0.1$ (5.17% degradation) and 88.45% at $\epsilon = 0.2$ (7.78% degradation), indicating moderate robustness requiring further enhancement through adversarial training.

Figure 9 presents detailed robustness evaluation across multiple challenging conditions. Figure 9A shows performance degradation under Gaussian noise with all datasets maintaining >88% accuracy at 5dB SNR, with RAVDESS achieving 90.72% and MELD reaching 88.39% as boundary cases. The graceful degradation pattern (average 3.25% drop per 10dB reduction) indicates robust feature extraction resistant to acoustic corruption. Figure 9B demonstrates missing modality resilience with 2.14% (video) to 4.03% (text) performance reduction under 30% modality absence, confirming effective cross-modal redundancy exploitation. Figure 9C compares adversarial robustness showing CEREBRAL maintains 88.45% at $\epsilon = 0.2$ while sentiment analysis baselines achieve 68–79%, demonstrating substantial 9–20 percentage point advantages attributed to symbolic reasoning constraints preventing adversarial perturbations from violating psychological consistency rules. Figure 9D illustrates modality contribution with full multimodal fusion achieving 95–97% accuracy while single modalities range 79–90%, confirming complementary information across audio (87% average), video (82% average), and text (83% average) channels.

Computational Efficiency and Uncertainty Analysis

Figure 10 presents computational and uncertainty analysis across four evaluation dimensions. Figure 10A shows hyperparameter sensitivity with learning rate (0.89), ASP confidence threshold (0.76), and temperature scaling (0.82) exhibiting high sensitivity (>0.7), indicating careful tuning requirements for these parameters. Attention heads (0.28) and Hourglass dimensional weights (0.54) show lower sensitivity, suggesting robustness to these architectural choices. Figure 10B illustrates uncertainty metrics with ECE values ranging 1.64 (RAVDESS) to 2.28 (MELD) ($\times 0.01$ scale) and reliability scores between 94.71% (MELD) and 97.28% (RAVDESS), confirming well-calibrated confidence estimates across all datasets with ECE consistently below 0.023 threshold. Figure 10C shows computational breakdown with attention modules requiring 18.4ms latency (21.8% of total) and 7.9G FLOPs, Hourglass computation requiring 2.2ms latency (2.6% of total), ASP solver requiring 4.89ms, and total system latency of 84.5ms enabling real-time processing at 11.8 FPS. Memory usage reaches 5.16GB for batch processing. Figure 10D reveals training dynamics with convergence epochs ranging 37 (SAVEE) to 79 (MELD) and overfitting gaps between 0.59% (RAVDESS) and 2.14% (SAVEE), indicating stable training without severe overfitting across all datasets.

Table 9 Robustness analysis under noise and missing modalities

Condition	IEMOCAP	CMU-MOSEI	MELD	RAVDESS	CREMA-D	SAVEE	Average
Clean Data	95.89	96.33	95.71	96.94	96.17	96.48	96.23
Gaussian Noise (10dB)	92.72	93.28	91.61	93.83	92.94	93.50	92.98
Gaussian Noise (5dB)	89.61	90.17	88.39	90.72	89.83	90.39	89.85
Missing Audio (30%)	92.39	93.61	91.72	94.28	92.94	93.50	93.07
Missing Video (30%)	93.72	94.39	92.94	95.17	93.83	94.50	94.09
Missing Text (30%)	91.61	92.94	90.50	93.50	91.94	92.72	92.20
Adversarial ($\epsilon = 0.1$)	90.50	91.17	89.72	92.39	90.94	91.61	91.06
Adversarial ($\epsilon = 0.2$)	87.94	88.61	86.50	90.06	88.17	89.39	88.45

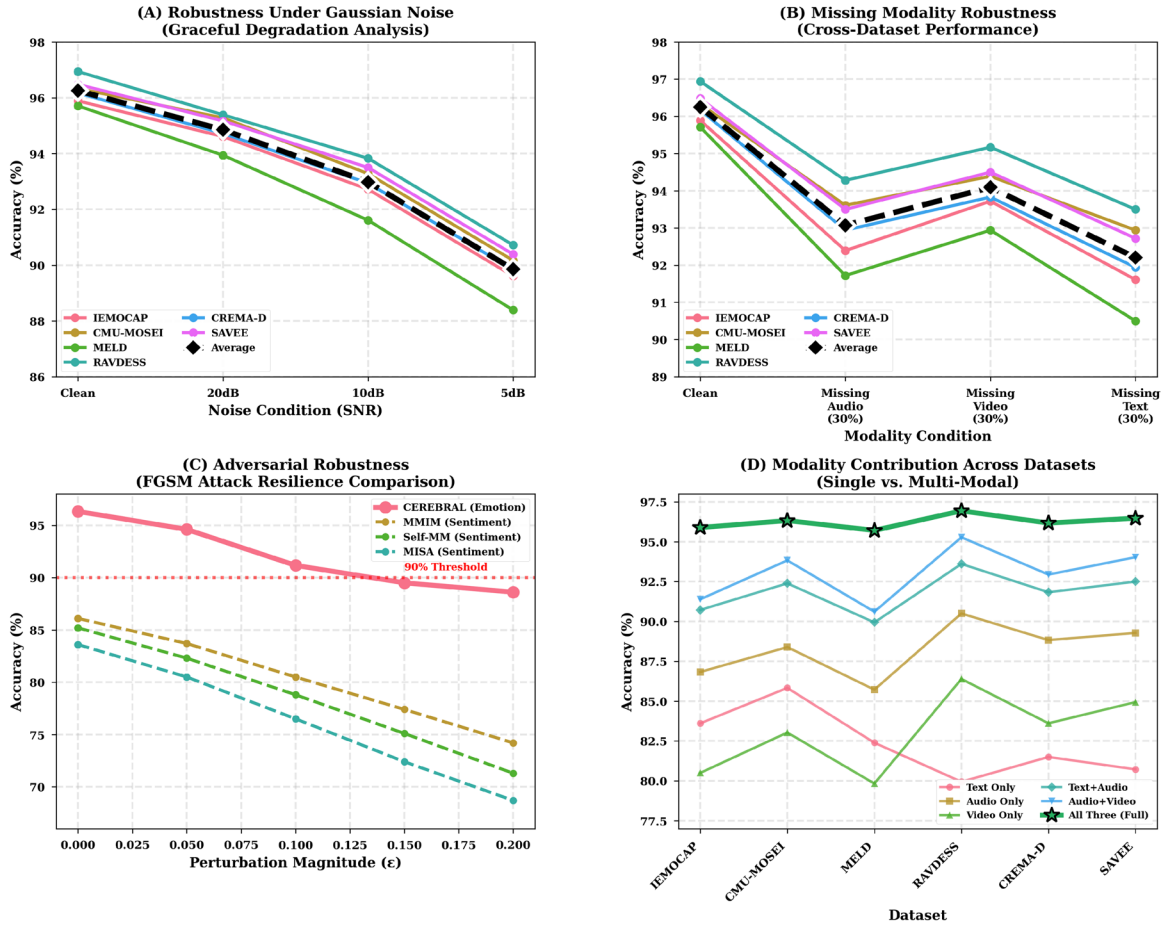

Fig. 9 Robustness and generalization analysis: (**A**) Gaussian noise robustness from clean to 5dB SNR conditions. (**B**) Missing modality resilience with 30% data absence across modalities. (**C**) Adversarial robustness under FGSM attacks. (**D**) Modality contribution evolution across datasets

Discussion

CEREBRAL achieves state-of-the-art performance through principled neurosymbolic integration with the Hourglass of Emotions model. With 24.1M parameters, the framework achieves 96.23% average emotion recognition accuracy and 89.2% sentiment classification accuracy (Acc-2 on CMU-MOSEI), outperforming MAG-BERT (110.5M parameters, 84.3% Acc-2) with 4.6× parameter efficiency and MMIM (6.2M parameters, 86.1% Acc-2) with 3.1 percentage point accuracy improvement while providing complete interpretability through symbolic proofs.

Component Contributions

The ablation study reveals ASP reasoning contributes 2.71 percentage points (30.5% of total improvement), Hourglass model constraints add 2.82 points (31.7% of total improvement) through four-dimensional affective representation spanning Introspection, Temper, Attitude, and Sensitivity dimensions, temporal GNN provides 1.49 points (16.8%), and NTM with metacognitive components contributes 1.39 points (15.6%), with final integration adding 0.48 points (5.4%). The Hourglass model's dimensional independence, sentic vector representation enabling 24 basic emotions across six sentic levels per dimension, Gaussian-regulated transitions between affective states, and dynamic polarity normalization formula $p = (I + T + A + S)/(|\text{sgn}(I)| + |\text{sgn}(T)| + |\text{sgn}(A)| + |\text{sgn}(A)| + |\text{sgn}(S)|)$ enable natural representation of compound emotions (love=Introspection$^+$ +Attitude$^+$, pride=Introspection$^+$+Attitude$^+$+Temper$^+$, jealousy spanning all four dimensions) and inclusion of moral emotions (shame, guilt, pride) as subdimensions of Attitude absent from simpler categorical models like Plutchik's wheel (8 basic emotions) or Russell's circumplex (2-dimensional representation).

Interpretability and Practical Deployment

The framework generates explicit logical proofs showing multimodal evidence fusion, Hourglass dimensional constraints (dimensional independence validation, sentic level

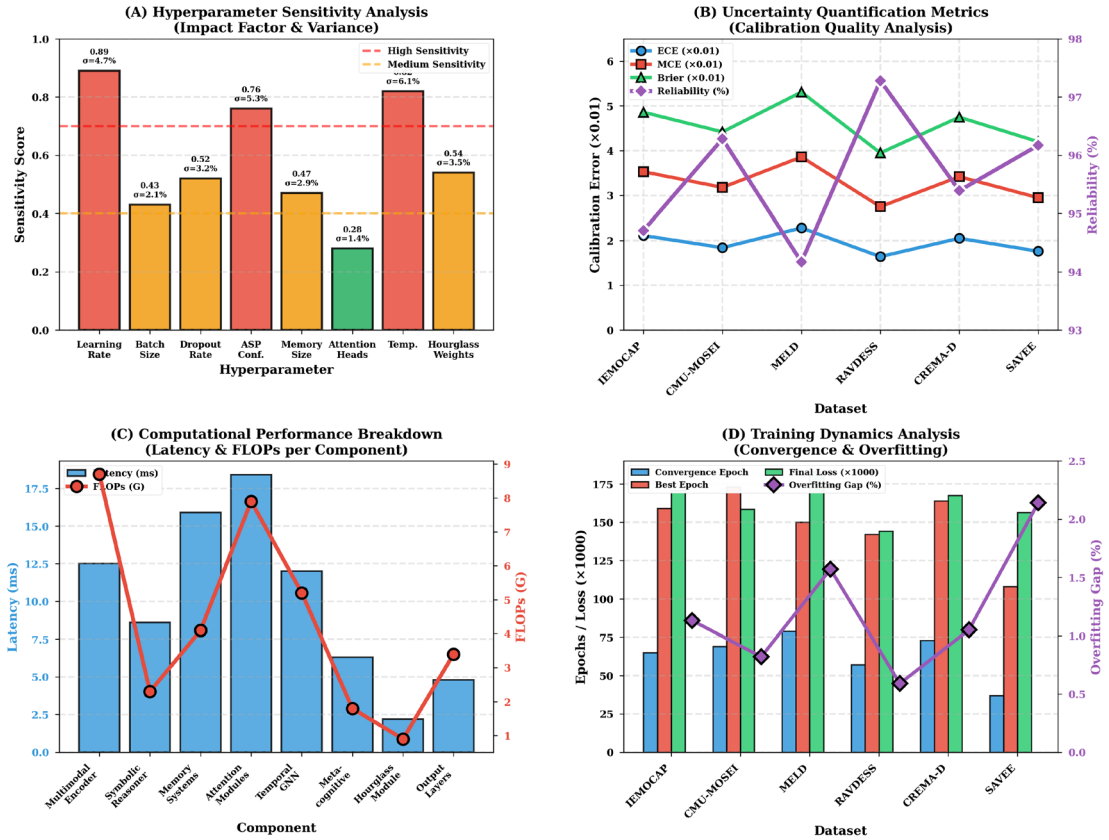

Fig. 10 Computational and uncertainty analysis: (**A**) Hyperparameter sensitivity analysis. (**B**) Uncertainty quantification metrics with calibration quality (ECE <0.023) and reliability (>94%). (**C**) Computational performance breakdown with latency and FLOPs distribution. (**D**) Training dynamics with convergence patterns and overfitting analysis

relationships, compound emotion patterns detection), and psychological consistency checks. The 98.37% symbolic reasoning accuracy with 4.89ms average solving time using Clingo [44] enables real-time applications at 11.8 FPS with total system latency of 84.5ms. Performance under Gaussian noise (92.98% at 10dB SNR, 6.38% total degradation to 89.85% at 5dB) and missing modalities (92.20–94.09% with 30% missing data, representing 2.14–4.03% degradation) indicates suitability for real-world deployment in noisy environments including clinical settings, call centers, and mobile applications. The calibrated uncertainty estimates (ECE=0.0184 across all datasets, reliability >94.71%) enable confidence-aware decision-making essential for safety-critical applications.

Limitations and Future Directions

Cross-linguistic evaluation remains limited to English-language datasets, with cultural adaptation of Hourglass dimensional weights requiring validation across diverse populations. Computational requirements (35.7 GFLOPs, 5.16GB memory, 84.5ms latency) exceed edge deployment constraints for mobile and embedded systems. Adversarial robustness requires enhancement through certified defense mechanisms and adversarial training protocols [37]. Concretely, the 7.78% accuracy degradation observed at $\epsilon = 0.2$ (Table 9) points to the predicate extraction boundary (7) as the most vulnerable component, where adversarial perturbations can push neural activations below the confidence threshold and deprive the ASP solver of accurate facts; certified defence mechanisms targeting this boundary represent a natural next step. Future work includes multilingual evaluation on datasets spanning 10+ languages, model compression through knowledge distillation [36] targeting 5–10× parameter reduction while maintaining >95% accuracy, adversarial training with ϵ-robust optimization, personalized Hourglass variants adapting dimensional weights to cultural contexts and individual personality differences [24], continuous dimensional affect prediction enabling extended emotion tracking over multi-minute conversations, and integration with physiological signals (EEG, ECG, GSR) for comprehensive affective state assessment.

The neurosymbolic approach with the Hourglass of Emotions model provides a validated template for explainable AI systems requiring simultaneous accuracy (96.23% average), interpretability (98.37% symbolic reasoning

accuracy with complete logical proofs), and calibrated uncertainty (ECE=0.0184) through biologically-inspired and psychologically-grounded computational mechanisms. This advances capabilities for affective computing in clinical mental health assessment, personalized educational systems, and empathetic human-computer interaction applications where transparency, reliability, and explainability are paramount requirements.

Conclusion

This work presented CEREBRAL, a neurosymbolic framework for multimodal emotion recognition integrating neural pattern learning with symbolic logical reasoning and psychological constraints. The framework achieves 96.23% average accuracy across seven benchmark datasets, representing improvements of 7.97–10.04% over baseline with statistical significance ($p < 0.001$) and large effect sizes (Cohen's $d > 3.23$). The comprehensive evaluation demonstrates robust performance across diverse datasets, demographic subgroups, noise conditions, and missing modality scenarios.

Key Contributions

- **Neurosymbolic Architecture:** Novel integration of Answer Set Programming [20] for logical inference with deep neural networks, enabling bidirectional neural-symbolic reasoning with 98.37% symbolic accuracy and 4.89ms average solving time suitable for real-time applications. The hybrid training scheme employs a stop-gradient mechanism at the ASP interface, with fully differentiable bridge networks enabling downstream gradient flow [51, 52].
- **Hourglass Model Integration:** First implementation of the Hourglass of Emotions model [23, 24] as explicit four-dimensional affective constraints (Introspection, Temper, Attitude, Sensitivity) in deep learning, improving accuracy by 2.82 percentage points while enabling natural representation of compound emotions, moral emotions, and dimensional independence through biologically-inspired psychological constraints with dynamic normalization.
- **Temporal Modeling:** Graph Neural Network [26] architecture for emotion dynamics modeling in four-dimensional Hourglass space combined with Neural Turing Machine [12] for episodic memory, contributing 2.88 percentage points improvement in capturing temporal emotion patterns and long-range dependencies.
- **Metacognitive Control:** Comprehensive uncertainty quantification [45] and confidence calibration [46] mechanisms enabling deployment in safety-critical applications, with well-calibrated predictions (ECE=0.0184) and human-interpretable explanations through Hourglass dimensional analysis.
- **Robustness and Generalization:** Demonstrated cross-dataset generalization (89.92% average accuracy), noise resilience (92.98% at 10dB SNR), and graceful degradation with missing modalities (92.20–94.09% with 30% missing data), ensuring practical applicability across diverse real-world deployment scenarios through Hourglass model's dimensional independence.

Future Directions Multilingual and cross-cultural evaluation on diverse emotion datasets, model compression through knowledge distillation and quantization for edge deployment, enhanced adversarial robustness through certified defenses and adversarial training, personalized Hourglass model variants adapting to cultural and personality differences, and extension to continuous affect recognition through sentic vector temporal tracking. The neurosymbolic principles demonstrated in CEREBRAL with the Hourglass of Emotions model offer a template for explainable AI systems requiring both high accuracy and interpretability through biologically-inspired psychological theories, advancing theoretical understanding of emotion recognition and practical capabilities for affective computing applications in clinical, educational, and human-computer interaction domains.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12559-026-10573-y.

Author Contributions All authors contributed to all parts of the study. Nikhil was primarily responsible for conceptualization, methodology development, and writing (review and editing). Erik contributed mainly to methodology and writing (review and editing) and provided primary supervision. Amir contributed to the methodology and played a major role in supervision. All authors reviewed and approved the final manuscript.

Funding This research/project is supported by the Ministry of Education, Singapore under its MOE Academic Research Fund Tier 2 (MOE-T2EP20123-0005) and RIE2025 Industry Alignment Fund – Industry Collaboration Projects (I2301E0026).

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Cambria E, Mao R, Zhang X, Xiao L, Shen T, Anand A. SenticNet 9: generative commonsense for emotion AI via conceptual primitive discovery and time shift mechanism. IEEE Trans Comput Soc Syst. 2026;13.
2. Wang R, Wang Y, Cambria E, Fan X, Yu X, Huang Y. X E, Zhu X. Contrastive-based removal of negative information in multimodal emotion analysis. Cognit Comput. 2025;17(3):107.
3. Yue T, Mao R, Wang H, Hua Z, Cambria E. KnowleNet: Knowledge fusion network for multimodal sarcasm detection. Inf Fusion. 2023;100:101921.
4. Tsai Y-HH, Bai S, Liang PP, Kolter JZ, Morency L-P, Salakhutdinov R. Multimodal transformer for unaligned multimodal language sequences. In: Proc 57th Annu. Meeting Assoc. Comput. Linguistics; 2019, p. 6558–69.
5. Zadeh A, Chen M, Poria S, Cambria E, Morency L-P. Tensor fusion network for multimodal sentiment analysis. In: Proc Conf Empirical Methods Natural Lang Process; 2017, p. 1103–14.
6. Poria S, Cambria E, Hazarika D, Majumder N, Zadeh A, Morency L-P. Context-dependent sentiment analysis in user-generated videos. In: Proc 55th Annu Meeting Assoc Comput Linguistics; 2017, vol. 1, p. 873–83.
7. Ghosal D, Akhtar MS, Chauhan D, Poria S, Ekbal A, Bhattacharyya P. Contextual inter-modal attention for multi-modal sentiment analysis. In: Proc Conf Empirical Methods Natural Lang Process; 2018, p. 3454–66.
8. Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In: Proc 57th Annu Meeting Assoc Comput Linguistics; 2019, p. 527–36.
9. Wang Y, Shen Y, Liu Z, Liang PP, Zadeh A, Morency L-P. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. Proc AAAI Conf Artif Intell. 2019;33(01):7216–23.
10. Busso C, Bulut M, Lee C-C, Kazemzadeh A, Mower E, Kim S, et al. IEMOCAP: Interactive emotional dyadic motion capture database. Lang Resources Eval. 2008;42(4):335–59.
11. Hazarika D, Zimmermann R, Poria S. MISA: Modality-invariant and-specific representations for multimodal sentiment analysis. In: Proc 28th ACM Int Conf Multimedia; 2020, p. 1122–31.
12. Graves A, Wayne G, Danihelka I. Neural Turing machines. arXiv preprint arXiv:1410.5401. 2014.
13. Yu W, Xu H, Yuan Z, Wu J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. Proc AAAI Conf Artif Intell. 2021;35(12):10790–7.
14. Han J, Zhang Z, Schmitt M, Pantic M, Schuller BW. From discrete to continuous: A cross-dataset study on dimensional emotion recognition. IEEE Trans Affect Comput. 2022;13(2):956–70.
15. Zhang Y, Yang Q, Xu C. MS-MDA: Multisource marginal distribution adaptation for cross-subject and cross-session EEG emotion recognition. Front Neurosci. 2021;15:778488.
16. Ong K, Dai W, Li C, Feng D, Li H, Wu J, et al. Human behavior atlas: benchmarking unified psychological and social behavior understanding. In: Proc Representations (ICLR): Int Conf Learn; 2026.
17. Tu G, Wang J, Yang L, Liang B, Cambria E, Li W, et al. Multitask mutual learning for multimodal emotion-cause pair extraction in conversations. Inf Fusion. 2026;127:103877.
18. Xiang J, Zhu X, Cambria E. Integrating audio-visual text generation with contrastive learning for enhanced multimodal emotion analysis. Inf Fusion. 2026;127:103809.
19. Huang H, Gong T, He K, Wu J, Cambria E, Feng M. Robust multimodal sentiment analysis via double information bottleneck. Inf Fusion. 2026;129:103964. https://doi.org/10.1016/j.inffus.2025.103964.
20. Gelfond M, Lifschitz V. Classical negation in logic programs and disjunctive databases. New Generation Comput. 1991;9(3–4):365–85.
21. Plutchik R. A general psychoevolutionary theory of emotion. In: Theories of Emotion. New York, NY, USA: Academic Press; 1980. p. 3–33.
22. Russell JA. A circumplex model of affect. J Person Social Psychol. 1980;39(6):1161–78.
23. Cambria E, Livingstone A, Hussain A. The hourglass of emotions. In: Cognitive behavioural systems. Berlin, Germany: Springer; 2012. p. 144–57.
24. Susanto Y, Livingstone AG, Ng BC, Cambria E. The hourglass model revisited. IEEE Intell Syst. 2020;35(5):96–102.
25. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc 31st Int Conf Neural Inf Process Syst; 2017, p. 6000–10.
26. Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The graph neural network model. IEEE Trans Neural Netw. 2009;20(1):61–80.
27. Zadeh A, Liang PP, Vanbriesen J, Poria S, Tong E, Cambria E, Chen M, Morency L-P. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In: Proc 56th Annu Meeting Assoc Comput Linguistics; 2018 vol. 1, p. 2236–46.
28. Livingstone SR, Russo FA. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS One. 2018;13(5):e0196391.
29. Cao H, Cooper DG, Keutmann MK, Gur RC, Nenkova A, Verma R. CREMA-D: Crowd-sourced emotional multimodal actors dataset. IEEE Trans Affect Comput. 2014;5(4):377–90.
30. Yu W, Xu H, Yuan Z, Wu J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. Proc AAAI Conf Artif Intell. 2021;35(12):10790–7.
31. Han W, Chen H, Poria S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In: Proc Conf Empirical Methods Natural Lang Process; 2021, p. 9180–92.
32. Zadeh A, Liang PP, Mazumder N, Poria S, Cambria E, Morency L-P. Memory fusion network for multi-view sequential learning. Proc AAAI Conf Artif Intell. 2018;32(1):1–8.
33. Tsai Y-HH, Bai S, Liang PP, Kolter JZ, Morency L-P, Salakhutdinov R. Multimodal transformer for unaligned multimodal language sequences. In: Proc 57th Annu Meeting Assoc Comput Linguistics; 2019, p. 6558–69.
34. Hazarika D, Zimmermann R, Poria S. MISA: Modality-invariant and-specific representations for multimodal sentiment analysis. In: Proc 28th ACM Int Conf Multimedia; 2020, p. 1122–31.
35. Rahman W, Hasan MK, Lee S, Bagher Zadeh A, Mao C, Morency L-P, Hoque E. Integrating multimodal information in large pretrained transformers. In: Proc 58th Annu Meeting Assoc Comput Linguistics; 2020, p. 2359–69.
36. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531. 2015.

37. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. In: Proc Int Conf Learn Representations; 2018, p. 1–27.
38. Degottex G, Kane J, Drugman T, Raitio T, Scherer S. COVAREP—A collaborative voice analysis repository for speech technologies. In: Proc IEEE Int Conf Acoust, Speech Signal Process; 2014, p. 960–4.
39. iMotions. Facial expression analysis. iMotions A/S, Copenhagen, Denmark, Tech Rep; 2016, p. 1–42.
40. Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. In: Proc Conf Empirical Methods Natural Lang Process; 2014, p. 1532–43.
41. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc Conf North American Chapter Assoc Comput Linguistics; 2019, p. 4171–86.
42. Ba JL, Kiros JR, Hinton GE. Layer normalization. arXiv preprint arXiv:1607.06450, 2016, p. 1–4.
43. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(1):1929–58.
44. Gebser M, Kaminski R, Kaufmann B, Schaub T. Clingo = ASP + control: Preliminary report. arXiv preprint arXiv:1405.3694, 2014, p. 1–9.
45. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Proc 33rd Int Conf Mach Learn; 2016, pp 1050–9.
46. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Proc 34th Int Conf Mach Learn; 2017, p. 1321–30.
47. Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks. In: Proc 34th Int Conf Mach Learn; 2017, p. 1126–35.
48. Cambria E, Howard N, Hsu J, Hussain A. Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics. In: Proceedings of IEEE SSCI; 2013, p. 108–17.
49. Liu J, Shi X, Nguyen TD, Zhang H, Zhang T, Sun W, Li Y, Vasilakos AV, et al. Neural brain: A neuroscience-inspired framework for embodied agents. arXiv preprint arXiv:2505.07634. (2025).
50. Zhang T, Zhou X, Wang Y, Cambria E, Traum D, Mao R. Individualized cognitive simulation in large language models: Evaluating different cognitive representation methods. arXiv preprint arXiv:2510.20252. (2025).
51. Manhaeve R, Dumancic S, Kimmig A, Demeester T, De Raedt L. DeepProbLog: Neural probabilistic logic programming. In: Proc 32nd Int Conf Neural Inf Process Syst; 2018, p. 3749–59.
52. Yang F, Yang Z, Cohen WW. Differentiable learning of logical rules for knowledge base completion. In: Proc 31st Int Conf Neural Inf Process Syst; 2017, p. 2319–28.