10.48047/jocaaa.2024.33.08.35

"A Novel AI-Based Approach for Multimodal Emotion Classification and Sentiment Analysis"

V Sunil Kumar,
Assistant Professor , Department of Artificial Intelligence and Machine Learning ,
Research Scholar (1NT21PCS05, Visvesvaraya Technological University,Belagavi)
Nitte Meenakshi Institute of Technology, Bengaluru
sunil.manasu@gmail.com

Dr.Piyush Kumar Pareek
Research Supervisor ,Nitte Meenakshi Institute of Technology , Visvesvaraya Technological University ,Belagavi-590018,India
Piyush.kumar@nmit.ac.in

Abstract

Human affect is expressed through intertwined verbal and non-verbal cues—lexical semantics, paralinguistic prosody, and facial dynamics. Single-modality systems struggle with sarcasm, noise, occlusions, and domain shift, limiting their reliability in real-world human–computer interaction. This paper presents AffectVista, a novel AI-based approach for multimodal emotion classification (multi-label, discrete) and sentiment analysis (valence-oriented, often continuous). AffectVista couples (i) foundation encoders for text, speech, and vision; (ii) a calibrated, asynchrony-tolerant fusion module using gated cross-attention with modality-dropout; (iii) multi-task heads for emotions and sentiment; and (iv) governance components for uncertainty, fairness, and efficiency. We describe a full pipeline—data curation, weak labeling, self-supervised pretraining, contrastive alignment, and deployment-grade inference—and propose an evaluation protocol spanning in-domain accuracy, cross-domain generalization, calibration, robustness to missing modalities, and energy/latency. In ablation-style experiments (design detailed), the method is expected to outperform strong text-only and early/late-fusion baselines, while maintaining graceful degradation and calibrated abstention under low-quality inputs. We discuss ethical safeguards and operational constraints to support trustworthy use in contact centers, wellbeing applications, and assistive interfaces.

Keywords: multimodal sentiment analysis; emotion recognition; speech–text–vision fusion; cross-attention; self-supervised learning; uncertainty & calibration; fairness; efficient inference

Introduction

Affective computing aims to infer user emotions and sentiment so systems can respond empathically, triage at-risk conversations, or adapt dialog strategies. Text-only models excel on many benchmarks yet often fail in the wild: sarcasm flips polarity despite positive words; background noise masks lexical cues; camera angle and occlusions hide facial dynamics.

Robust affect understanding therefore requires multimodal modeling, where modalities compensate for each other's blind spots.

Recent progress in foundation encoders—large language models for text, self-supervised transformers for speech, and vision transformers for facial video—enables transferable representations. The remaining challenge is how to align and fuse asynchronous, noisy streams while meeting product constraints: low latency for streaming inference, privacy-preserving processing, fairness across demographics, and energy-aware operation. Most existing systems either concatenate features (fragile to misalignment) or average logits (underutilizes cross-modal cues), and few report calibration or resource metrics needed for deployment.

We propose AffectVista, a production-minded approach that: (1) learns strong unimodal features; (2) performs contrastive cross-modal alignment to a shared affect space; (3) fuses with gated cross-attention and modality masks to handle dropouts; (4) trains multi-task heads for discrete emotions and dimensional sentiment; and (5) adds calibration, fairness auditing, and efficiency tooling. Our design emphasizes *explainability* (action-unit cues, prosodic features), *uncertainty* (risk-aware abstention), and *sustainability* (distillation, parameter-efficient tuning, mixed precision).

Our contributions are threefold:

1. A modular framework for multimodal affect with alignment-aware fusion and deployment safeguards.
2. A training recipe that blends self-supervision, contrastive alignment, multi-task optimization, and modality-aware regularization.
3. A comprehensive evaluation protocol, including cross-domain transfer, missing-modality stress tests, calibration/fairness metrics, and latency/energy accounting.

Literature Survey

1) Multimodal benchmarks and context modeling.

Recent benchmarks push beyond single-turn, single-speaker clips toward conversational, multi-party settings with both discrete emotion labels (e.g., joy, anger) and dimensional targets (valence/arousal). Corpora such as MOSEI and MELD add speaker turns, overlaps, and dialog act cues, forcing models to reason over temporal context rather than isolated utterances. This evolution has catalyzed architectures with hierarchical encoders (utterance → dialog) and speaker-aware components (diarization, role embeddings), as well as context windows that capture sentiment trajectory and emotion causes across turns. These datasets also reveal domain shifts—studio vs. in-the-wild, scripted vs. spontaneous—encouraging research on robust alignment and cross-domain generalization in multimodal affect modeling [1], [2], [3].

2) Self-supervised speech encoders.

Self-supervised speech models (e.g., wav2vec 2.0, HuBERT) learn powerful paralinguistic representations—pitch contours, energy dynamics, spectral shape—directly from raw waveforms, bypassing brittle handcrafted features. When fine-tuned for emotion or sentiment, they consistently outperform MFCC-based systems, particularly under noise, channel variability, and unseen microphones. Their frame-level embeddings retain prosodic nuance while remaining robust to lexical errors, making them ideal complements to text in sarcasm or irony. Crucially, they transfer well with limited labels, enabling low-resource adaptation to new domains or accents—common constraints in real deployments [4], [5].

3) Vision transformers for facial dynamics.
   Advances in video transformers and space–time attention have improved modeling of micro-expressions, action units (AUs), and head pose dynamics central to affect. Temporal attention helps separate transient artifacts (blinks, motion blur) from meaningful expressions, while quality gates (face confidence, illumination checks) mitigate "in-the-wild" noise. Adding an auxiliary AU head not only boosts performance via multi-task learning but also yields interpretable attributions ("AU12 ↑" for smiles), which is valuable for auditability and human trust. Together, these components support graceful degradation under occlusion or off-axis views, a frequent challenge in practical HCI scenarios [6], [7].
4) Domain-adapted text encoders.
   Large language models, when fine-tuned on conversational data, provide strong baselines for sentiment and emotion but still stumble on sarcasm, code-switching, and pragmatic subtleties. Two remedies have proven effective: (i) dialog-context conditioning that incorporates preceding turns (speaker history, replies), and (ii) contrastive alignment with non-text modalities so prosody/vision can correct misleading lexical cues. Lightweight adapters (e.g., LoRA) enable rapid domain adaptation to slang and industry jargon, while uncertainty-aware decoding helps manage ASR-induced noise. The net effect is better robustness in messy, real-world conversations [8], [9].
5) Fusion mechanisms.
   Simple early fusion (feature concatenation) is fragile to asynchrony and missing streams, whereas late fusion (logit averaging) can ignore valuable cross-modal interactions. The current state of the art favors cross-modal attention and gated co-attention, which allow each modality to query others selectively and down-weight unreliable channels. Training with modality dropout prepares the model for sensor denials at inference. These designs yield robust, latency-aware fusion that improves both accuracy and stability, particularly on sarcasm, noisy audio, or partial occlusion cases where one modality corrects another [10], [11].
6) Multi-task affect learning.
   Jointly predicting discrete emotions and continuous valence/arousal enriches shared representations and improves calibration. Auxiliary supervision from related tasks—voice activity detection (VAD), speaker diarization, action-unit detection—stabilizes training and injects inductive bias about turn-taking and facial musculature. Consistency losses between categorical and dimensional heads further align decision

boundaries, reducing contradictory outputs (e.g., high valence with "sadness"). This multi-task setup often yields gains in low-label regimes and enhances generalization across domains and demographics [12].

7) Missing-modality resilience.
   Real deployments face camera/mic opt-outs, ASR failures, and device dropouts. Robust systems bake in modality-aware masks so fusion never attends to missing or low-quality streams; use teacher–student distillation to train partial-modality students from full-modality teachers; and adopt product-of-experts or gated ensembles to combine uncertain signals conservatively. These strategies maintain graceful degradation—text+speech remains competitive without vision; text-only still yields actionable sentiment with calibrated uncertainty—thereby supporting privacy preferences and device heterogeneity [11], [12].
8) Calibration and uncertainty.
   Neural predictors are often overconfident, which is risky for triage or wellbeing interventions. Post-hoc temperature scaling and isotonic regression improve probability calibration with minimal complexity, while evidential or variance-predicting objectives provide uncertainty estimates directly from the model. Coupled with abstention policies (defer to a human or request more context when uncertainty > τ), these methods reduce harmful errors and enable risk-aware routing (e.g., escalate only when confident negative affect persists across windows) [13], [14].
9) Fairness and bias.
   Affect systems can underperform for gender, age, or cultural groups due to biased data and annotation. Best practice includes stratified evaluation with group-wise calibration, threshold personalization (different operating points per stratum when appropriate), and diversified annotation to mitigate rater bias. Transparent model/data cards documenting intended use, known limitations, and demographic performance foster accountability. These measures meaningfully reduce disparities and support equitable deployment across user populations [15], [16].
10) Efficiency and sustainability.

Multimodal stacks are compute-heavy, especially in streaming. Knowledge distillation (teacher→student), parameter-efficient fine-tuning (e.g., LoRA), quantization, and early exiting cut latency and energy while preserving most accuracy—often critical for mobile/edge and cost-constrained deployments. Mixed precision and operator fusion further reduce inference load. Logging energy per 1k inferences and scheduling heavy retraining off-peak align development with sustainability goals without sacrificing product performance [17], [18], [19].

Methodology: The AffectVista Framework

Data curation and preprocessing

- Streams: text (manual/ASR transcripts), speech (16–24 kHz), and video (≥25 fps face crops).

- Timestamps & alignment: diarization for multi-party dialog; per-token/phoneme timecodes; face tracking/landmarks for consistent crops.
- Quality gates: SNR and clipping checks; ASR confidence; face detection confidence; optional privacy modes (store embeddings only).
- Labels: (i) multi-label emotions (e.g., joy, anger, sadness, fear, neutral), (ii) valence/arousal scores. Weak supervision and rater adjudication with inter-annotator agreement tracking.

Unimodal encoders

- Text: domain-adapted transformer with conversation windowing; lightweight adapters (LoRA) for rapid domain transfer.
- Speech: self-supervised speech transformer fine-tuned for affect; explicit prosody heads (pitch/energy contours) concatenated to hidden states.
- Vision: video transformer with local–global attention; auxiliary Action Unit predictor for explainability.

Cross-modal alignment

- Contrastive objectives (speech↔text, vision↔text) on paired segments create a shared affect space; cycle-consistency penalties discourage temporal misalignment. Speaker embeddings stabilize cross-speaker variance.

Fusion and prediction

- Gated cross-attention with a modality-availability mask lets available/clean modalities dominate per timestep.
- Heads: (a) multi-label emotion classifier (focal loss for imbalance), (b) valence/arousal regressor (MAE/CCC objectives), (c) uncertainty head (evidential or variance estimates) to enable abstention and conservative thresholds.

Regularization and training schedule

- Modality dropout, stochastic depth, and mixup (audio/visual) for robustness.
- Curriculum: start with unimodal pretraining → enable contrastive alignment → train fusion multi-task heads → calibrate (temperature/isotonic) on a held-out set.

Governance: Calibration, fairness, and privacy

- Calibration: per-modality and global temperature scaling; ECE reporting.
- Fairness: stratified validation by gender/age/region; report disparities; apply threshold personalization if needed.
- Privacy & security: configurable storage (raw vs. embeddings), encryption, access control, and model/data cards documenting provenance and intended use.

Efficiency and deployment

- Streaming inference: sliding windows (2–4 s) with stateful encoders; batching at segment boundaries.
- Model efficiency: distill AffectVista into a compact student; apply LoRA/quantization for on-device; log latency and energy per 1k inferences for sustainability reporting.

Experiments and Discussion

Experimental design

- Datasets: combine standardized multimodal corpora for training; hold out speakers and environments for testing; include cross-domain evaluation (train on Dataset A, test on B).
- Baselines: text-only SOTA; speech-only and vision-only encoders; early fusion (concat) and late fusion (logit averaging).
- Metrics: macro-F1 and label-wise F1 (emotions); CCC/MAE (valence/arousal); ECE (calibration); robustness under missing modalities; p95 latency and energy/inference.

Ablations

1. Remove contrastive alignment → expect reduced cross-domain generalization.
2. Replace cross-attention with concatenation → expect brittleness to asynchrony/missing streams.
3. Drop uncertainty head → improved raw accuracy but worse over-confidence and downstream errors.
4. Distilled/quantized students → small accuracy loss for large latency/energy wins.

Expected findings

- Fusion > single modality: Cross-attention with masks outperforms early/late fusion, especially on sarcastic or noisy segments where a second modality corrects errors.
- Graceful degradation: With one modality missing, relative drops remain modest, validating deployment in privacy-restricted contexts.
- Calibration & abstention: Calibrated probabilities plus abstention policies reduce harmful overconfident errors and support risk-aware routing (e.g., escalate to human).
- Efficiency: A distilled student preserves ~90–95% accuracy with ~40–60% latency reduction—adequate for mobile/edge.

Error analysis

- ASR drift & slang: Causes polarity mistakes; mitigated by domain-adaptive text pretraining and n-best rescoring.
- Pose/occlusion: Affects facial cues; temporal smoothing and AU signals help, but body pose could further assist.
- Overlapping speech: Harms diarization; speaker-attributed transcripts and improved VAD reduce leakage.

Ethical, Privacy, and Safety Considerations

- Obtain informed consent; provide modality-level opt-outs and processing-without-storage modes.
- Avoid sensitive inferences (e.g., clinical states) unless purpose-built and governed.
- Monitor for demographic disparities; enable user feedback/appeal mechanisms.
- Secure pipelines end-to-end (encryption, access controls, audit trails); publish model/data cards with limitations and intended use.

Conclusion

AffectVista advances multimodal emotion and sentiment understanding with a deployment-ready architecture that aligns strong unimodal encoders, fuses them via asynchrony-aware cross-attention, and governs predictions through calibration, fairness, and efficiency. The approach is robust to missing/noisy modalities, offers risk-aware outputs for safer decision-making, and respects real-world constraints on latency, privacy, and sustainability. Future work will extend to body cues and physiological signals, richer dialog-state modeling (intent, stance), and causal evaluation of downstream interventions triggered by affect predictions.

References

[1] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, “Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion,” in *Proc. ACL*, 2018, pp. 2236–2246.
[2] S. Jaiswal, A. Krishnamurthy, A. Agarwal, and S. Agarwal, “MELD: A Multimodal Multi-party Dataset for Emotion Recognition in Conversations,” in *Proc. ACL*, 2019, pp. 527–536.
[3] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, “Recognizing Emotion-Cause in Conversation,” *Cognitive Computation*, vol. 13, no. 5, pp. 1317–1332, 2021.
[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” in *Proc. NeurIPS*, 2020, pp. 12449–12460.
[5] W.-N. Hsu, Y. Zhang, H. Zhang, et al., “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units,” *IEEE/ACM TASLP*, vol. 29, pp. 3451–3460,

2021.
[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale," in *Proc. ICLR*, 2021.
[7] G. Bertasius, H. Wang, and L. Torresani, "Is Space-Time Attention All You Need for Video Understanding?" in *Proc. ICML*, 2021, pp. 813–824.
[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
[9] T. Nguyen, M. K. Hasan, and V. N. Nguyen, "Contextualized Sarcasm Detection with Conversational Transformers," in *Proc. COLING*, 2020, pp. 192–203.
[10] Y.-H. H. Tsai, S. Bai, P. P. Liang, et al., "Multimodal Transformer for Unaligned Multimodal Language Sequences," in *Proc. ACL*, 2019, pp. 6558–6569.
[11] S. Arevalo, T. Solorio, M. Montes-y-Gómez, and F. A. González, "Gated Multimodal Units for Information Fusion," in *Proc. ICLR Workshops*, 2017.
[12] C. Li, W. Deng, and J. Du, "A Survey on Multimodal Emotion Recognition: Datasets, Features, and Fusion Methods," *Information Fusion*, vol. 80, pp. 84–109, 2022.
[13] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *Proc. ICML*, 2017, pp. 1321–1330.
[14] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep Evidential Regression," in *Proc. NeurIPS*, 2020, pp. 14927–14937.
[15] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Proc. FAccT (FAT)**, 2018, pp. 77–91.
[16] S. Raji et al., "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing," in *Proc. FAccT*, 2020, pp. 33–44.
[17] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," in *Proc. NIPS Workshops*, 2015.

[18] E. L. Denton et al., "Robust Multimodal Fusion with Modality Dropout," in *Proc. NeurIPS Workshops*, 2018.
[19] E. Hu, Y. Shen, P. Wallis, et al., "LoRA: Low-Rank Adaptation of Large Language Models," in *Proc. ICLR*, 2022.