Semi-Supervised Multimodal Emotion Recognition in Low Resource Egyptian Arabic via Fusion Strategy Comparison and Contrastive Learning

Anonymous ACL submission

Abstract

Multimodal Emotion Recognition (MER) in under-resourced languages presents a difficult challenge, primarily due to the scarcity of annotated data and the intricate task of effectively fusing diverse information streams. This paper introduces a novel semi-supervised framework tailored specifically for Egyptian Arabic (EA), a dialect often underrepresented in MER research. The proposed approach integrates textual, acoustic, and visual features, employing a combination of self-supervised pre-training and contrastive learning to reduce dependence on extensive labeled datasets. The study conducts a thorough investigation into various fusion strategies including early, late, and a proposed hybrid method to discern their optimal impact on MER performance. Through rigorous experimentation on the EGY-MER dataset, the findings compellingly demonstrate that this semi-supervised framework, particularly when enhanced with hybrid fusion and contrastive learning, substantially surpasses the performance of traditional unimodal and fully supervised baselines. This research not only significantly advances MER capabilities for low-resource EA but also offers profound insights into effective multimodal fusion and semi-supervised learning techniques that are broadly applicable to other linguistically diverse and under-resourced contexts.

1 Introduction

Emotion recognition remains a complex task even for humans, given that emotions are inherently subtle, variegated, and context-dependent (Wu et al., 2025). Most of the work in literature has traditionally focused on unimodal approaches analyzing either linguistic content, acoustic features, or visual cues in isolation(Kapase· and Uke, 2025). However, such unimodal systems are fundamentally constrained. For instance, textual sentiment analysis can fail to detect sarcasm or irony, while facial expressions may not align with speech, and physiological signals may fluctuate due to factors unrelated to emotion, such as fatigue or illness. Despite delivering useful results, unimodal systems do not reflect the true nature of the human emotional experience, which is naturally multimodal(Thakur et al., 2025). Multimodal Emotion Recognition (MER) addresses these limitations by combining information from multiple sensory streams—speech, text, vision, and physiological signals—to produce richer and more robust inferences. This integrated approach mirrors human perception, which relies on synchronizing facial cues, vocal intonation, body language, and contextual information to interpret emotional states. MER systems offer resilience to noise and ambiguity, outperforming unimodal systems, especially in real-world conditions where input modalities may be incomplete, contradictory, or asynchronous(Kapase· and Uke, 2025)(Bravo et al., 2025)(Thakur et al., 2025)(Wu et al., 2025).

In Arabic-speaking environments, the application of MER presents both promising opportunities and unique challenges. Chief among them is the diglossic nature of Arabic, where Modern Standard Arabic (MSA) coexists with a wide array of regional dialects that differ sharply in pronunciation, morphology, syntax, and vocabulary. Cultural variation further complicates matters, as emotional displays in Arabic-speaking societies are modulated by social norms, gender, and formality undermining the direct transfer of Western-centric emotion models. This underlines the need for culturally calibrated annotation schemes and models attuned to the emotional registers of Arabic speakers(Dahou et al., 2025). Despite this, the Arab world remains largely underrepresented in affective computing, and one of the key obstacles to progress is the lack of multimodal Arabic datasets; existing corpora fall short in scope and modality diversity, making them insuf-

ficient as training resources for robust MER systems(Haouhat et al., 2023)(Al-Azani and El-Alfy, 2025)(Aladeemy et al., 2024)(Abu Shaqra et al., 2023)(Al Roken and Barlas, 2023)(Alqarafi et al., 2017). While recent literature shows advances in Arabic ER and MER, most existing models target MSA or adopt a general multi-dialect approach. These challenges are further intensified when focusing on Egyptian Arabic (EA), the most widely understood and spoken dialect in the Arab world , which exhibits distinct phonology, flexible syntax, frequent code-switching, and heavy reliance on idiomatic expressions. From an emotional perspective, EA is particularly layered; intonation, sarcasm, and culturally specific vocal fillers play an outsized role in conveying affect, often overriding literal semantic meaning. This necessitates specialized architectures capable of handling EA's morphological richness while robustly fusing acoustic and visual cues to resolve the ambiguities inherent in informal spoken Egyptian communication — yet the optimal fusion of these heterogeneous modalities for low-resource dialectal speech remains a critically under-explored research area(Dahou et al., 2025)(Al Roken and Barlas, 2023).

This paper introduces a robust framework designed specifically for Egyptian Arabic MER. The current limitations of current systems are addressed by evaluating five distinct fusion strategies: early, late, hybrid, attention-based, and transformer-based fusion. The proposed approach leverages pre-trained encoders, including AraBERT for text, Wav2Vec2.0 for speech, and Swin Transformer for visual features, adapted via Low-Rank Adaptation (LoRA) to suit low-resource constraints. The proposed framework incorporates a semi-supervised contrastive pre-training stage, which aligns modality representations in a shared latent space before supervised fine-tuning. This approach is particularly effective for Egyptian Arabic, where labeled data is scarce. This work establishes a comprehensive baseline for Egyptian Arabic MER and provides practical insights into the design of multimodal systems for low-resource, culturally diverse contexts.

2 Related Work

MER research has been evolving through different paradigms in terms of architecture from basic signal integration as in traditional MER systems to reliability aware modeling, contextual conversational graphs, and large-scale semantic alignment. Early fusion known as feature level concatenation tries to capture primitive correlations however its performance is hindered by temporal synchronization issues between a high resolution modality and a low resolution modality not to mention the dilemma of the curse of dimensionality. Late fusion known as decision level fusion on the other hand aims for robust performance against noise through training independent classifiers for each modality. In the Arabic domain, Shaqra et al. (2019) utilized the naturalistic AVANEmo dataset to demonstrate that late fusion significantly outperforms unimodal systems. Nevertheless, Dixit and Satapathy (Dixit and Satapathy, 2024) showed that late fusion fails to capture the intricate cross modal correlations essential for resolving ambiguities in natural conversation. Poria et al. (2017) demonstrated that context-aware fusion across text, audio, and video substantially outperforms any single modality on benchmarks such as IEMOCAP and CMU-MOSI (Zadeh et al., 2016). To avoid the fusion trap; research shifted towards model-level integration. The focus here is on how modalities interact within deep hidden layers. Al Roken and Barlas (2023) integrated a deep learning model utilizing CNNs and ResNet showing that hybrid integration is superior in terms of capturing the high emotional intensity in Arabic culture. Despite these advances in Arabic context, a critical missing component in dialectal Arabic is the robustness that is shown by (Mittal et al., 2020). The M3ER framework, introduced a "modality check step" using Canonical Correlation Analysis (CCA) to weigh the reliability of each modality on a per-sample basis, effectively generating "proxy" features for noisy or missing inputs. (Shou et al., 2024) proposed AGF-IB, utilizing Information Bottleneck theory and adversarial alignment to eliminate modality heterogeneity. Attention-based and transformer-based fusion enable selective cross-modal querying (Tsai et al., 2019) and global cross-modal self-attention (Rahman et al., 2020) respectively and are considered the current golden standard in high resource languages. Nevertheless, these heavy models based on transformers face the parameter paradox. High performing models require around 15.9 million parameters. On limited size datasets, these models are prone to catastrophic overfitting. In terms of literature corpus in Egyptian Arabic shows a crit-

ical restriction as most research rely on unimodal speech emotion detection. Abdel-Hamid (2020) and Safwat et al. (2023) both established foundational Egyptian speech datasets but these advances remain inherently unimodal lacking the visual and lexical modalities. A major recent advancement is the EGY-MER dataset (Hussein et al., 2025) that integrates synchronized visual, textual and audio modalities establishing baseline MER research purely focused on Egyptian Arabic.

2.1 Self-Supervised and Contrastive Learning

Self-supervised pre-training is considered the standard answer for low resource settings. SimCLR (Chen et al., 2020) established InfoNCE contrastive loss as a powerful representation learning signal, while Supervised Contrastive learning (Khosla et al., 2020) extended it to incorporate label structure. In speech, WavLM (Chen et al., 2022) applies masked prediction to raw waveforms with learned layer-wise aggregation. For text, MARBERT (Abdul-Mageed et al., 2021) is pre-trained on one billion dialectal Arabic tweets capturing Egyptian Arabic morphology and colloquial patterns that MSA trained models miss. For vision, combining CLIP (Radford et al., 2021), DINOv2 (Oquab et al., 2024), and ResNet-50 features provides complementary representations of facial expressions. This reveals a profound gap in truly multimodal frameworks for the Egyptian dialect, as most existing benchmarks remain either unimodal (SER) or lack textual alignment. This paper addresses these limitations by introducing a framework specifically tailored for Egyptian Arabic.

3 Methodology

This study proposes two experiments that share the same dataset and seven-class evaluation protocol but differ in data scale, encoder choice, and training strategy. Experiment 1 asks: *which fusion architecture is most effective?* Experiment 2 asks: *given more data and stronger encoders, how much does contrastive pre-training and ensembling help?*

3.1 Dataset Description

Both experiments utilize the EGY-MER dataset which is considered the first Egyptian multimodal dataset integrating synchronized audio, video and text(Hussein et al., 2025). The dataset consists

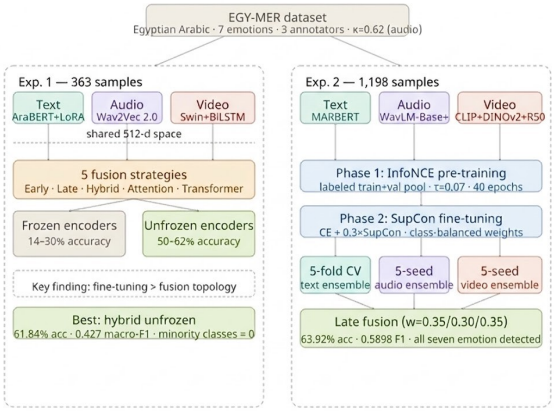

Figure 1: Proposed Methodology

of 412 processed videos out of these, 108 videos have been annotated. Initial corpus consisted of 363 labeled utterances while the extended corpus currently consists of 1998 labeled utterance. The dataset is collected from multiple Egyptian talk shows and dramas capturing the daily reactions of Egyptians. The dataset is labeled in basic six emotions (Happiness, Sadness, Anger, Fear, disgust and Surprise) with addition to Neutral emotion. EGY-MER is one of the few datasets in the Arabic context that covers rare emotions such as disgust. The dataset has an inter-annotator agreement of (Cohen's $\kappa \approx 0.62$) for Audio modality, (Cohen's $\kappa \approx 0.61$) for Video modality and (Cohen's $\kappa \approx 0.45$) for text which is in line with the fact that interpreting emotions from words and expressions is inherently subjective. Figure 1 illustrates the two experiments

3.2 Experiment 1: Fusion Architecture Comparison

3.2.1 Encoder Stack

All encoders project to a shared $d = 512$ embedding space.

Text (AraBERT). Transcripts are normalised (diacritics removed, alef forms unified) and tokenized with AraBERT's wordpiece tokenizer (Antoun et al., 2020). The 12-layer encoder (768-d, 12 heads) is used to produce a mean pooled representation projected to 512-d. LoRA (Hu et al., 2022) (rank 16, query and value projections) is utilized to reduce trainable parameters from ≈135M to ≈2M.

Audio (Wav2Vec 2.0). Raw 16 kHz waveforms are processed end-to-end by Wav2Vec 2.0 (Baevski et al., 2020). A CNN feature extrac-

tor yields ≈500 frame-level embeddings, mean-pooled with attention masking and projected to 512-d.

Video (Swin Transformer + BiLSTM). Sixteen frames at 224×224 are each encoded by Swin Transformer (base, ImageNet) to 1024-d (Liu et al., 2021). A two-layer bidirectional LSTM (256 units per direction) aggregates the sequence temporally; mean pooling yields the 512-d representation.

3.2.2 Fusion Strategies

Early Fusion. Token-level features are concatenated: $X_{\text{fused}} = [X_t; X_a; X_v] \in R^{(n_t+n_a+n_v)\times d}$, processed by a 4-layer, 8-head transformer with attention pooling.

Late Fusion. A gating network combines per-modality logits: $g = \sigma(W_g[h_t; h_a; h_v])$, $y_{\text{final}} = \sum_m g_m \cdot y_m$.

Hybrid Fusion. Parallel early-path (2-layer joint transformer) and late-path (per-modality MLPs) branches balanced by a learned gate.

Attention-Based Fusion. Bidirectional cross-attention between each modality pair produces augmented representations concatenated for MLP classification.

Transformer-Based Fusion. All modality tokens plus a learnable `[CLS]` are processed by a 6-layer, 8-head transformer; `[CLS]` is classified.

3.2.3 Training

AdamW with encoder LR 1×10^{-5} and fusion LR 1×10^{-4}, 500-step linear warmup, cosine decay. Combined loss: $\mathcal{L} = \lambda_1\mathcal{L}_{\text{CE}} + \lambda_2\mathcal{L}_{\text{aux}} + \lambda_3\mathcal{L}_{\text{InfoNCE}}$, label smoothing 0.1, modality dropout $p = 0.2$. Early stopping on validation macro-F1 (patience 10). Two encoder regimes are compared: *frozen* and *unfrozen*.

3.3 Experiment 2: Production Pipeline

Experiment 2 fixes late fusion as the topology, motivated by its robustness to missing modalities and near competitive performance in Experiment 1.

3.3.1 Upgraded Encoders

Text (MARBERT). Features concatenate `[CLS]`, mean-pooled, and max-pooled hidden states (2304-d). Bottom 8 layers frozen; top 4 fine-tuned at 2×10^{-5}. Multi-dropout (five 0.3-dropout heads, averaged logits) reduces variance.

Audio (WavLM-Base+). Learned layer-wise weighting combines all 13 transformer layers: $h = \sum_{l=1}^{13} \text{softmax}(\mathbf{w})_l \cdot h_l$. Final embedding concatenates mean and std of the weighted sequence (1536-d). Progressive unfreezing: bottom 6 WavLM layers stay frozen; top 6 unfreeze at epoch 3 with LR 4×10^{-5}.

Video (Multi-Feature MSW Transformer). Per-frame features from CLIP ViT-L/14 (768-d), DINOv2 (1024-d), and ResNet-50 (2048-d) are concatenated (3584-d) and processed by a Multi-Scale Window transformer with SE-block recalibration across three temporal scales (16, 8, and 4 central frames).

3.3.2 Phase 1: Self-Supervised Pre-training

Each encoder is independently pre-trained on the labeled train and validation pool using InfoNCE contrastive loss (van den Oord et al., 2018) ($\tau = 0.07$, 40 epochs):

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\mathbf{z}_1^\top \mathbf{z}_2/\tau)}{\sum_{k=1}^{N} \exp(\mathbf{z}_1^\top \mathbf{z}_k/\tau)}. \quad (1)$$

Two augmented views per sample form positive pairs. Audio: additive noise, time masking, speed perturbation. Video: random frame masking and feature-level Gaussian noise. Text: independent multi-dropout forward passes through MARBERT. The projection head is discarded after pre-training; encoder weights are retained as initialization for Phase 2.

3.3.3 Phase 2: Supervised Contrastive Fine-tuning

Each modality classifier is fine-tuned with:

$$\mathcal{L} = \mathcal{L}_{\text{CE}}(\text{class weights}) + 0.3 \cdot \mathcal{L}_{\text{SupCon}}, \quad (2)$$

where SupCon (Khosla et al., 2020) pulls together all same emotion samples in the projection space, providing denser supervision under label scarcity. Class-balanced weights address corpus imbalance.

3.3.4 Ensembling and Late Fusion

Video: five-seed ensemble (seeds 42, 1337, 2024, 777, 999), Lookahead optimizer (Zhang et al., 2019) + OneCycleLR, weights proportional to per-seed validation macro-F1.

Text: five-fold stratified CV on pooled train+val; test probabilities averaged across folds.

Late fusion weights grid-searched (step 0.05) over validation macro-F1: $p_{\text{final}} = w_v p_v + w_a p_a + w_t p_t$, $\sum w = 1$.

4 Results

4.1 Experiment 1

As the results show in Table 1 that unfreezing encoders improves accuracy significantly for early, late and hybrid fusion respectively. This indicates that pre-trained representations from MSA or English data are poorly aligned with Egyptian Arabic dialect emotion data. All configurations fail completely on Happiness and Surprise (0.000 F1) and most fail on Fear, revealing that 363 labeled samples are insufficient for minority class learning regardless of fusion strategy.

Configuration	Acc. (%)	Macro F1	Wtd. F1
Early (frozen)	30.26	0.278	0.341
Early (unfrozen)	55.26	0.338	0.538
Late (frozen)	19.74	0.207	0.210
Late (unfrozen)	50.00	0.363	0.509
Hybrid (frozen)	14.47	0.157	0.193
Hybrid (unfrozen)	**61.84**	**0.412**	**0.605**
Attention (unfrozen)	56.58	0.355	0.560
Transformer (unfrozen)	56.58	0.343	0.546

Table 1: Experiment 1 test results (76 test samples from 363-sample split).

4.2 Experiment 2: Unimodal Baselines

As indicated by Table 2, video and audio achieve identical accuracy (57.73%) but video produces higher macro-F1 (0.5264 vs. 0.4729), indicating more balanced per-class recognition. Text lags at 45.36%, consistent with lower annotator agreement on dialect transcripts.

Modality	Accuracy	Macro F1
Video	0.5773	0.5264
Audio	0.5773	0.4729
Text	0.4536	0.3777

Table 2: Experiment 2 unimodal test performance (180 test samples from 1,198-sample split).

4.3 Experiment 2: Ablation

From Table 3, four core findings can be drawn:

SSL alone does not improve macro-F1. Without a discriminative fine-tuning objective, contrastive pre-training structures the embedding space but does not sharpen class boundaries (0.4925 vs. baseline 0.5094).

SupCon alone improves F1 without changing accuracy. SupCon raises macro-F1 from 0.5094 to 0.5270 at identical accuracy, indicating it specifically benefits minority classes by pulling same-emotion embeddings together.

SSL + SupCon gives the largest accuracy gain. The combined system achieves 65.98% accuracy (+4.1 points over baseline), confirming that pre-training and fine-tuning objectives are complementary. The fusion weights shift heavily toward audio (0.60), suggesting WavLM representations benefit most from SSL initialization.

Ensembling achieves the best macro-F1. The final ensemble (63.92%, 0.5898 macro-F1) trades 4 points of accuracy against SSL+SupCon for substantially better minority-class coverage. What is worth mentioning is that this proposed approach manages to detect all 7 emotions without failing for any minority class which is a first on the EGYMER dataset.

Condition	Acc.	Macro F1	Weights (V,A,T)
Baseline	0.6186	0.5094	0.20, 0.30, 0.50
SupCon only	0.6186	0.5270	0.20, 0.30, 0.50
SSL only	0.6186	0.4925	0.20, 0.40, 0.40
SSL + SupCon	0.6598	0.5134	0.30, 0.60, 0.10
Final ensemble	**0.6392**	**0.5898**	0.35, 0.30, 0.35

Table 3: Experiment 2 ablation on test set (180 samples). Final ensemble row uses multi-seed video, 5-fold text, and grid-searched fusion weights without SSL/SupCon.

5 Discussion

Comparing the two experiments. Experiment 1 and Experiment 2 are not directly comparable because they differ in both data scale (363 vs. 1,198 samples) and methodology. The improvement in results reflects a combination of more data, stronger encoders, and improved training. The most important qualitative difference is that Experiment 2 achieves non-zero F1 on all seven classes, which can be attributed primarily to the larger labeled pool providing sufficient minority class examples for the class-balanced loss to operate on.

Encoder fine-tuning dominates architecture choice. The +25–47% accuracy gains from unfreezing in Experiment 1 show that for Egyptian Arabic MER, encoder adaptation matters more than fusion topology. With limited computational

power, it should be prioritized full encoder fine-tuning over architectural complexity.

Why text underperforms. The lagging in text compared to video and audio across both experiments can be directly lined to the lower annotator agreement on text labels as it introduces more label noise. MARBERT's dialectal pre-training may still be insufficient for short emotional utterances, and grid searched fusion weights consistently assign low weight to text, confirming the model itself learns to down weight text predictions.

6 Conclusion

This study successfully addresses the pressing need for robust Multimodal Emotion Recognition in low-resource linguistic settings. This study proposes and evaluates a novel semi-supervised framework specifically tailored to Egyptian Arabic; a currently rarely represented dialect in MER research, effectively overcoming the inherent limitations imposed by scarce labeled data. The comprehensive exploration of fusion strategies highlights the superior efficacy of a hybrid approach, which adeptly combines textual, acoustic, and visual information. The integration of self-supervised pre-training and contrastive learning proved pivotal in extracting rich, meaningful representations from limited resources, thereby significantly enhancing the model's generalization capabilities. The experimental results, validated on the EGY-MER dataset, provide compelling evidence that the proposed methodology not only outperforms existing unimodal and fully supervised baselines but also establishes a viable and promising pathway for advancing MER in other under-resourced linguistic contexts. Future research directions will focus on benefiting from the unlabeled corpus that is currently larger in size aiming for more generalized better perfroming models tailored for Egyptian Arabic.

Limitations

The two experiments use different data scales, preventing clean attribution of gains to methodology versus data size. SSL pre-training uses only the labeled pool; incorporating the full unlabeled corpus is a direct extension that is currently being studied and implemented. Fear, Happiness, and Surprise remain underrepresented despite class balancing.

References

Lamiaa Abdel-Hamid. 2020. Egyptian arabic speech emotion recognition using prosodic, spectral and wavelet features. *Speech Communication*, 122:19–30.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARbert & MARbert: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 7088–7105.

Ftoon Abu Shaqra, Rehab Duwairi, and Mahmoud Al-Ayyoub. 2023. A multi-modal deep learning system for arabic emotion recognition. *International Journal of Speech Technology*, 26(1):123–139.

Sadam Al-Azani and El-Sayed M El-Alfy. 2025. A review and critical analysis of multimodal datasets for emotional ai. *Artificial Intelligence Review*, 58(10):334.

Noora Al Roken and Gerassimos Barlas. 2023. Multimodal arabic emotion recognition using deep learning. *Speech Communication*, 155:103005.

Amani A Aladeemy, Ali Alzahrani, Mohammad H Algarni, Saleh Nagi Alsubari, Theyazn HH Aldhyani, Sachin N Deshmukh, Osamah Ibrahim Khalaf, Wing-Keung Wong, and Sameer Aqburi. 2024. Advancements and challenges in arabic sentiment analysis: A decade of methodologies, applications, and resource development. *Heliyon*, 10(21).

Abdulrahman S Alqarafi, Ahsan Adeel, Mandar Gogate, Kia Dashitpour, Amir Hussain, and Tariq Durrani. 2017. Toward's arabic multi-modal sentiment analysis. In *International Conference in Communications, Signal Processing, and Systems*, pages 2378–2386. Springer.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, pages 9–15.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Luis Bravo, Ciro Rodriguez, Pedro Hidalgo, and Cesar Angulo. 2025. A systematic review on artificial intelligence-based multimodal dialogue systems capable of emotion recognition. *Multimodal Technologies and Interaction*, 9(3):28.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. In *IEEE*

Journal of Selected Topics in Signal Processing, volume 16, pages 1505–1518.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607.

Abdelghani Dahou, Abdelhalim Hafedh Dahou, Mohamed Amine Chéragui, Amin Abdedaiem, Mohammed A. A. Al-qaness, Mohamed Abd Elaziz, Ahmed A. Ewees, and Zhonglong Zheng. 2025. A survey on dialect arabic processing and analysis: Recent advances and future trends. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 24(8):84:1–84:45.

Chhavi Dixit and Shashank Mouli Satapathy. 2024. Deep cnn with late fusion for real time multimodal emotion recognition. *Expert Systems with Applications*, 240:122579.

Abdelhamid Haouhat, Slimane Bellaouar, Attia Nehar, and Hadda Cherroun. 2023. Towards arabic multimodal dataset for sentiment analysis. In *Fourth International Conference on Intelligent Data Science Technologies and Applications, IDSTA 2023, Kuwai, Kuwait, October 24-26, 2023*, pages 126–133. IEEE.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Amany Hussein, May Hussien, Ammar Hassona, Wolfgang Minker, Mohammed A-M Salem, and Nada Sharaf. 2025. Egy-mer: Establishing the first egyptian arabic multimodal emotion recognition dataset for affective computing.

Ajay Babasaheb Kapase· and Nilesh Uke. 2025. A comprehensive review in affective computing: an exploration of artificial intelligence in unimodal and multimodal emotion recognition systems. *International Journal of Speech Technology*, pages 1–23.

Prannay Khosla, Yonglong Tian, Phillip Guo, Kihyuk Sohn, Zhiqiang Song, Dilip Krishnan, Xinlei Chen, Armand Joulin, Diane Bouchacourt, and 1 others. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1359–1367.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and 1 others. 2024. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 873–883.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369.

Sarah Safwat, Mohammed A-M Salem, and Nada Sharaf. 2023. Building an egyptian-arabic speech corpus for emotion analysis using deep learning. In *Pacific Rim International Conference on Artificial Intelligence*, pages 320–332. Springer.

Ftoon Abu Shaqra, Rehab Duwairi, and Mahmoud Al-Ayyoub. 2019. The audio-visual arabic dataset for natural emotions. In *2019 7th International Conference on Future Internet of Things and Cloud (FiCloud)*, pages 324–329. IEEE.

Yuntao Shou, Tao Meng, Wei Ai, Fuchen Zhang, Nan Yin, and Keqin Li. 2024. Adversarial alignment and graph fusion via information bottleneck for multimodal emotion recognition in conversations. *Information Fusion*, 112:102590.

Priyanka Thakur, Nirmal Kaur, Naveen Aggarwal, and Sarbjeet Singh. 2025. A comprehensive review of unimodal and multimodal emotion detection: Datasets, approaches, and limitations. *Expert Systems*, 42(9):e70103.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding.

You Wu, Qingwei Mi, and Tianhan Gao. 2025. A comprehensive review of multimodal emotion recognition: Techniques, challenges, and future directions. *Biomimetics*, 10(7):418.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. volume 31, pages 82–88.

Michael R Zhang, James Lucas, Geoffrey Hinton, and Jimmy Ba. 2019. Lookahead optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems*, volume 32.