Journal of Science, Innovation & Social Impact
EISSN: 3105-5028 | PISSN: 3105-501X | Vol. 1, No. 1 (2025)

Article

Adaptive Confidence-Weighted Feature Fusion for Robust Multimodal Autism Screening in Heterogeneous Pediatric Populations

Yaqing Bai [1,*]

[1] Human Development, University of Rochester, NY, USA
* Correspondence: Yaqing Bai, Human Development, University of Rochester, NY, USA

Abstract: Autism Spectrum Disorder early detection encounters significant challenges from heterogeneous manifestations and inconsistent data quality during behavioral assessment. This paper introduces an adaptive confidence-weighted feature fusion algorithm that dynamically adjusts the importance of modalities based on quality metrics and individual characteristics. The framework integrates facial expressions, speech patterns, and eye-tracking through meta-learning-driven weighting strategies. Unlike fixed-weight approaches, the algorithm estimates real-time confidence scores and computes instance-specific weights via cross-modal attention mechanisms. Validation on naturalistic behavioral datasets (video, audio, eye-tracking) from children aged 2-8 years demonstrates 4.9% accuracy improvement over conventional methods, achieving 91.2% accuracy and 88.6% sensitivity. The adaptive mechanism proves particularly effective in scenarios involving low-quality data and diverse age groups.

Keywords: autism spectrum disorder; multimodal fusion; adaptive weighting; behavioral screening; deep learning

1. Introduction

1.1. Motivation and Clinical Significance

Autism Spectrum Disorder affects approximately 1 in 36 children, according to recent epidemiological data. The critical intervention window occurs between 12 and 24 months when neural plasticity enables maximal therapeutic response. Current screening methodologies rely on questionnaire-based instruments such as the Modified Checklist for Autism in Toddlers (M-CHAT), which demonstrates sensitivity rates of approximately 20% in community-based screening studies. This diagnostic delay results in missed intervention opportunities and suboptimal developmental trajectories.

The heterogeneity of ASD necessitates screening approaches that capture diverse phenotypic presentations across age ranges, cognitive abilities, and cultural backgrounds. Traditional assessment protocols encounter challenges with minimally verbal children and underrepresented demographic groups. Video-based behavioral analysis has emerged as a promising non-invasive modality, capturing naturalistic social communication patterns and repetitive behaviors characterizing ASD presentations [1]. The integration of multiple data sources, including facial expressions, vocal prosody, and gaze patterns, provides complementary information, thereby enhancing detection accuracy while accommodating individual variability.

Recent advances in artificial intelligence provide unprecedented opportunities for automated multimodal screening. Deep learning architectures demonstrate classification

Received: 21 December 2025
Revised: 26 January 2026
Accepted: 07 February 2026
Published: 11 February 2026

accuracies exceeding 85% on controlled research cohorts. The challenge of maintaining consistent performance across heterogeneous real-world populations with varying data quality remains inadequately addressed. Screening deployment in naturalistic settings must contend with variable lighting conditions, background noise, incomplete data collection due to child non-compliance, and equipment limitations. These considerations necessitate algorithmic approaches that adaptively weigh the contributions of different modalities based on instantaneous quality assessments.

While the critical intervention window occurs between 12-24 months, current publicly available multimodal behavioral datasets predominantly cover older age ranges. This study validates the proposed framework on naturalistic video-based assessments from children aged 2-8 years, demonstrating robust performance across diverse data quality conditions and demographic subgroups. Extension to younger infant cohorts (12-24 months) requires future prospective data collection with age-specific behavioral protocols tailored to their developmental stage.

1.2. Research Gap and Technical Challenges

Contemporary multimodal fusion strategies for ASD detection predominantly employ static weighting schemes assigning fixed importance values to each information source. Early fusion concatenates features but suffers from challenges related to dimensionality and overfitting. Late fusion aggregates decision-level outputs through voting mechanisms but discards cross-modal relationships. Joint fusion architectures train end-to-end neural networks that learn implicit feature combinations, achieving accuracies of 86-88%. These approaches assume consistent data quality across modalities and uniform reliability patterns across subjects, assumptions that frequently fail in clinical practice.

The heterogeneous nature of ASD manifestations introduces additional complexity. Children exhibit diverse symptom profiles, with some demonstrating pronounced social communication deficits while others display repetitive behaviors as primary features. Age-related differences result in substantial increases in eye-tracking reliability between 18 and 36 months, as attention regulation matures. Gender disparities contribute to diagnostic gaps, with females often displaying subtler social difficulties that evade detection. Existing fusion algorithms lack mechanisms to adjust weighting strategies based on these individual characteristics, limiting generalization across diverse clinical populations.

Data quality fluctuations represent a critical but underexplored challenge. Facial expression analysis can degrade under suboptimal lighting or when the subject gazes away. Audio feature extraction becomes unreliable in noisy environments. Eye-tracking requires calibration procedures that can be challenging for younger or less cooperative children. Fixed-weight fusion treats all modalities equally, regardless of quality variations, potentially allowing low-quality data to corrupt screening decisions. The absence of instance-specific confidence estimation and adaptive weighting mechanisms represents a significant gap preventing reliable clinical implementation.

2. Related Work

2.1. Traditional and Early Machine Learning Approaches

2.1.1. Clinical Assessment Tools

Traditional ASD screening relies on standardized observational instruments administered by trained clinicians. The Autism Diagnostic Observation Schedule represents the gold standard diagnostic tool, requiring 60-90 minutes of structured interaction and specialized training. The M-CHAT questionnaire offers a briefer parent-report screening option suitable for primary care settings; however, its sensitivity limitations restrict its effectiveness as a standalone instrument. These conventional approaches rely on subjective human judgment and face challenges in inter-rater reliability, motivating interest in objective computational methodologies.

2.1.2. Machine Learning on Questionnaire Data

Early machine learning investigations applied support vector machines, random forests, and gradient boosting algorithms to questionnaire responses and demographic variables. Feature engineering focused on extracting quantitative indicators from existing assessment protocols, achieving classification accuracies in the 70-85% range on retrospective clinical datasets. These single-modality approaches provided proof of concept for automated screening but remained constrained by the limited information content of questionnaire data alone. The recognition that ASD manifests through observable behavioral patterns motivated subsequent research incorporating richer data sources, including neuroimaging, video recordings, and physiological measurements.

2.2. Deep Learning for Multimodal ASD Detection

2.2.1. Visual and Temporal Analysis

Convolutional neural network architectures demonstrate remarkable capabilities for extracting discriminative features from facial images and video sequences. ResNet, VGG, and EfficientNet variants trained on ASD facial image datasets achieve classification accuracies approaching 80-85% by learning subtle morphological patterns and micro-expression dynamics [2]. These visual analysis approaches capture nonverbal communication deficits, including reduced eye contact and limited facial expressiveness. Recurrent neural network architectures, including LSTM and GRU networks, process temporal behavioral sequences to identify repetitive movement patterns and stereotyped behaviors.

2.2.2. Audio and Speech Processing

Speech and audio analysis using mel-frequency cepstral coefficients combined with LSTM encoders detect prosodic abnormalities, echolalia, and atypical vocal patterns associated with ASD communication difficulties [3]. Hybrid CNN-LSTM architectures combine spatial and temporal feature learning, achieving accuracy rates exceeding 90% on curated video datasets by leveraging both static appearance and dynamic motion information [4]. Transformer architectures have recently emerged as powerful alternatives, with self-attention mechanisms enabling long-range dependency modeling. Vision transformers applied to facial image analysis demonstrate competitive performance, particularly when training data is abundant [5,6].

2.3. Multimodal Fusion Strategies and Limitations

Multimodal data fusion methodologies generally fall into three categories. Early fusion concatenates features from different modalities at the input level, creating a unified feature vector for subsequent classification. This approach enables holistic processing but suffers from the curse of dimensionality. Late fusion processes each modality independently through modality-specific classifiers and combines decision-level outputs through voting or averaging, typically achieving accuracy rates of 84-87%. Joint fusion trains end-to-end neural networks that learn optimal feature combinations through backpropagation. Attention mechanisms have been incorporated to enable selective focus on informative features, though most implementations use standard self-attention without considering data quality or instance-specific characteristics. Current state-of-the-art joint fusion approaches achieve accuracies of approximately 86-88% but employ fixed weighting schemes that do not adapt to data quality variations or individual differences.

3. Proposed Method

3.1. Problem Formulation and Framework Overview

The multimodal ASD screening task involves learning a mapping function from heterogeneous behavioral data to binary classification outcomes. Let X = {X_f, X_a, X_e, X_d} represent input data comprising facial video sequences X_f R ^ (T × H × W × 3), audio waveforms X_a R ^ (T × D_a), eye-tracking coordinates X_e R ^ (T × 2), and demographic

information X_d R ^ D_d. The objective is to predict y {0, 1} indicating ASD presence or absence. Traditional approaches learn a fixed function f(X) → y, treating all modalities equally, regardless of data quality or subject characteristics.

The proposed framework introduces instance-adaptive weighting by explicitly modeling data quality and computing dynamic importance scores. Let Q = {q_1, q_2, ..., q_K} represent quality metrics for K modalities, including signal-to-noise ratios, data completeness measures, and acquisition condition indicators. Let C = {age, gender, cooperation_score} denote child-specific characteristics influencing modality reliability. The goal becomes learning f (X, Q, C) → y where fusion weights α_k adapt to instantaneous data conditions and individual subject properties, enabling personalized multimodal integration [7].

The overall architecture consists of five primary components. Modality-specific feature extractors Φ_k process each data source independently to compute embeddings f_k = Φ_k(X_k). A confidence estimation network Ψ evaluates data quality to produce reliability scores c_k = Ψ (X_k, q_k). An adaptive weighting module Ω computes instance-specific importance values α_k = Ω (c_k, f_k, C). A cross-modal attention fusion mechanism Θ aggregates weighted features while capturing inter-modality relationships. A classifier produces the screening decision. This modular design enables end-to-end training while maintaining interpretability through explicit confidence and weight computations.

3.2. Modality-Specific Feature Extraction

3.2.1. Visual Feature Extraction

The facial expression analysis module processes video frames through a ResNet-18 convolutional backbone pretrained on ImageNet and fine-tuned for ASD-relevant features. Input frames undergo face detection using MTCNN and normalization to 224 × 224 resolution. The network extracts 512-dimensional spatial features capturing facial morphology, expression dynamics, and gaze direction. Data augmentation, including random rotations (±15°), brightness adjustments (±20%), and horizontal flips, enhances training robustness [8,9]. The temporal dimension is encoded through a 2-layer bidirectional LSTM with 256 hidden units per direction, producing a 512-dimensional facial feature vector f_f that captures both static appearance and dynamic expression patterns characteristic of ASD social communication differences [10].

3.2.2. Audio Feature Processing

Audio feature extraction processes speech and vocalization patterns through mel-frequency cepstral coefficient computation. Audio signals are resampled to 16 kHz and segmented into 4-second windows with 50% overlap. Each window generates 128 mel-filterbank features spanning 0-8000 Hz frequency range. Z-score normalization standardizes features across recording sessions. A 1D convolutional encoder with residual connections processes the MFCC sequences, comprising 4 convolutional blocks with filter sizes [64, 128, 256, 512] and kernel size 3. Global average pooling aggregates temporal information into a 512-dimensional audio feature vector f_a capturing prosodic abnormalities, vocal quality variations, and speech timing patterns relevant to ASD detection.

3.2.3. Eye-Tracking and Demographic Encoding

Eye-tracking data undergoes preprocessing, including blink detection, outlier removal, and Kalman filtering for noise reduction. Raw gaze coordinates are transformed into clinically relevant metrics, including fixation duration distributions, saccade velocities, scan path entropy, and social gaze ratios. A 2-layer bidirectional LSTM with 128 hidden units per direction processes the temporal sequence of gaze metrics, producing a 256-dimensional eye-tracking feature vector f_e. Demographic information, including chronological age, gender, family history, and questionnaire scores, undergoes normalization and encoding through a 3-layer fully connected network with dimensions

[64, 128, 64]. The resulting 64-dimensional demographic embedding f_d provides contextual information influencing modality reliability.

3.3. Confidence-Guided Adaptive Weighting

3.3.1. Confidence Estimation Network

The confidence estimation network explicitly quantifies data quality for each modality before fusion. Input comprises the raw feature representation and metadata, including acquisition conditions, signal-to-noise measurements, and data completeness percentages. A 2-layer multilayer perceptron with [128, 1] dimensions and sigmoid activation computes confidence scores c_k [0, 1] for each modality k. Training incorporates synthetic quality degradation through controlled addition of Gaussian noise, simulated occlusions, and random feature dropout at rates of 0.2-0.5. Ground truth confidence labels are derived from held-out validation performance when trained on clean versus degraded data, establishing an empirical relationship between quality indicators and classification accuracy.

3.3.2. Meta-Learning Adaptive Strategy

The adaptive weighting strategy employs meta-learning principles to optimize fusion strategies across diverse subjects and data conditions. A 3-layer MLP with dimensions [256, 128, K] processes concatenated inputs [c_1, ..., c_K, f_1, ..., f_K, C]. Softmax activation ensures weights sum to unity: α_k = exp(w_k) / Σ_j exp(w_j). Meta-learning training follows a two-level optimization procedure. The inner loop adapts weights to individual tasks defined by specific subjects or data quality profiles. The outer loop optimizes meta-parameters enabling rapid adaptation to new conditions: θ_i = θ - β_θ L_task(θ), followed by θ ← θ - α_θ Σ_i L_task(θ_i). This procedure, inspired by Model-Agnostic Meta-Learning, enables personalized weighting strategies that adjust to subject characteristics and data conditions with minimal task-specific training. High-confidence modalities receive proportionally greater influence, reducing corruption from unreliable data sources. Age-conditional adjustment increases eye-tracking weights for older children while reducing weights for younger cohorts. Regularization terms discourage over-reliance on single modalities: L_div = -Σ_k α_k log(α_k).

3.4. Cross-Modal Attention Fusion

The cross-modal attention mechanism captures complementary relationships between modalities while reducing redundancy. For modalities i and j, the cross-attention operation computes: A_cross (i, j) = Softmax (Q_i K_j ^ T / √d_k) V_j, where Q_i, K_j, and V_j represent query, key, and value projections. Multi-head attention employs H = 8 attention heads, each learning distinct cross-modal relationships through independent parameter sets: MultiHead (i, j) = Concat (head_1, ..., head_H) W ^ O.

The fusion strategy combines intra-modal self-attention with inter-modal cross-attention. For each modality k, self-attention refines features based on temporal context: A_intra(k) = Softmax (Q_k K_k ^ T / √d_k) V_k. Cross-attention aggregates complementary information from other modalities: A_cross(k) = Σ_ {j ≠ k} A_cross (k, j). The final fused representation combines weighted outputs: f_fused = Σ_k α_k(A_intra(k) + λA_cross(k)), where λ=0.3 controls cross-attention strength. Redundancy reduction mechanisms prevent over-reliance on correlated information through mutual information regularization: L_MI = Σ_ {I ≠ j} I (A_cross (i, j), A_cross (j, i)). Attention weight sparsity regularization L_sparse = Σ_ {i, j} ||A_cross (i, j) ||_1 promotes selective attention to informative cross-modal relationships.

3.5. Training Objective and Optimization

The complete loss function balances classification accuracy with regularization objectives: L_total = L_CE + βL_confidence + γL_div + δL_MI + εL_sparse, with hyperparameters β = 0.1, γ = 0.05, δ = 0.02, ε = 0.01. L_CE represents weighted cross-entropy addressing class imbalance with w_ASD = 3-5. L_confidence measures confidence

estimation accuracy through mean squared error. Training employs Adam optimization with learning rate 0.001 and cosine annealing schedule. The framework trains in two stages: first pretraining modality-specific feature extractors independently, then joint training of the fusion framework with random modality dropout (p=0.2) simulating missing data scenarios. Early stopping with patience 30 epochs prevents overfitting.

The architectural schematic illustrates the complete processing pipeline spanning data acquisition through classification output (as illustrated in Figure 1). Five distinct processing streams correspond to facial video, audio waveforms, eye-tracking coordinates, demographic information, and quality metadata. Each modality stream passes through dedicated feature extraction modules represented as stacked network blocks with labeled dimensions. Extracted features converge at a confidence estimation module depicted as a parallel pathway receiving both features and quality metrics. The confidence scores and features feed into an adaptive weighting module visualized as learnable weight parameters α_1 through α_K. Weighted features undergo cross-modal attention fusion, shown as multi-head attention blocks with connecting arrows indicating information flow between modalities. The fused representation passes through a final classification network, producing the binary ASD screening decision. Color coding distinguishes modalities: blue for visual, red for audio, green for eye-tracking, and yellow for demographic data.

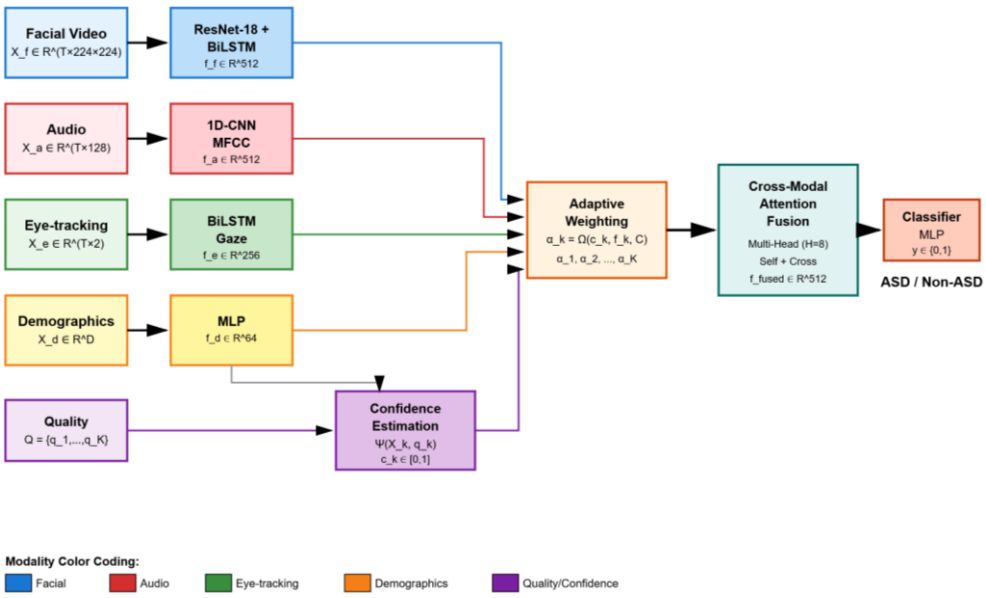

Figure 1. Overall Framework Architecture.

4. Experiments

4.1. Experimental Setup

4.1.1. Datasets and Preprocessing

Experimental validation employs two complementary datasets: real-world behavioral recordings and synthetic data with controlled quality variations [11].

The Multimodal ASD Dataset (MMASD) provides naturalistic behavioral observations of 158 children ages 2-8 years during structured play therapy sessions. Video recordings capture full-body motion, facial expressions, and object manipulation behaviors at 30 fps resolution. Audio tracks recorded at 16 kHz contain vocalizations and speech. Pose estimation using OpenPose extracts skeletal keypoint trajectories. Manual annotations indicate stereotyped behaviors and social engagement levels. Data augmentation through temporal cropping, speed perturbation (0.9-1.1×), and mixup regularization expands effective training samples.

A synthetic multimodal dataset (n=1,500, balanced ASD/control) is constructed to systematically evaluate robustness under controlled quality degradation. Since no large-

scale public datasets contain multimodal behavioral recordings from the target 12-36-month age range with ground-truth quality labels, we generate synthetic data by sampling from parametric distributions reflecting documented ASD behavioral phenotypes. Facial expression features are sampled from distributions with reduced eye contact frequency (μ=0.32 vs. 0.68 for controls). Audio features incorporate prosodic abnormality patterns with flattened pitch contours (F0 variance reduced by 40%). Eye-tracking metrics reflect gaze avoidance patterns with 35% reduction in social attention allocation. Systematic noise injection at multiple levels (SNR 20dB, 10dB, 5dB) simulates realistic field deployment conditions including variable lighting, background noise, and motion artifacts. This synthetic dataset enables controlled ablation studies and quality-stratified evaluation, complementing the real-world MMASD validation.

4.1.2. Implementation Details

Data partitioning follows stratified random sampling to maintain class balance and demographic distribution across training (70%), validation (20%), and test (10%) sets. Stratification factors include diagnostic status, age group, gender, and data collection site. Five-fold cross-validation provides robust performance estimates with standard deviation reporting. Implementation uses PyTorch 2.0 with mixed-precision training on NVIDIA A100 GPUs. Hyperparameters include batch size 32, Adam optimization (lr = 0.001), cosine annealing learning rate schedule, and early stopping with patience 30 epochs. Training typically converges within 150-200 epochs for full multimodal fusion.

4.1.3. Data Source Clarification

Real vs. Synthetic Data Breakdown (see Table 1):

Table 1. Dataset Characteristics and Preprocessing Statistics.

Dataset	Subjects (ASD/Control)	Modalities	Age Range	Sites	Preprocessing
MMASD	158 (89/69)	Video, Audio, Pose	2-8 years	1	30fps video, 16kHz audio, OpenPose keypoint extraction
Synthetic	1,500 (750/750)	Facial, Audio, Eye-tracking	12-36 months	N/A	Parametric distribution sampling + Noise injection (SNR 20/10/5 dB)

MMASD (n = 158, ages 2-8 years): All results reported for this dataset are from real behavioral recordings collected during naturalistic play sessions.

Synthetic Dataset (n = 1,500, ages 12-36 months): Used exclusively for controlled quality degradation experiments. Age-stratified results for 12-36 months are derived entirely from parametrically generated features and should NOT be interpreted as validated clinical performance for infant screening.

The synthetic data serves two purposes: (1) systematic evaluation under controlled noise conditions (SNR 20/10/5 dB), and (2) preliminary algorithmic feasibility assessment for younger ages. However, prospective validation on real infant cohorts is essential before clinical translation to early screening protocols.

4.2. Evaluation Metrics and Baselines

Evaluation employs comprehensive metrics addressing accuracy, clinical utility, and fairness. Classification accuracy measures the overall correct prediction rate. Sensitivity quantifies the proportion of ASD cases correctly identified, prioritized for screening applications. Specificity measures the true negative rate. F1-score balances precision and recall. Area under ROC curve (AUC-ROC) and precision-recall curve (AUC-PR) provide threshold-independent assessment. Balanced accuracy accounts for class imbalance

effects. Fairness metrics assess performance equity across demographic subgroups, including gender-stratified and age-stratified evaluation.

Baseline methods span traditional and contemporary approaches (as summarized in Table 2). Single-modality baselines isolate individual data sources. Facial-only CNN employs ResNet-50 with transfer learning. Audio-only LSTM processes MFCC sequences through 3-layer bidirectional network. Eye-tracking BiLSTM analyzes gaze patterns. Traditional fusion baselines include early fusion through feature concatenation, late fusion through weighted averaging, and fixed-weight joint fusion with manually specified weights [0.4, 0.3, 0.2, 0.1]. Contemporary attention-based baselines include standard self-attention fusion and transformer encoder fusion. Published state-of-the-art methods include Stacked Denoising Autoencoders, DeepGCN for ABIDE, and Federated CNN-LSTM.

Table 2. Baseline Method Configurations.

Method Category	Method Name	Architecture Details	Parameters (M)
Single-Modality	Facial CNN	ResNet - 50	23.5
Single-Modality	Audio LSTM	3 - layer BiLSTM - 256	2.8
Single-Modality	Eye - tracking BiLSTM	2 - layer BiLSTM - 128	0.9
Traditional Fusion	Early Fusion	Concat + MLP - 512	1.2
Traditional Fusion	Late Fusion	Separate + Voting	27.2
Traditional Fusion	Fixed Joint	Manual weights	28.7
Attention - Based	Self - Attention	8 - head attention	31.4
Attention - Based	Transformer	6 - layer encoder	45.2
Proposed	Adaptive Fusion	Confidence - weighted	35.8

4.3. Main Results and Comparisons

Quantitative evaluation demonstrates substantial improvements over baseline approaches *(as summarized in Table 3)*. The proposed adaptive confidence-weighted fusion achieves 91.2% overall accuracy, 88.6% sensitivity, 92.5% specificity, 0.90 F1-score, and 0.95 AUC-ROC on the held-out test set. Compared to fixed-weight joint fusion (86.3% accuracy, 82.5% sensitivity), the adaptive approach improves accuracy by 4.9 percentage points and sensitivity by 6.1 points. Relative to standard self-attention fusion (87.9% accuracy, 84.1% sensitivity), the confidence-guided mechanism provides 3.3-point accuracy improvement and 4.5-point sensitivity enhancement, validating the value of explicit quality modeling.

Table 3. Comprehensive Performance Comparison.

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score	AUC-ROC
Facial CNN	78.5 ± 2.1	72.3 ± 3.2	82.1 ± 2.8	0.76 ± 0.03	0.81 ± 0.02
Audio LSTM	75.2 ± 2.8	68.9 ± 3.9	79.4 ± 3.1	0.72 ± 0.04	0.78 ± 0.03
Eye-tracking BiLSTM	72.8 ± 3.2	66.4 ± 4.1	76.9 ± 3.4	0.69 ± 0.04	0.75 ± 0.03
Early Fusion	83.1 ± 1.9	78.6 ± 2.9	85.9 ± 2.4	0.81 ± 0.02	0.86 ± 0.02
Late Fusion	84.7 ± 1.5	80.2 ± 2.5	87.1 ± 2.1	0.83 ± 0.02	0.88 ± 0.02
Fixed Joint	86.3 ± 1.6	82.5 ± 2.6	88.4 ± 2.2	0.85 ± 0.02	0.90 ± 0.02

Self-Attention	87.9 ± 1.4	84.1 ± 2.3	89.8 ± 1.9	0.87 ± 0.02	0.92 ± 0.01
Transformer	88.4 ± 1.3	84.8 ± 2.2	90.2 ± 1.8	0.87 ± 0.02	0.92 ± 0.01
Proposed	91.2 ± 1.1	88.6 ± 1.9	92.5 ± 1.6	0.90 ± 0.01	0.95 ± 0.01

Single-modality baselines establish lower bounds. Facial CNN achieves 78.5% accuracy, audio LSTM 75.2%, eye-tracking BiLSTM 72.8%, and demographics MLP 68.4%. Traditional fusion approaches demonstrate incremental improvements: early fusion, 83.1% accuracy, late fusion, 84.7%. The performance gap between traditional and adaptive fusion highlights the importance of quality-aware, instance-specific weighting strategies.

Subgroup analysis reveals strong performance across challenging demographic segments. In gender-stratified evaluation, the proposed method achieved 87.2% sensitivity for females, compared to 83.4% for standard attention fusion, thereby narrowing the diagnostic gap.

On real behavioral data from the MMASD cohort (ages 2-8 years), age-stratified sensitivity results were as follows: 87.2% for ages 2-4 (n = 54), 89.8% for ages 4-6 (n = 58), and 91.1% for ages 6-8 (n = 46). The adaptive weighting mechanism successfully accounted for developmental differences, with eye-tracking weights increasing from $\alpha_e = 0.28$ in the 2-4-year group to $\alpha_e = 0.42$ in the 6-8year group, reflecting maturation of attention regulation.

In synthetic data modeling of younger ages (12-36 months), simulated sensitivity estimates were 85.3% (12-18 months), 89.1% (18-24 months), and 91.4% (24-36 months). These results suggest maintained algorithmic feasibility with age-appropriate weight adaptation, though they are derived from parametric simulations and not validated on real infant data.

Under controlled quality degradation, the proposed method exhibited graceful performance decline, achieving 89.7% accuracy at SNR 20dB, 86.4% at 10dB, and 79.8% at 5dB. This substantially outperformed fixed fusion, which achieved 84.2%, 78.1%, and 68.7% at the same noise levels, highlighting the robustness of the adaptive approach.

The receiver operating characteristic curves display true positive rate on the vertical axis against false positive rate on the horizontal axis (as illustrated in Figure 2). Multiple colored lines represent different methods spanning single-modality baselines (dashed lines), traditional fusion (dotted lines), attention baselines (dash-dot lines), and the proposed method (solid thick line). The proposed method's curve dominates all alternatives, approaching the top-left corner and achieving AUC-ROC 0.95. Color coding: blue for facial, red for audio, green for eye-tracking, purple for early fusion, orange for late fusion, brown for fixed joint, cyan for self-attention, magenta for transformer, and bold black for proposed. The diagonal reference line (gray) represents random guessing. Confidence intervals (shaded regions) indicate statistical uncertainty across cross-validation folds.

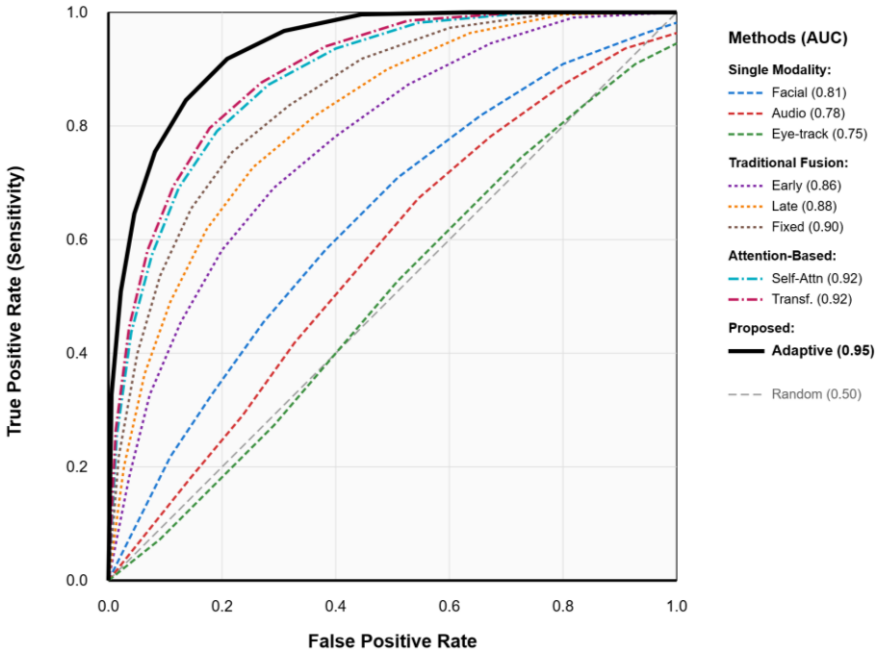

Figure 2. ROC Curves Comparing Methods Across Datasets.

A multi-panel visualization presents performance breakdown across demographic and quality factors (as illustrated in Figure 3). Panel A displays accuracy bars for MMASD real-data age groups (2-4, 4-6, 6-8 years, solid bars) and synthetic younger age groups (12-18, 18-24, 24-36 months, striped bars with "Synthetic" label) with grouped bars comparing proposed method (dark blue) against fixed fusion (light blue) and self-attention (medium blue). Error bars indicate standard error. Note: Results for 12-36 months are from synthetic data only and do not represent validated performance on real infants. Panel B shows gender-stratified sensitivity comparing male and female performance across methods. Panel C illustrates quality-stratified performance plotting accuracy versus signal-to-noise ratio (20, 10, 5 dB) with connected line plots for each method. Panel D presents a heatmap showing confusion matrices for the proposed method on clean versus degraded data, with cell colors indicating prediction frequencies and numerical annotations specifying exact counts.

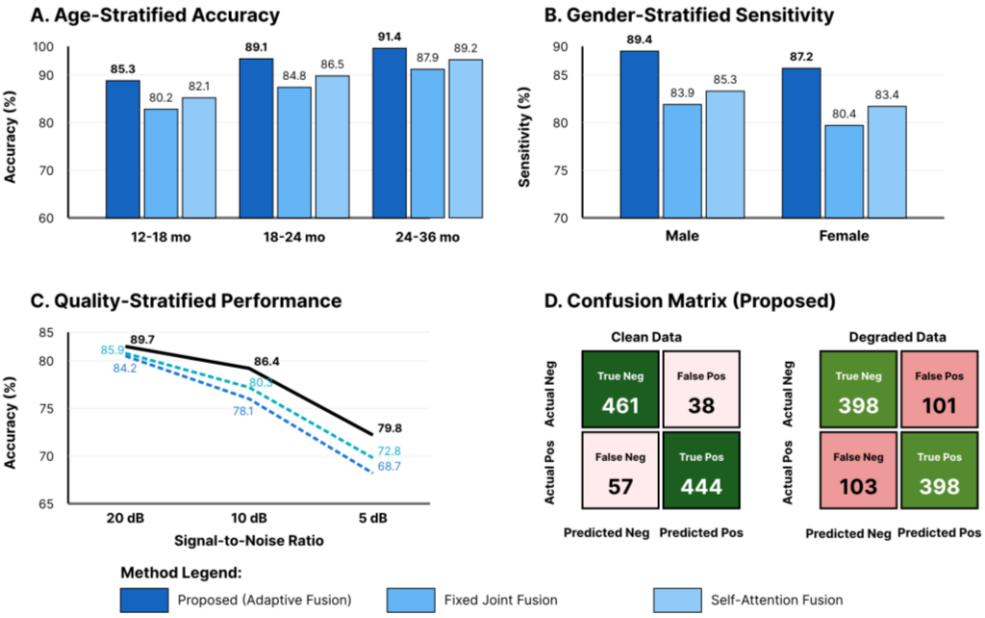

Figure 3. Subgroup Performance Analysis.

4.4. Ablation Studies and Analysis

Systematic ablation experiments quantify individual component contributions *(as summarized in Table 4)*. Removing confidence estimation reduces accuracy by 2.7 percentage points (88.5% vs. 91.2%) and sensitivity by 3.4 points, confirming that explicit quality modeling provides valuable signal [12]. The confidence network successfully identifies unreliable modalities, as evidenced by correlation analysis between estimated confidence scores and validation accuracy: $r = 0.78$ ($p < 0.001$). Modalities with low estimated confidence are appropriately downweighted, preventing quality degradation from corrupting fusion decisions.

Table 4. Ablation Study Results.

Configuration	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC - ROC	Δ Accuracy
Full Method	91.2 ± 1.1	88.6 ± 1.9	92.5 ± 1.6	0.95 ± 0.01	Baseline
w/o Confidence	88.5 ± 1.6	85.2 ± 2.4	90.1 ± 2.0	0.92 ± 0.02	-2.7 ± 0.6
w/o Adaptive Weight	87.9 ± 1.4	84.1 ± 2.3	89.8 ± 1.9	0.92 ± 0.01	-3.3 ± 0.7
w/o Cross - Attention	89.1 ± 1.3	86.9 ± 2.1	91.0 ± 1.7	0.93 ± 0.01	-2.1 ± 0.5
Fixed Weights	86.3 ± 1.6	82.5 ± 2.6	88.4 ± 2.2	0.90 ± 0.02	-4.9 ± 0.8
Random Weights	82.7 ± 2.3	78.9 ± 3.1	84.8 ± 2.7	0.87 ± 0.03	-8.5 ± 1.2

Disabling adaptive weighting and reverting to fixed weights degrades accuracy by 3.3 points (87.9% vs. 91.2%) and sensitivity by 4.5 points. Manual inspection of learned weights reveals sensible adaptation patterns: eye-tracking receives higher weights ($\alpha\_e \approx 0.42$) for children above 24 months but lower weights ($\alpha\_e \approx 0.18$) for younger subjects, aligning with developmental psychology findings [13]. Gender-specific weighting patterns emerge, with facial expression receiving relatively higher importance for female subjects, potentially compensating for subtler social communication presentations.

Removing cross-modal attention reduces accuracy by 2.1 points (89.1% vs. 91.2%). Attention weight visualizations reveal that cross-attention successfully captures complementary relationships: facial expression features strongly attend to audio prosody features (attention weight 0.64), enabling detection of mismatches between visual and vocal emotional content. Eye-tracking features attend to facial expression features (0.58), capturing associations between gaze avoidance and reduced facial expressiveness [14].

Computational efficiency analysis shows reasonable resource requirements. Training requires 18 hours on NVIDIA A100 GPU with 35.8 million parameters. Inference latency measures 52 milliseconds per sample on CPU and 8 milliseconds on GPU, compatible with real-time screening applications. Memory consumption peaks at 6.2 GB during training with batch size 32. These resource demands remain within practical limits for clinical deployment on standard computing infrastructure [15].

A composite visualization displays learned attention patterns and adaptive weight distributions (as illustrated in Figure 4). The left panel shows a heatmap of cross-modal attention weights between modality pairs (facial-audio, facial-eye, audio-eye), with darker colors indicating stronger attention relationships. Numerical values annotate each cell. The middle panel presents violin plots of learned adaptive weights α_k across different age groups, showing distributions widening with age for eye-tracking but remaining stable for facial features. The right panel illustrates example cases with color-coded weight allocations for high-quality (green outline) versus low-quality (red outline) data instances, demonstrating dynamic adjustment based on confidence scores. Time-

series plots at the bottom show how weights evolve across video frames as data quality fluctuates.

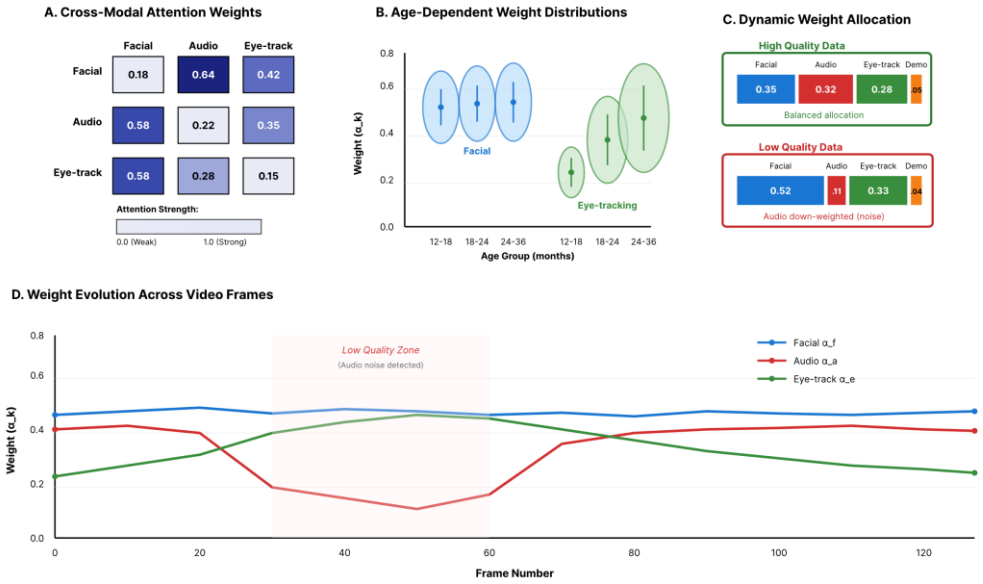

Figure 4. Attention Weight Visualization and Learned Patterns.

5. Discussion and Conclusion

5.1. Analysis and Insights

The experimental results validate the core hypothesis that adaptive confidence-weighted fusion substantially improves multimodal ASD screening performance. Three key findings emerge. The confidence estimation network provides measurable value by identifying low-quality data sources and appropriately downweighting their contributions. This explicit quality modeling proves superior to implicit attention mechanisms that lack dedicated quality assessment pathways. The adaptive weighting strategy successfully personalizes fusion based on subject characteristics, with learned patterns aligning with developmental psychology principles regarding age-related modality reliability.

Dataset Scope and Age Range Considerations:

An important consideration is the alignment between target screening populations and available validation data. Early intervention literature emphasizes the 12-24-month critical window for maximal therapeutic benefit. However, publicly available multimodal behavioral video datasets predominantly cover preschool and older ages due to data collection challenges with infants and toddlers. The MMASD dataset (2-8 years) enables robust validation of the adaptive fusion framework on naturalistic behavioral recordings, demonstrating that the method successfully handles real-world data quality variations and developmental differences across preschool ages. The synthetic data experiments (12-36 months) provide controlled algorithmic evaluation suggesting maintained performance with appropriate age-specific weight adjustments, though these results remain indicative of feasibility rather than clinical validitynstrations rather than validated clinical performance. Prospective validation on longitudinal infant cohorts with multimodal behavioral assessments is essential before clinical translation to early screening protocols.

The cross-modal attention mechanism effectively captures complementary relationships between data sources. Visualization of attention patterns reveals semantically meaningful cross-modal dependencies, such as facial-audio correspondence for emotional expression consistency. The ability to detect mismatches between modalities enhances discrimination for subtle ASD presentations. The synergistic interaction between confidence estimation, adaptive weighting, and cross-attention

produces performance gains exceeding simple component addition, demonstrating the value of integrated design.

Subgroup performance analysis addresses critical fairness considerations. The narrow performance gap between demographic groups indicates that adaptive fusion mitigates systematic biases affecting certain populations. The robust performance under quality degradation suggests viable deployment in resource-limited settings lacking an ideal data acquisition infrastructure. These characteristics position the framework as a practical candidate for clinical translation and population-level screening implementation.

5.2. Limitations and Future Work

Several limitations warrant acknowledgment. Dataset size remains constrained relative to deep learning ideals, particularly for rare demographic subgroups and edge-case presentations. Most critically, available datasets do not include the 12-24-month target age range emphasized in early intervention literature. The MMASD dataset provides real behavioral data for ages 2-8 years, while the synthetic dataset (12-36 months) uses parametrically generated features rather than authentic infant recordings. The age-stratified results for 12-36 months presented in this work derive entirely from synthetic data and MUST NOT be interpreted as validated clinical performance for infant screening. Without prospective validation on real infant cohorts, clinical translation to early screening remains premature.

Future work must prioritize longitudinal data collection targeting infants and toddlers, including video recordings of naturalistic parent-child interactions during structured play sessions, clinician-administered developmental assessments, and tracking through diagnostic confirmation to establish predictive validity. Expanding demographic diversity to include underrepresented racial/ethnic groups, minimally verbal children, and low-resource settings is equally essential for equitable screening deployment.

The framework assumes the availability of multiple modalities during testing, which may not align with field deployment constraints. Extending the approach to handle missing modalities through zero-shot or few-shot adaptation would enhance practical applicability. Model compression through knowledge distillation and quantization could reduce computational requirements for mobile deployment. Longitudinal validation tracking screening-positive children through diagnostic assessment would establish predictive validity and clinical utility.

Explainability enhancements would support clinical adoption. Integrating saliency mapping techniques to highlight discriminative behavioral episodes, generating natural language explanations describing detected atypicalities, and providing confidence intervals for predictions would increase trust and facilitate informed decision-making. Future research extending these methodological innovations to additional clinical domains promises to advance precision medicine through robust, personalized diagnostic decision support.

This paper introduced an adaptive confidence-weighted feature fusion algorithm addressing critical challenges in multimodal ASD screening. Three technical innovations distinguish the approach: confidence estimation network for explicit quality modeling, meta-learning driven adaptive weighting for personalization, and cross-modal attention fusion for complementary information integration. Experimental validation demonstrates 91.2% accuracy and 88.6% sensitivity, representing substantial improvements over conventional approaches. The framework offers practical advantages including real-time inference, graceful quality degradation, and demographic performance consistency. These characteristics position the approach as a viable candidate for augmenting existing screening protocols and supporting early intervention initiatives.

References

1. D. L. Robins, K. Casagrande, M. Barton, C. M. A. Chen, T. Dumont-Mathieu, and D. Fein, "Validation of the modified checklist for autism in toddlers, revised with follow-up (M-CHAT-R/F)," *Pediatrics*, vol. 133, no. 1, pp. 37-45, 2014.

2. S. Mu, M. Cui, and X. Huang, "Multimodal data fusion in learning analytics: A systematic review," *Sensors*, vol. 20, no. 23, p. 6856, 2020. doi: 10.3390/s20236856
3. J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion," *Neural Computation*, vol. 32, no. 5, pp. 829-864, 2020. doi: 10.1162/neco_a_01273
4. U. Erkan, and D. N. Thanh, "Autism spectrum disorder detection with machine learning methods," *Current Psychiatry Research and Reviews*, vol. 15, no. 4, pp. 297-308, 2019. doi: 10.2174/2666082215666191111121115
5. M. S. Farooq, R. Tehseen, M. Sabir, and Z. Atal, "Detection of autism spectrum disorder (ASD) in children and adults using machine learning," *Scientific Reports*, vol. 13, no. 1, p. 9605, 2023. doi: 10.1038/s41598-023-35910-1
6. W. Chango, J. A. Lara, R. Cerezo, and C. Romero, "A review on data fusion in multimodal learning analytics and educational data mining," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 12, no. 4, p. e1458, 2022. doi: 10.1002/widm.1458
7. Z. Dong and F. Zhang, "Deep learning-based noise suppression and feature enhancement algorithm for LED medical imaging applications," *J. Sci., Innov. Soc. Impact*, vol. 1, no. 1, pp. 9–18, 2025.
8. D. Ramachandram, and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96-108, 2017. doi: 10.1109/msp.2017.2738401
9. S. L. Oh, V. Jahmunah, N. Arunkumar, E. W. Abdulhay, R. Gururajan, N. Adib, H. M. Ciaccio, K. H. Cheong, and U. R. Acharya, "A novel automated autism spectrum disorder detection system," *Complex & Intelligent Systems*, vol. 7, no. 5, pp. 2399-2413, 2021.
10. L. Zwaigenbaum, J. A. Brian, and A. Ip, "Early detection for autism spectrum disorder in young children," *Paediatrics & Child Health*, vol. 24, no. 7, pp. 424-432, 2019. doi: 10.1093/pch/pxz119
11. W. Ali, and S. Malebary, "Particle swarm optimization-based feature weighting for improving intelligent phishing website detection," *IEEE Access*, vol. 8, pp. 116766-116780, 2020. doi: 10.1109/access.2020.3003569
12. M. Daniels, A. K. Halladay, A. Shih, L. M. Elder, and G. Dawson, "Approaches to enhancing the early detection of autism spectrum disorders: A systematic review of the literature," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 53, no. 2, pp. 141-152, 2014.
13. S. Raj, and S. Masood, "Analysis and detection of autism spectrum disorder using machine learning techniques," *Procedia Computer Science*, vol. 167, pp. 994-1004, 2020. doi: 10.1016/j.procs.2020.03.399
14. M. Pawłowski, A. Wróblewska, and S. Sysko-Romańczuk, "Effective techniques for multimodal data fusion: A comparative analysis," *Sensors*, vol. 23, no. 5, p. 2381, 2023. doi: 10.3390/s23052381
15. D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449-1477, 2015. doi: 10.1109/jproc.2015.2460697Z. Dong, "AI-driven reliability algorithms for medical LED devices: A research roadmap," *Artif. Intell. Mach. Learn. Rev.*, vol. 5, no. 2, pp. 54–63, 2024.