Intelligent Agents with Emotional Intelligence: Current Trends, Challenges, and Future Prospects

Anonymous authors
Paper under double-blind review

Abstract

The development of agents with emotional intelligence is becoming increasingly vital due to their significant role in human-computer interaction and the growing integration of computational systems across various sectors of society. Affective computing aims to design intelligent systems that can recognize, evoke, and express human emotions, thereby emulating human emotional intelligence. While previous reviews have focused on specific aspects of this field, there has been limited comprehensive research that encompasses emotion understanding, elicitation, and expression, along with the related challenges. This survey addresses this gap by providing a holistic overview of core components of artificial emotion intelligence into one cohesive map for researchers. It covers emotion understanding through multimodal data processing, as well as affective cognition, which includes cognitive appraisal, emotion mapping, and adaptive modulation in decision-making, learning, and reasoning. Additionally, it addresses the synthesis of emotional expression across text, speech, and facial modalities to enhance human-agent interaction. This paper identifies and analyzes the key challenges and issues encountered in the development of affective systems, covering state-of-the-art methodologies designed to address them. Finally, we highlight promising future directions, with particular emphasis on the potential of generative technologies to advance affective computing.

1 Introduction

Developing intelligent agents that possess human-level intelligence is a key goal in the field of Human-Computer Interaction (HCI) and general artificial intelligence Jeon (2017). A crucial aspect of achieving this goal is the incorporation of emotional intelligence, which is essential for human cognition and social interaction, into these intelligent agents. Emotional intelligence encompasses three interrelated capabilities. First, emotion understanding involves accurately detecting and interpreting affective signals; for example, recognizing when a user is feeling frustrated during an interaction by analyzing their tone of voice or facial expressions. Second, emotion elicitation and experiences refer to interpreting the causes, context, and implications of emotions for both the individual and the interaction. For instance, an agent can perform a cognitive appraisal of the surrounding environment and contextual factors to infer its internal emotional state, which subsequently guides appropriate decision-making in the given situation. Third, emotion expression encompasses the capacity to generate, modulate, and convey appropriate emotional responses in a socially meaningful way, such as responding with a reassuring message or a sympathetic tone when the user is upset. Affective Computing, coined by Rosalind Picard Picard (2000), emerged as a discipline dedicated to equipping machines with emotional intelligence, enabling them to recognize, interpret, and respond to human emotions. By embedding emotional intelligence into intelligent agents, affective computing facilitates more naturalistic, adaptive, and socially competent interactions, which in turn enhance user trust, engagement, and satisfaction Zall & Kangavari (2024). Such emotionally intelligent systems not only improve usability but also enable advanced functionalities, including personalized assistance, empathetic dialogue, and context-aware decision-making.

Figure 1 presents the conceptual framework for intelligent agents with emotional intelligence. Building upon prior research, three fundamental components are identified as the core elements underpinning the development of emotional intelligence, described as follows:

1. **Emotion Understanding:** This stage involves analyzing the affective features embedded in user input, enabling the agent to accurately detect and interpret the user's emotional state during interaction Afzal et al. (2024); Zhao et al. (2025).

2. **Affective Cognition:** In this phase, the agent assesses emotional events through cognitive reasoning processes to ensure accurate and context-sensitive interpretation. Subsequently, it performs emotional elicitation modeling to generate an internal affective state, which modulates higher-order cognitive functions such as learning, inference, and decision-making Liu et al. (2024); Raggioli et al. (2025). This internal affective regulation drives the agent's adaptive behavior, resulting in responses aligned with the user's emotional context.

3. **Emotional Expression Synthesis:** Finally, the agent externalizes its emotional states through multimodal communication channels, including text, speech, and visual expressions, where cross-modal synchronization is essential for fostering coherence, authenticity, and naturalness in emotional interactions Chen et al. (2023b); Abilbekov et al. (2024).

Emotion understanding in intelligent agents encounters substantial challenges in accurately detecting subtle or ambiguous affective signals, especially within diverse cultural contexts or in the presence of environmental noise Rahmani et al. (2023) Kamran et al. (2023). These challenges arise from practical limitations, including datasets with restricted size, noise, imbalance, and suboptimal quality Ye et al. (2024); Oh & Kim (2024). Further obstacles are presented by learning model deficiencies, such as low accuracy, limited interpretability, poor generalizability, and the lack of standardized evaluation metrics Umair et al. (2024) Cambria et al. (2024). Multimodal integration introduces additional complexity, particularly in the combination of heterogeneous data types and the management of missing modalities Kumar et al. (2024); Geetha et al. (2024); Alsaadawi et al. (2024); Khan et al. (2024). The inherent complexity of emotions, including their diversity, overlap, and the challenges associated with experimental design, further complicates emotion understanding Ye et al. (2024); Sharma et al. (2024). Moreover, contemporary large-scale models are affected by issues such as hallucinations, high annotation costs, and limited contextual comprehension Farquhar et al. (2024); Schuller et al. (2024). Recent advancements, such as data augmentation, synthetic data generation, transfer, and semi-supervised learning, model interpretability methods, enhanced evaluation standards, and multimodal fusion, present promising avenues for addressing these challenges. Affective cognition, a core component of emotional intelligence in intelligent agents, extends beyond emotion recognition to encompass reasoning about emotions and the generation of contextually appropriate responses through the integration of cognitive and affective theories of mind. Modeling affective cognition is inherently challenging due to the complexity of capturing contextual and causal factors, often necessitating sophisticated cognitive frameworks enriched with domain-specific knowledge.

This process involves identifying events and mental states that elicit emotions (emotion elicitation) and interpreting the resulting behavioral and cognitive outcomes (emotional experiences), as described in cognitive appraisal theories and data-driven approaches Zall & Kangavari (2024); Liu et al. (2024); Jokinen & Oulasvirta (2025). However, the development of affective cognition models faces several significant challenges, including limited and ambiguous datasets, difficulties in computing cognitive appraisal variables, and scalability constraints in computational cognitive models Gandhi et al. (2024); Somarathna et al. (2022); Bayro & Jeong (2025). Additional challenges arise in large language model (LLM)-based systems, such as contextual misinterpretation and limitations in emotional reasoning capabilities Tak & Gratch (2024); Khan et al. (2025); Raggioli et al. (2025). To address these issues, recent studies have proposed solutions including advanced cognitive architectures, virtual reality–based data collection, reinforcement learning integration, and explainable emotion-alignment frameworks. This study investigates these challenges and solutions to advance the development of emotionally intelligent agents capable of naturalistic and socially appropriate interactions.

Emotional text, speech, and face synthesis have become essential components of affective computing. They allow agents to generate responses that are contextually appropriate and emotionally resonant. In the realm

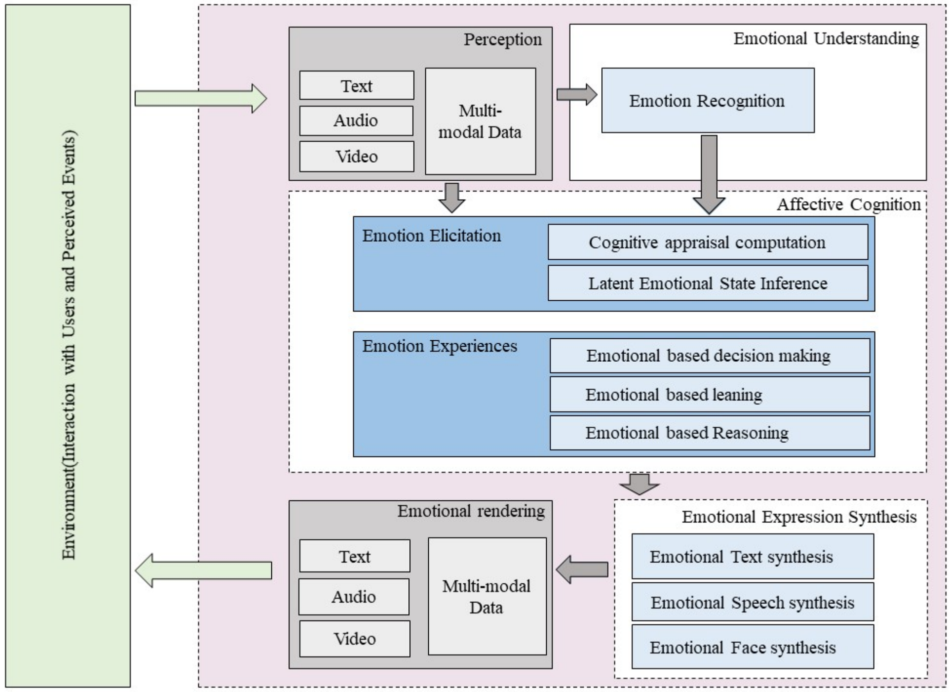

Figure 1: Overview of intelligent agent with emotional intelligence

of text synthesis, advancements like style transfer, conditional generation, and fine-tuned LLMs enhance expressiveness. However, challenges such as data sufficiency, consistency, and multimodal alignment still exist. Frameworks like MOPO and EmoBench aim to address these issues Firdaus et al. (2023); Resendiz & Klinger (2024); Sabour et al. (2024). Emotional Speech Synthesis (ESS) and voice conversion increasingly utilize non-parallel and diffusion-based methods, such as EmoConv-Diff, to improve scalability and emotional fidelity Zhou et al. (2020a); Prabhu et al. (2024). Simultaneously, emotional face synthesis utilizes GANs, diffusion models, and multimodal systems such as UniPortrait and EmotiveTalk to address challenges related to data imbalance and synchronization. These advancements enable the creation of lifelike and coherent multimodal affective expressions.

1.1 Research questions

To guide our investigation into the integration of emotional intelligence in intelligent agents, we pose the following research questions:

- *RQ1*: How can intelligent agents robustly detect and interpret affective signals across diverse environments despite data and model limitations?
- *RQ2*: What are effective approaches for modeling emotion elicitation and experience in affective cognition, given challenges in ambiguity, computation, and scalability?
- *RQ3*: How can intelligent agents generate contextually appropriate and emotionally coherent responses across multiple modalities?

1.2 Contribution

This study investigates the fundamental components of emotional intelligence in intelligent agents, with an emphasis on their functional roles, the specific challenges that impede effective implementation, and the solutions proposed in existing literature. Through a systematic examination of these dimensions, the study aims to elucidate how emotional intelligence can be effectively integrated into artificial agents to enable more natural, adaptive, and human-like interactions. By synthesizing current research findings, this work highlights both the progress achieved and the critical gaps that remain unaddressed in the pursuit of genuine emotional intelligence in intelligent agents. Despite the substantial potential of emotionally intelligent agents, persistent challenges across core capabilities continue to constrain their performance and applicability in real-world environments.

Table 1 provides a comprehensive comparison between this survey and recent studies in the field of affective computing published from 2024 to 2025. The comparison is organized along two principal dimensions: (1) emotional aspects, encompassing emotion recognition, elicitation, and expression, and (2) modalities, including text, speech, and visual or facial cues. This structure underscores the distinctive contribution of the present study, which offers an integrative perspective by addressing all relevant emotional dimensions across multiple modalities—thereby delivering a more holistic and unified understanding compared to previous research. The primary contributions of this study can be summarized as follows:

Table 1: Comparison of the present survey with recent studies in affective computing (2024–2025) based on covered emotional aspects and modalities

	Emotion			**Modality**		
	Recognition	Elicitation	Expression	Text	Speech	Vision/Facial
Recent Trends of Multimodal Affective Computing: A Survey from an NLP Perspective Hu et al. (2024)	✓	–	–	✓	✓	✓
A Review of Human Emotion Synthesis Based on Generative Technology Ma et al. (2025)	✓	–	✓	✓	✓	–
Affective Computing in the Era of Large Language Models: A Survey from the NLP Perspective Zhang et al. (2024e)	✓	–	✓	✓	–	–
Artificial Emotion: A Survey of Theories and Debates on Realising Emotion in Artificial Intelligence Li et al. (2025b)	–	✓	–	–	–	–
Emotion recognition and generation: a comprehensive review of face, speech, and text modalities Mobbs et al. (2025)	✓	–	✓	✓	✓	✓
Intelligent Agents with Emotional Intelligence: Current Trends, Challenges, and Future Prospects	✓	✓	✓	✓	✓	✓

- To the best of our knowledge, this work represents the first comprehensive review of intelligent agents equipped with the full spectrum of emotional intelligence capabilities—namely, emotion understanding, affective cognition, and emotional expression, providing a cohesive framework for advancing naturalistic and empathetic human-computer interaction (HCI).
- It offers an in-depth analysis and categorization of the key challenges that hinder the effective realization of these three core capabilities within intelligent agents.
- It systematically evaluates recent methodological advancements proposed to address these challenges and delineates promising future research directions for the development of emotionally intelligent systems.

1.3 Paper organization

The remainder of this paper is organized as follows. Section 2 describes the systematic literature review methodology. Section 3 examines challenges and emerging solutions in emotion understanding. Section 4 explores challenges and emerging solutions in affective cognition, focusing on emotion elicitation in interactive systems and emotional experiences. Sections 5 to 7 address challenges and emerging solutions in emotional expression synthesis across multiple modalities: Section 5 focuses on Emotional Text Synthesis (ETS), Section 6 on ESS, and Section 7 on emotional face synthesis. Section 8 offers a comprehensive discussion of challenges and future research directions. Finally, Section 9 concludes with a summary of contributions and a vision for developing emotionally intelligent agents.

2 Methodology: Systematic Literature Review Process

This paper adheres to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to ensure transparency and reproducibility in the study selection process. We adopted the PRISMA framework to systematically identify, screen, and include relevant literature on emotional intelligence in intelligent agents. Our focus is on three core capabilities: emotion understanding (emotion recognition), affective cognition (emotional elicitation and experiences), and emotional expression synthesis.

2.1 Identification

We conducted a comprehensive literature search across several electronic databases, including Google Scholar, IEEE Xplore, ACM Digital Library, Scopus, Web of Science, and arXiv. The search terms were crafted to encompass key themes and included combinations such as: ("affective computing" OR "emotional AI" OR "emotion recognition" OR "emotion synthesis") AND ("textual emotion" OR "speech emotion" OR "facial expression" OR "multimodal emotion") AND ("cognitive architectures" OR "computational models of emotion" OR "appraisal theories" OR "reinforcement learning"). Our search focused on publications from 2017 to 2025 to capture the latest advancements in deep learning-based approaches. Initially, no date restrictions were applied to include foundational works; however, we prioritized publications from 1990 to 2025 to emphasize modern developments. We also identified additional records through backward citation searching (reviewing the references of key papers) and forward citation searching (using tools like Google Scholar's "Cited by" feature). Grey literature, such as preprints and conference proceedings, was included if it was peer-reviewed or highly cited.

2.2 Screening

The titles and abstracts of all 2,500 records were screened for relevance. Studies were excluded at this stage if they were clearly outside the scope of affective computing (e.g., psychological studies on human emotion, medical studies on emotional disorders) or focused solely on human subjects without application to intelligent agents. This screening process reduced the number of records to 500.

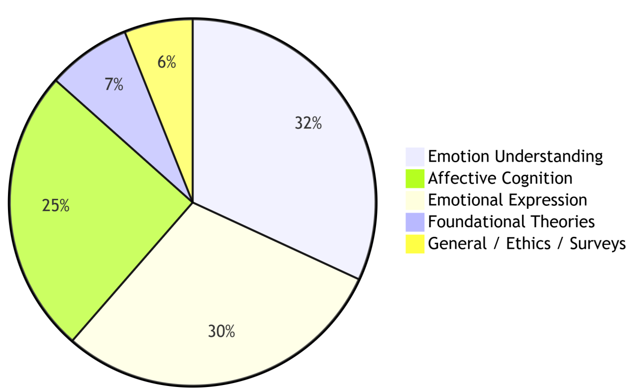

Figure 2: Distribution of the 298 included studies across the three core capabilities of emotional intelligence in intelligent agents.

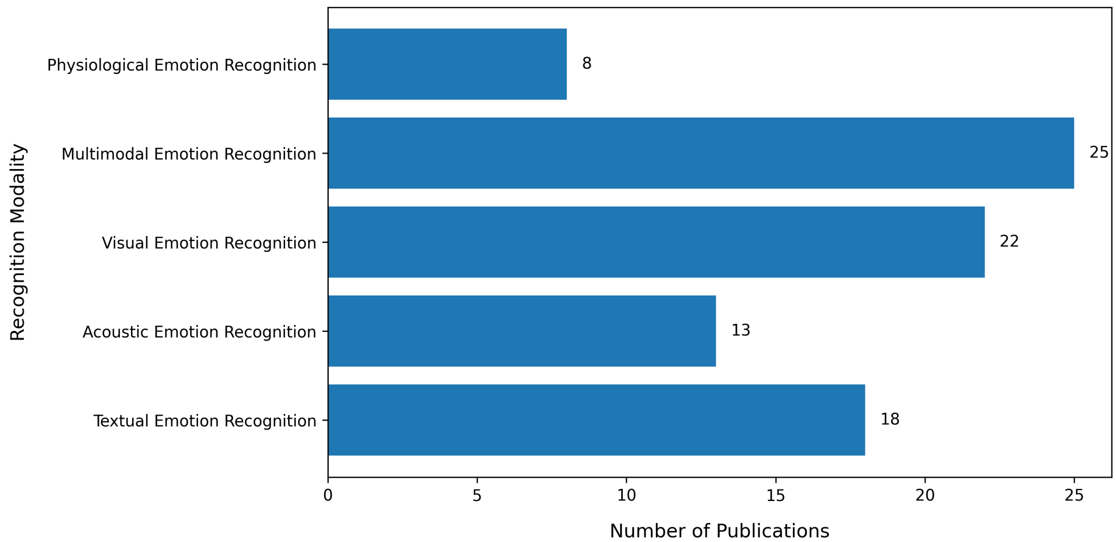

Figure 3: Breakdown of emotion understanding research by modality

2.3 Eligibility

The full text of the 500 remaining reports was retrieved and assessed for eligibility based on the following criteria:

- Inclusion criteria:
  The review incorporated studies published in English, including peer-reviewed journal articles, conference proceedings, books, and preprints from reputable repositories. To be included, a publication was required to explicitly address emotional intelligence in intelligent agents and demonstrate relevance to at least one of the three core capabilities examined in this study: (i) emotion understanding, such as emotion recognition from textual, acoustic, visual, or physiological data; (ii) affective cognition, including computational models of emotion elicitation, internal emotional state representation, or cognitive–affective architectures; and (iii) emotional expression synthesis, encompassing the generation of affective responses through textual, acoustic, or visual modalities. Both empirical studies presenting quantitative evaluations and theoretical or review articles offering insights into AI architectures, models, datasets, or applications were considered. Duplicate records identified across multiple databases were systematically removed to ensure a unique set of studies for analysis.

- Exclusion criteria:
  Non-English publications were excluded to maintain consistency in analysis and interpretation. Studies that did not explicitly focus on intelligent agents—such as works limited to general human psychology without a clear artificial intelligence context—were omitted. Additionally, research addressing unrelated topics, including non-affective AI tasks (e.g., conventional pathfinding or image classification without emotional components), was excluded. Low-quality sources, such as non-peer-reviewed blogs, opinion pieces, and outdated technical reports lacking empirical validation or academic citations, were also filtered out. Finally, all duplicate publications across the searched databases were removed to maintain a non-redundant corpus for the systematic review.

After this rigorous assessment, 202 reports were excluded, leaving 298 studies that formed the core basis of our survey.

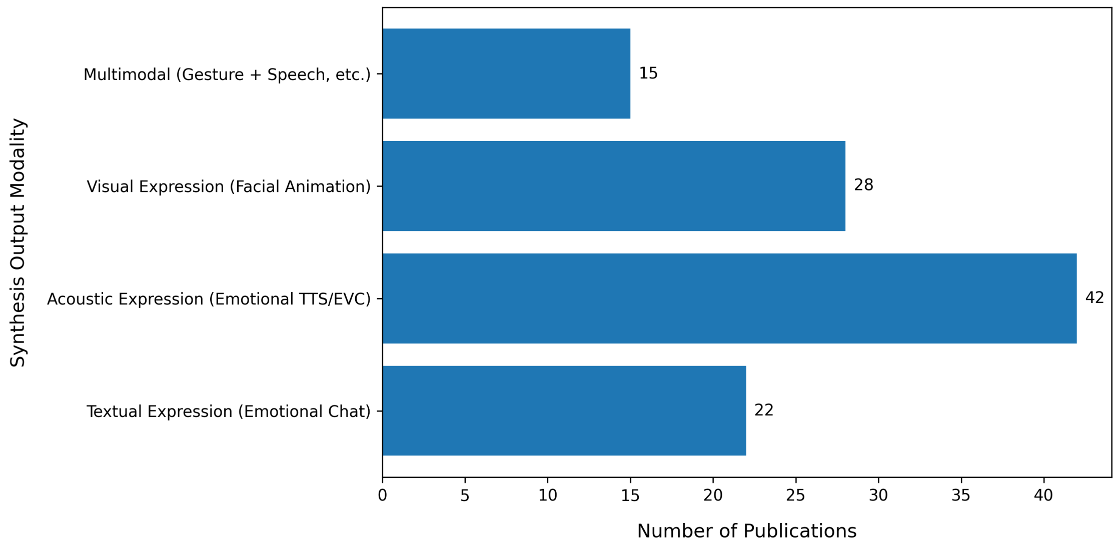

Figure 4: Breakdown of emotional expression synthesis research by output modality.

2.4 Inclusion

The final 298 studies were subjected to an in-depth analysis to synthesize key findings, identify emerging trends, compare methodologies, and explore challenges and future directions outlined in this paper. To offer a comprehensive view of the research landscape, we categorized these studies based on their primary focus across the three core capabilities of emotional intelligence in intelligent agents. As depicted in Fig. 2, the distribution of the 298 studies highlights that 32% of the research effort is devoted to emotion understanding, which includes techniques for recognizing emotions from textual, acoustic, visual, and physiological data. Emotional expression synthesis, focusing on generating emotional responses in text, speech, or visual forms, accounts for 30% of the corpus. The remaining 25% addresses affective cognition, exploring mechanisms for eliciting, representing, and integrating emotions into cognitive architectures. The residual 13% comprises foundational works and general surveys, providing a broader context for the field. A more detailed breakdown of each capability is presented in Figs. 3, 5, and 4, which highlight specific sub-categories within these domains. Fig. 3 illustrates the modalities within emotion understanding, revealing a strong emphasis on multimodal recognition frameworks. Fig. 4 delineates the output modalities for emotional expression synthesis. Lastly, Fig. 5 categorizes the diverse theoretical and computational models in affective cognition, ranging from cognitive architectures to appraisal theories. This structured analysis serves as the foundation for the detailed state-of-the-art review presented in the subsequent sections of this paper.

3 Emotion Understanding

Emotion recognition is a fundamental task in artificial intelligence (AI) that aims to understand and interpret human emotions expressed through various forms of data Afzal et al. (2024). This process identifies specific emotional states using facial expressions Canal et al. (2022), vocal tones Wani et al. (2021), language Deng & Ren (2021), and physiological signals Dadebayev et al. (2022). Emotion recognition is crucial in enhancing HCI, mental health monitoring, and other applications where understanding human emotional responses is essential Nayak et al. (2021). The inputs for emotion recognition are diverse and multimodal, making the task challenging and an intriguing opportunity. Text-based inputs, such as social media posts, reviews, or dialogues, are processed using Natural Language Processing (NLP) techniques to identify emotional nuances in the language. Machine learning models analyze words, phrases, sentence structure, and context to predict the underlying emotional tone of a text. In contrast, vision-based inputs leverage computer vision techniques to analyze facial expressions, body movements, and gestures, often using deep learning models for recognition. These visual cues offer valuable data to help AI systems accurately interpret emotions. This is especially

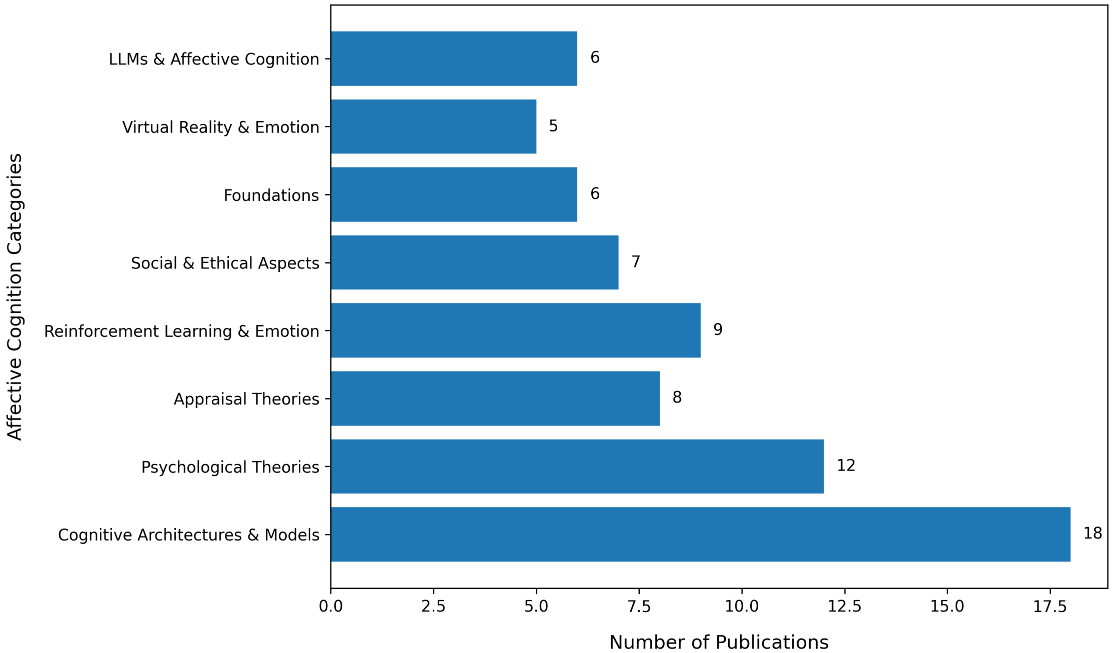

Figure 5: Categorization of affective cognition research by theoretical and computational focus.

important in real-time applications such as video conferencing or emotion-aware robotics. In addition to text and vision, sound or speech signals are critical for emotion recognition. Audio inputs, such as tone of voice, pitch, volume, and rhythm, are key indicators of emotional states. Finally, physiological signals, including heart rate, skin conductance, and facial muscle activity, provide additional insights into emotional responses, obtained through wearable devices or biosensors. These physiological markers reflect the autonomic nervous system's reactions to emotional stimuli. They are instrumental in applications like mental health monitoring and affective computing, where a deeper understanding of emotional states is necessary for more personalized interactions.

3.1 Approaches

Figure 6 illustrates the overall emotion recognition framework, highlighting the flow of data through successive processing stages and model learning components. The process begins with input data acquisition, which may involve multiple modalities, including textual data (e.g., written language), visual data (images or video), acoustic data (speech or audio signals), and physiological signals (e.g., heart rate or electroencephalography (EEG)). Each modality offers complementary cues for inferring emotional states. For example, textual data can convey sentiment through lexical and semantic patterns, while facial expressions captured in visual data provide observable emotional indicators. Similarly, physiological signals such as heart rate variability can reflect stress or relaxation levels. The diversity of input modalities reflects the multifaceted nature of human emotions and motivates the integration of multimodal data for robust emotion recognition. Following data acquisition, preprocessing is applied to clean and standardize the raw inputs, which often contain noise, artifacts, or inconsistencies. Preprocessing techniques vary by modality and may include background noise removal in audio signals, normalization and tokenization in textual data, or resizing and cropping in visual data. This stage is important for improving data quality and ensuring that learning models are not adversely affected by irrelevant variations or noise. Feature extraction aims to derive informative representations from the preprocessed data. In textual modalities, this may involve extracting emotion-related lexical features or syntactic patterns. Visual data can yield features related to facial expressions or motion dynamics, while acoustic features such as pitch, intensity, and spectral characteristics are commonly used to infer emotional states. For physiological signals, features such as heart rate variability or EEG frequency bands are extracted to capture underlying affective responses. These features constitute the foundational representations used for emotion classification. Modality processing determines how information from different data sources is handled. In unimodal processing, each modality is analyzed independently using a dedicated learning model. In contrast, multimodal processing leverages complementary information across modalities to improve recognition performance through data fusion. Two primary fusion strategies are commonly employed:

- **Feature-level fusion:** Features from multiple modalities are combined at an early stage to form a unified representation that is subsequently input to a learning model.

- **Decision-level fusion:** Each modality is processed independently, and the resulting predictions are aggregated at a later stage to produce a final emotion classification.

Finally, learning models map the extracted and fused features to emotional outcomes. These models are trained to identify patterns within the feature space and to classify emotional states based on learned representations.

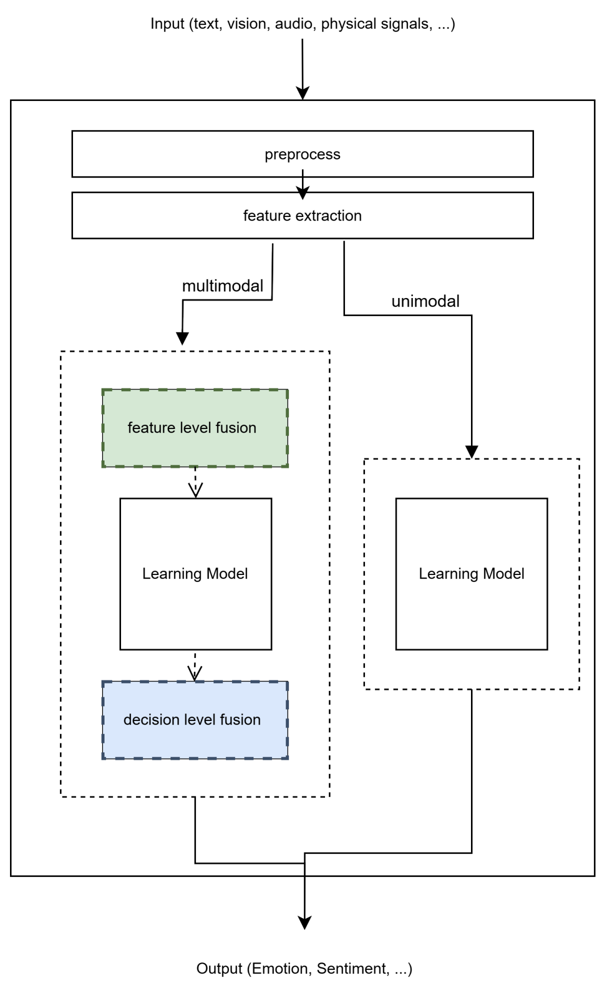

Figure 6: The overall framework of emotion recognition.

Figure 7: Challenges in emotion understanding

3.2 Challenges

In this section, we aim to address the existing challenges in emotion recognition and explore the solutions that have been proposed for each challenge. We will review relevant studies that address these issues and their suggested remedies. These challenges arise from the complex nature of human emotions and the limitations of current technologies used to process and interpret them. As depicted in Figure 7, we classify these challenges into three major domains: data, learning models, and problem nature. Table 2 summarizes these categories, outlining their respective sub-issues and highlighting studies that address these challenges. Next, we will discuss the challenges and the methods proposed to overcome them.

Table 2: Summary of challenges and solutions in emotional understanding

Challenges	Sub-challenge	Solutions
Data-related	Small size of data	Semi-supervised graph contrastive learning (Ye et al., 2024); GAN-based synthetic data generation (Schuller et al., 2024); BERT-based augmentation techniques (Koufakou et al., 2023)
	Noise in data	Advanced data augmentation methods (Chowdary et al., 2023); Uncertainty-aware multimodal fusion (Tellamekala et al., 2023); Noise-robust CNN architectures (Oh & Kim, 2024)
	Data imbalance	SMOTE and Tomek Links (Ghafourian et al., 2022); GAN-based augmentation (Meng et al., 2024); Cost-sensitive learning (Li & Deng, 2016)
	Low data quality	Multi-source transfer learning (Sarkar et al., 2023); Probabilistic uncertainty modeling (Lo et al., 2023); Differential entropy features (Uyanık et al., 2022)
Learning-related	Low accuracy	Multi-modal integration (Umair et al., 2024); Ensemble learning methods Younis et al. (2022); Unsupervised representation learning Ross et al. (2023)
	Interpretability	Attention mechanisms Cortiñas-Lorenzo & Lacey (2023); Grad-CAM visualization Sharma et al. (2024); Neurosymbolic AI frameworks Cambria et al. (2024); Coarse-to-fine training Lian et al. (2024)
	Standardized metrics	Zero-shot evaluation Schuller et al. (2024); Quantitative quality metrics Languré & Zareei (2024)
	Generalization	Subject-independent models Younis et al. (2022); Hybrid transformer approaches Zanwar et al. (2022); Emotion-specific pretraining Aroyehun et al. (2023)
Problem Nature	Variety of emotions	Graph contrastive learning Ye et al. (2024); Multi-modal gating mechanisms Sharma et al. (2024); Sensory knowledge integration Zhao et al. (2025)
	Overlap of emotions	Uniform label annotation Du et al. (2024); Causal intervention Yang et al. (2023); Graph attention networks Li et al. (2023b)
	Experiment design	Real-world multimodal studies Younis et al. (2022); Cross-subject evaluation Hu et al. (2021); Continuous emotion assessment Bota et al. (2019)
Multi-modal	Common representation	Rule-based conversion Kumar et al. (2024); Attention mechanisms Geetha et al. (2024); Semantic alignment networks Ezzameli & Mahersia (2023)
	Computational efficiency	Sparse cross-modal attention Dai et al. (2021); Hierarchical processing Wei et al. (2022); Multiplicative fusion Mittal et al. (2020)
	Modality gap	Trimodal integration Alsaadawi et al. (2024); Cross-attention mechanisms Rajan et al. (2022); Contactless data fusion Khan et al. (2024)

Continued on next page

Table 2: Summary of Challenges and Solutions in Emotional Understanding (Continued)

Challenges	Sub-challenge	Solutions
	Optimal fusion	Novel feature extractors Middya et al. (2022); Shifted Window Transformers Kim & Hong (2024); Speaker-aware networks Guo et al. (2024)
LLM/FM	Annotation cost	Synthetic data generation Schuller et al. (2024); LLM-based labeling Li (2024); Coarse-to-fine training Lian et al. (2024)
	Contextual understanding	Enhanced prompting Amin et al. (2024); Multimodal datasets Zhang et al. (2023c); Unified feature conversion Kumar et al. (2024)
	Hallucinations	Semantic entropy detection Farquhar et al. (2024); Fact-checking methods Sahoo et al. (2024); Reinforcement learning Li et al. (2024a)

3.2.1 Data-Related Challenges

Challenges related to data availability and quality, including limited dataset sizes, noisy or low-quality samples, and class imbalance, significantly constrain the accuracy, robustness, and generalizability of emotion recognition systems. Addressing these challenges is critical for the development of reliable affective computing models.

Small dataset size: The limited size of datasets represents a major obstacle for emotion recognition models, as it restricts their ability to generalize across diverse emotional expressions and populations. Effective emotion recognition requires large-scale, diverse datasets to enable capturing the complexity and variability of human emotion. However, many existing datasets are small, domain-specific, and biased toward particular demographic groups, which undermines their practical applicability. To mitigate this issue, several studies have explored data-efficient learning and data augmentation strategies. For instance, Ye et al. (2024) proposes a semi-supervised dual-stream self-attentive adversarial graph contrastive learning framework for EEG-based emotion recognition. By leveraging a small set of labeled samples alongside a larger pool of unlabeled data, this approach improves recognition performance under low-resource conditions. The dual-stream architecture effectively captures both structural and non-structural features, alleviating the limitations imposed by insufficient labeled data.

In a complementary direction, generative approaches have been investigated to expand dataset size and diversity. Schuller et al. (2024) examines the use of generative techniques Bond-Taylor et al. (2021), highlighting the role of adversarial learning and generative adversarial networks (GANs) Hajarolasvadi et al. (2020) in synthesizing realistic affective data. By constructing large-scale synthetic datasets that approximate authentic emotional expressions, such methods can partially overcome the scarcity of annotated emotional data. In the context of textual emotion recognition, Koufakou et al. (2023) explores multiple data augmentation strategies to address both small dataset size and class imbalance. Their study applies contextual word replacement using BERT-based models Devlin et al. (2019), where words are substituted based on surrounding context to generate semantically coherent training samples. Additionally, synthetic data generation using pre-trained language models is employed to enhance dataset diversity and improve model generalization. Traditional augmentation techniques, such as Easy Data Augmentation (EDA) Wei & Zou (2019), including synonym replacement, word insertion, deletion, and word shuffling, have also demonstrated effectiveness. By enriching datasets with balanced and diverse examples, these methods substantially improve the performance of deep learning models, like RoBERTa Liu et al. (2019), especially in low-resource scenarios.

Noise in data: Noise or irrelevant information that disrupts the recognition process can affect emotion recognition systems, especially those relying on sensors or video recordings. Noise may arise from poor lighting, background interference in video data, or ambient noise in audio recordings. This noise complicates distinguishing relevant emotional cues from distracting information. Chowdary et al. (2023) addresses this issue by employing data augmentation techniques during the training of deep learning models. Data augmentation includes strategies such as translations, normalizations, cropping, adding noise, and scaling, which effectively increase the variety and quantity of training data. By incorporating these modifications, the model becomes more robust to variations and noise encountered in real-world data. This approach enhances the model's ability to generalize well across different scenarios, improving facial emotion recognition accuracy even in noise conditions. Tellamekala et al. (2023) addresses noise in data, especially within the visual modality, by introducing an uncertainty-aware multimodal fusion approach that quantifies modality-wise aleatoric uncertainty. During the test phase, this paper simulates noise by applying face masking to half of the evaluation sequences. The proposed COLD fusion framework adapts to noise by dynamically estimating fusion weights. When the visual modality is obscured, COLD fusion increases reliance on the audio modality, effectively compensating for the loss of visual information. Oh & Kim (2024) presents a noise-robust deep learning model for emotion classification using facial expressions, addressing challenges posed by image noise, such as lighting variations, facial angles, and distortions. To make the model more intelligent and adaptable, it applies data augmentation tricks including image rotations, brightness shifts, and strategic cropping, mimicking real-world variations. State-of-the-art deep networks, such as CNNs O'Shea & Nash (2015), ResNet He et al. (2016), and VGG Simonyan & Zisserman (2014), enhanced with attention mechanisms Vaswani et al. (2017), facilitate the model to focus on the most informative facial regions, even under poor conditions. By integrating transfer learning with auto-encoder-based noise reduction, the model effectively learns from high-quality features while filtering out unwanted distortions. This boosts accuracy and enhances the reliability of emotion recognition in noisy environments.

Data imbalance: Another challenge in emotion recognition is data imbalance, where common emotions, such as happiness, are more prevalent than rare emotions like surprise. This imbalance can lead to model bias, resulting in better performance for recognizing common emotions while struggling with less frequent ones. Data imbalance can undermine the stability and fairness of emotion recognition systems. Ghafourian et al. (2022) employs a combination of SMOTE (Synthetic Minority Over-sampling Technique)Chawla et al. (2002) and Tomek Links. SMOTE generates synthetic samples for underrepresented emotions by interpolating between existing minority-class samples, ensuring more diverse and realistic training data. Meanwhile, Tomek Links removes borderline majority-class samples that may confuse the model, improving class separation. This dual approach balances the dataset without duplicating data, enabling deep learning models such as VGG-16 and ResNet-50 to learn more effectively from all emotion categories and improving overall classification accuracy. Data augmentation using GANs Hajarolasvadi et al. (2020) produces new samples for less common emotional categories, which improves data distribution Meng et al. (2024). This paper employs sampling strategies to promote balanced learning and avoid incorrect conclusions that can result from imbalanced data. Another strategy is loss sensitivity, which involves designing a class-sensitive loss function that pays more attention to less prominent classes. Finally, feature integration is utilized, where a deep variational auto-encoder merges complementary semantic information from multiple modalities, such as text, audio, and images, to enhance emotion recognition. Li & Deng (2016) reviews and compares several methods in imbalanced learning, which are categorized into algorithmic modification techniques and data distribution modification methods. Algorithmic approaches include Cost-Sensitive Learning (SVM-Weight)Iranmehr et al. (2019), which assigns higher misclassification penalties to minority emotions. This ensures that models pay more attention to these underrepresented categories. Data distribution approaches involve under-sampling to remove excess samples from majority classes and Over-Sampling (Random Over-Sampling and SMOTE) to generate new samples for minority classes synthetically. Iranmehr et al. (2019) introduces virtual facial sample generation (VFSG), a novel technique that enhances data diversity by creating synthetic facial images with variations in lighting and angles, offering a more realistic and robust solution than standard over-sampling methods.

Low data quality: Low-quality data, such as blurry images, can harm the performance of emotion recognition systems. These algorithms rely on high-quality and precise data, so any decrease in quality can result in incorrect predictions and reduce the overall performance of the model. To tackle this issue, Sarkar et al. (2023) incorporates multi-source transfer learning (MSTL) Lee et al. (2019) combined with multivariate correlation analysis (MCA)Abdi (2003) for facial emotion recognition. This method focuses on extracting relevant and high-quality features from multiple pre-trained models that have been trained on diverse, high-quality datasets, rather than relying solely on noisy or low-quality target data. The MCA technique selects only the most correlated and valuable features, which helps minimize the negative effects of noisy or low-resolution images. The approach also reduces the risk of negative transfer, ensuring that only beneficial knowledge from source domains is transferred, while avoiding the inclusion of noisy or irrelevant patterns. This technique enhances model performance in low-data and few-shot scenarios by effectively integrating knowledge from multiple data sources. It enables offsetting the poor quality of the target dataset without needing explicit data denoising or super-resolution methods. Lo et al. (2023) proposes probabilistic data uncertainty learning, which models the uncertainty caused by poor resolution. They use the Emotion Wheel theory to handle label ambiguity by representing emotions in a continuous space, which enables the model to better distinguish between similar expressions. These methods contribute to improving recognition accuracy and uncertainty estimation in low-resolution settings. Uyanık et al. (2022) addresses the challenge of low-quality EEG data by leveraging Differential Entropy (DE)Duan et al. (2013) as a robust feature extraction method, which is less sensitive to noise and variations in the signal. It also applies preprocessing techniques, such as band-pass filtering, to remove unwanted frequencies and reduce artifacts. To mitigate the impact of noisy data, the study employs machine learning models such as Support Vector Machine (SVM) Hearst et al. (1998) and Neural Networks, which can generalize and filter out unreliable patterns from the dataset. Combining these techniques enhances the accuracy of automated emotion recognition in virtual reality environments, even when dealing with low-quality EEG signals.

3.2.2 Model-Related Challenges

Learning models are crucial for emotion recognition, as they map the given input, such as facial expressions, vocal cues, or physiological measurements, into corresponding emotional states. However, developing robust models for this task involves several challenges.

Low accuracy: Low accuracy in emotion recognition models often stems from overfitting, where models that excel on training data struggle to generalize to unseen samples. This problem is especially common when training on small or imbalanced datasets, which often fail to capture the full complexity and diversity of human emotions. Relying on a single data modality (such as facial movements or text) limits a model's ability to capture the multifaceted nature of emotions and can restrict accuracy. To combat this, researchers are developing innovative methods to boost the performance of emotion recognition systems, including multimodal fusion, more diverse datasets, robust learning strategies, and domain adaptation techniques. Umair et al. (2024) fuses multiple data streams, including facial expressions, voice, and text, to overcome the limitations of unimodal systems. Ensemble learning methods have been utilized, as noted in Younis et al. (2022), which have significantly improved accuracy. These solutions highlight the importance of integrating diverse methods and ensemble techniques to boost performance. Ross et al. (2023) addresses the challenge of low accuracy by proposing an unsupervised multi-modal representation learning framework by utilizing stacked convolutional auto-encoders to learn latent representations from wearable bio-signals, specifically electrocardiogram (ECG) and electrodermal activity (EDA) data.

Interpretability: Understanding decision-making in complex deep learning models remains a significant challenge, particularly in sensitive domains such as healthcare and psychological analysis, where trust and transparency are critical. Deep learning architectures, such as CNNs O'Shea & Nash (2015) and Recurrent Neural Networks (RNNs) Schmidt (2019), are referred to as "black boxes" because their internal reasoning is difficult to interpret. While these models can produce acceptable results, explaining how or why a particular prediction was made remains difficult. This lack of transparency can be problematic, especially in applications where clarity is essential, such as medical or psychological contexts. To address this, Cortiñas-Lorenzo & Lacey (2023) categorizes explainability approaches into three types: Pre-model (before training), In-model (during training), and Post-model (after training). Pre-model approaches focus on designing meaningful features and reducing data complexity. In contrast, in-model methods aim to develop architectures that are inherently interpretable, such as attention mechanismsVaswani et al. (2017) that highlight salient portions of the input. Post-model techniques, such as Grad-CAMSelvaraju et al. (2017) or sensitivity analysis, enable visualization and analysis of model decisions after training. Sharma et al. (2024) employs a mechanism called Grad-CAM (Gradient-weighted Class Activation Mapping) to analyze and visualize the attention maps related to the predictions of the model. This method highlights the regions in meme images that the model references when making emotional classifications. This interpretative framework demonstrates how the model correlates visual elements of memes with their associated emotions. It provides insights into which aspects of a meme, such as facial expressions or contextual elements, contribute to the predicted emotions. Cambria et al. (2024) introduces a neurosymbolic AI framework that combines commonsense knowledge representation with hierarchical attention networks to enhance transparency in affective computing. Unlike black-box deep learning models, this framework employs a three-step normalization process, including syntactic and pragmatic, to map input text to interpretable conceptual primitives. This approach ensures that the decision-making process is both understandable and explainable. By combining symbolic AI, which utilizes structured reasoning through commonsense knowledge graphs, with sub-symbolic AI that employs pattern recognition via deep learning, the model achieves traceable and trustworthy sentiment classification and personality prediction. Lian et al. (2024) tackles the challenge of interpretability in Explainable Multimodal Emotion Recognition (EMER) through a two-stage training framework. By integrating multimodal inputs, including audio, video, and text, the approach strengthens the link between the inputs and the predicted emotions, enabling more transparent reasoning about the outputs. In the first stage, the model learns a coarse mapping using a large-scale, coarsely labeled dataset (EMER-Coarse), which facilitates the identification of general emotional trends. In the second stage, the model utilizes a smaller, manually checked dataset (EMER-Fine) to refine these mappings, ensuring better alignment with reliable human-annotated labels. This systematic method not only improves emotion recognition accuracy but also provides a framework for tracing predictions to the supporting multimodal evidence, enhancing the model's interpretability.

Lack of standardized evaluation metrics: The lack of standardized evaluation metrics for emotion recognition complicates the comparison of results across various studies and hinders the assessment of the actual effectiveness of proposed models. This inconsistency can lead to uneven reporting, hindering meaningful cross-study comparisons. Schuller et al. (2024) discusses the challenges arising from the absence of standardized evaluation metrics in affective computing, particularly in assessing the capabilities of foundation models (FMs). This paper advocates for developing new methods and metrics that provide a rigorous scientific evaluation of these emerging models. It emphasizes the need for comparative benchmarks, such as the use of established datasets for emotion recognition tasks. Additionally, it highlights the importance of refined approaches, like zero-shot classification, to ensure that model performance can be assessed in a meaningful way. Similarly, Languré & Zareei (2024) addresses the lack of standardized evaluation metrics in text emotion detection (TED). The study proposes a comprehensive framework for assessing dataset quality based on 14 quantitative metrics across four key dimensions: representativity, readability, structure, and part-of-speech (POS) tag distribution. By systematically measuring factors such as data balance, linguistic complexity, lexical diversity, and syntactic structure (POS tag distribution), the framework ensures that datasets are evaluated using consistent and reproducible criteria. It demonstrates that variations in these quality metrics impact model performance, reinforcing the need for standardization in TED research. By introducing these metrics and validating their influence through experiments with BiLSTM Huang et al. (2015) and Bert models Devlin et al. (2019), the paper establishes a foundation for more rigorous dataset assessment, enabling more reliable and comparable TED model evaluations.

Generalization: A critical challenge in emotion recognition systems lies in their limited ability to generalize across diverse domains, including cultural variations, environmental contexts, and different input modalities (e.g., speech versus facial expressions). This domain shift problem often leads to performance degradation when models trained on one dataset are applied to new scenarios. Recent work by Younis et al. (2022) emphasizes the need for subject-independent predictive models to improve cross-domain robustness. These models do not rely on the unique characteristics of specific individuals, making them more versatile for real-world applications. They are designed to identify emotions across different populations with varying physiological and psychological characteristics, enhancing their utility in diverse settings. In Zanwar et al. (2022), a hybrid approach is proposed that combines transformer-based models, such as Bert Devlin et al. (2019) and RoBERTa Liu et al. (2019), with psycholinguistic features extracted using BiLSTM Huang et al. (2015) networks. This combination improves the model's generalization in unseen datasets by leveraging state-of-the-art language representations and deeper contextual information. Similarly, in Aroyehun et al. (2023), a novel pre-training method called eMLM (emotion-specific Masked Language

Model) is introduced. This method masks emotion-related words during training, encouraging the model to learn broader contextual relationships rather than relying solely on specific terms. Combining eMLM with fine-tuning allows the LEIA model to adapt effectively to domain shifts and perform well on new datasets. Ross et al. (2023) reduces the reliance on human-annotated labels, enabling the aggregation of multiple datasets into a more extensive and diverse training set. This approach enhances the model's generalization ability across different emotional contexts, improving classification performance. The results demonstrated that the proposed method achieved state-of-the-art accuracy across multiple datasets compared to baseline techniques, illustrating its effectiveness in overcoming the limitations associated with lower accuracy in existing models.

3.2.3 Problem Nature Challenges

The nature of the problem in emotion recognition poses significant challenges, as emotions are inherently subjective, overlapping, and context-dependent. Unlike fixed categorical tasks, emotions exist on a continuous spectrum, making it difficult to establish clear distinctions between them. Additionally, emotional expressions can vary widely between individuals and cultures, and environmental factors influence them, leading to inconsistencies in labeling. The presence of multiple modalities adds another layer of complexity, as different modalities can convey conflicting emotional cues. These factors make emotion recognition a complex task that requires adaptive models capable of handling ambiguity, variability, and contextual dependencies.

Variety of emotions: The variety of human emotions poses a significant challenge for emotion recognition systems, which need to address the complex range of emotional states that individuals experience. While basic emotions such as happiness, sadness, anger, and fear serve as foundational categories, human emotional experiences extend beyond these simple labels. Furthermore, emotions can manifest differently in each individual due to factors like personality, context, and socio-cultural norms. These complexities make it difficult to distinguish between emotions. Consequently, emotion recognition systems risk oversimplifying and misinterpreting the depth and context of human emotions. Ye et al. (2024) proposes a semi-supervised learning framework designed specifically for EEG-based emotion recognition. This framework employs graph contrastive learning You et al. (2020) to enhance the model's ability to differentiate between various emotions. The framework leverages advanced learning strategies to tackle emotional diversity, enabling it to more accurately detect and represent emotions in EEG signals, even when labeled data are scarce. Sharma et al. (2024) proposes a multimodal neural framework designed to tackle the challenge of emotional diversity in understanding human emotions conveyed through memes. This framework specifically models enhanced visual cues related to emotions and utilizes a gating mechanism for effective integration of multiple modalities, ultimately improving emotion recognition in memes. Zhao et al. (2025) presents a model that integrates sensory knowledge into the T5 framework Xue et al. (2020) to enhance emotion classification. This model embeds sensory information within the attention mechanism, which enhances contextual understanding and increases sensitivity to subtle emotional states. Notably, the model employs an adapter approach that facilitates the joint training of contextual and sensory information through a unified loss function.

Overlap of emotions: The overlap and coexistence of emotions complicate the accurate identification and classification of emotional states. For example, emotions such as fear and surprise share similar physiological responses and facial expressions, making it difficult to discern where one emotion ends and another begins. The subjective nature of emotional experiences, the dynamic shifts between different emotional states, and cultural differences in emotional expression heighten this ambiguity. Du et al. (2024) presents a dataset featuring uniformly independent labels across different modalities, which enables a more accurate understanding of emotions and tackles the challenges posed by mixed and overlapping emotions in emotion recognition. This approach addresses the contradictions frequently encountered in emotional expression. It implements robust annotation methods and evaluates inter-annotator agreement Braylan et al. (2022) to ensure the reliable labeling of mixed emotions. These improvements enhance the dataset's quality for training emotion recognition models, thereby increasing the technology's capability to interpret complex human emotions. Yang et al. (2023) offers a causal inference-based approach using the Contextual Causal Intervention Module (CCIM) to remove spurious correlations between context and emotion labels. Traditional emotion recognition models often misclassify emotions due to context bias, where certain emotions are disproportionately associated with specific backgrounds. The CCIM mitigates this by applying causal intervention to disentangle genuine emotional cues from misleading contextual dependencies. It constructs a confounded dictionary that clusters contextual features and re-weights their influence using a backdoor adjustment technique. This ensures emotions are classified based on intrinsic expressions rather than external biases. GraphMFTLi et al. (2023b) leverages a graph-based multi-modal fusion approach that effectively integrates intra-modal and inter-modal contextual information. GraphMFT develops three separate heterogeneous graphs: Visual-Acoustic, Visual-Textual, and Acoustic-Textual. This structure facilitates a more refined and adaptive fusion of different modalities. The model employs graph attention networks (GATs) to assign distinct importance weights to contextual and cross-modal relationships, allowing it to capture nuanced differences between emotions such as "frustration" and "anger" or "excitement" and "happiness." The improved GAT architecture also mitigates the over-smoothing problem often found in deep graph networks, preserving distinct emotional features. By incorporating speaker embeddings and learning dynamic interactions across different modalities, the model reduces the misclassification of overlapping emotions, thereby improving the overall accuracy of emotion detection.

Proper design of experiments: Proper experimental design significantly affects the reliability, generalizability, and accuracy of resulting models. It is essential to consider the diversity and complexity of human emotions, which can vary widely among individuals and cultures. Poorly designed experiments can lead to models that are inaccurate or biased due to cultural differences. Replicating real-world emotional dynamics in controlled settings is inherently challenging, as emotions are shaped by numerous complex and often unpredictable factors. Addressing these challenges necessitates careful attention to these elements to ensure that trained models accurately represent real-world scenarios and are applicable in practice. In this context, the article Younis et al. (2022) conducts a real-world study that captures participants' natural emotional responses in their environment. It develops subject-independent models by integrating multi-modal data, combining physiological and environmental data to create a comprehensive dataset. Traditional emotion recognition models typically use predefined video clips to elicit emotions, so all participants watch the same videos. This setup can introduce unwanted noise patterns in the data, leading models to focus on features related to the video content instead of capturing authentic emotional responses. To address this challenge, Hu et al. (2021) introduces a novel experiment setup that mitigates the impact of stimulus materials on classification accuracy in cross-subject studies. The paper proposes an innovative approach to data partitioning, where the training and testing sets are sourced from different video-induced datasets, instead of using multiple participants who are watching the same videos. This method ensures that the model learns generalizable emotion patterns and reduces the risk of overfitting to specific stimuli. The study evaluates this approach using public emotion datasets and demonstrates that it enhances cross-subject emotion recognition. As a result, the findings are more reliable and applicable to real-world scenarios. Bota et al. (2019) addresses the challenges of designing experiments for emotion recognition, especially when comparing controlled laboratory settings to real-life situations. It identifies several issues, such as the difficulty in accurately annotating emotional responses in unconstrained environments, the degradation of signal quality due to noise from uncontrolled variables, and the significant impact of individual factors like mood and cultural background. To tackle these challenges, the authors suggest focusing on unconstrained scenarios that feature a diverse pool of participants and conducting continuous evaluations of emotional responses. This approach aims to enhance the validity, reliability, and generalizability of findings in emotion recognition research.

3.2.4 Multi-Modal Challenges

Fusing multiple modalities is a crucial step in developing robust emotion understanding models. These models integrate diverse data sources, such as audio, visual, and textual cues, each offering distinct insights into emotional states. Effective fusion requires precise alignment of modalities to preserve their individual contributions, a process that can be computationally costly, especially with large-scale datasets.

Furthermore, the inherent differences in how each modality encodes emotional information can complicate the optimization of their interactions. To ensure these sources complement rather than conflict with one another, fusion strategies must be carefully designed to balance performance and efficiency. Addressing these challenges demands advanced architectures and substantial computational resources, ultimately enhancing the system's ability to generalize across varied emotional expressions and contexts.

Common representation space: Achieving a common representation space in emotion recognition requires integrating diverse modalities, each capturing emotions in distinct ways. These modalities do not naturally align, creating gaps that make unifying them into a single representation challenging. Moreover, the variability of emotional expressions across individuals and cultures adds another layer of complexity, demanding advanced techniques to align these heterogeneous data sources effectively for accurate emotion recognition. Kumar et al. (2024) developed a rule-based system that converts non-verbal cues into text to address this issue. These textual representations are combined into prompts and processed by a pre-trained LLM, allowing for more effective emotion recognition. This approach facilitates multimodal integration while maintaining flexibility to incorporate additional modalities in the future. Geetha et al. (2024) highlights the challenges associated with integrating various emotional cues, which often have different feature representations and temporal dynamics. It discusses advanced deep learning techniques, such as attention mechanisms and transformers, that can effectively focus on relevant features from multiple modalities to create a unified representation. Moreover, specialized architectures improve the interpretation of complex emotional states by incorporating spatiotemporal context. These approaches offer potential solutions for achieving a cohesive common representation in emotion recognition systems. Ezzameli & Mahersia (2023) employs a fusion and alignment method to create a shared representation that effectively associates different modalities, which is essential for multi-modal learning. It reviews various fusion techniques, including feature-level, decision-level, and model-level fusion, highlighting the importance of identifying relationships between different modalities and integrating data to enhance recognition accuracy. The Semantic Alignment Network (SAN)Hou et al. (2023) utilizes a Cross-Modal Alignment (CMA) module that projects heterogeneous features into a unified semantic embedding space, reducing semantic discrepancies and enabling effective fusion of emotionally relevant information. SAN further uses attention mechanisms to dynamically weight each modality's contribution, improving alignment precision and the recognition of subtle emotional nuances.

Computational resource and training cost: Emotion recognition systems often process data from multiple sources, such as audio, video, and physiological signals, which requires substantial computational resources for feature extraction, data alignment, and model training. Training advanced algorithms or deep learning models on such large and complex datasets is both time- and resource-intensive, demanding high-performance hardware. These requirements significantly increase financial costs and development time, posing challenges to the scalability and widespread adoption of emotion recognition systems. Dai et al. (2021) addresses this issue in multimodal emotion recognition by proposing a fully end-to-end model that integrates feature extraction and emotion recognition into a single pipeline, eliminating the need for hand-crafted feature engineering. To enhance efficiency, it introduces a sparse cross-modal attention mechanism, which selectively focuses on the most relevant features from different modalities, reducing redundant computations and unnecessary processing. Additionally, the model is optimized to learn directly from raw multi-modal data, removing the dependency on separately pre-processed datasets and reducing training overhead. This method effectively lowers computational costs while preserving high accuracy, making it well-suited for real-time applications and environments with limited resources. A fully end-to-end system for fast and efficient video-based emotion recognition is proposed in Wei et al. (2022). This paper incorporates a hierarchical attention mechanism that enhances the contribution of the audio modality without significantly increasing computational overhead, thereby optimizing computational efficiency. Additionally, the authors introduce a single-branch inference module for visual processing, which replaces traditional multi-branch architectures with a simplified yet effective structure. This change reduces both computational complexity and storage requirements. By integrating data preprocessing and multi-modal learning into a unified system, this system eliminates redundant computations and speeds up inference. Mittal et al. (2020) presents a multiplicative fusion strategy that adaptively adjusts the influence of each modality based on its reliability, thereby reducing redundant computations and avoiding the processing of noisy or unreliable data. By suppressing less informative modalities and prioritizing the most reliable ones, the model maintains high accuracy while optimizing computational efficiency, leading to significant reductions in both training and inference costs. Wu et al. (2025) adopts a prompt learning approach, which enables efficient fine-tuning with fewer labeled data and eliminates the need for extensive training. Task-specific prompts guide pre-trained models, substantially decreasing computation time and memory usage. Furthermore, their text–audio fusion strategy improves feature integration, reduces redundancy, and enhances inference efficiency.

Modality Gap: The modality gap refers to the differences and mismatches between various data sources, such as audio, video, and physiological signals, each capturing emotional information in distinct ways. These variations arise from factors such as sensor limitations, environmental conditions, and the inherent characteristics of each modality. For instance, facial expressions may offer clear visual cues, while voice tone or heart rate might convey subtler or complementary aspects of emotion. This gap poses a challenge, as features extracted from different modalities may not align well or may even conflict, hindering effective fusion for accurate emotion recognition. Bridging the modality gap requires advanced techniques to align, integrate, and harmonize heterogeneous data sources into a unified and reliable representation. Alsaadawi et al. (2024) addresses this issue by advocating for a trimodal affective computing approach that integrates textual, vocal, and visual data to enhance the accuracy and understanding of emotions. It outlines various data fusion strategies for combining features from multiple modalities, bridging their differences. It demonstrates improvements in emotion classification performance by employing advanced algorithms designed to process these diverse cues simultaneously. Khan et al. (2024) investigates contactless data collection methods and proposes better solutions to integrate these diverse modalities. It focuses on bridging the modality gap by developing frameworks for effectively combining different sensing techniques, emphasizing the need for fusion strategies to handle the disparities in the type, quality, and granularity of data from various sources. The paper presents novel methods to align and integrate multimodal cues by leveraging the strengths and addressing the weaknesses of each modality, ensuring meaningful contributions from all modalities and enhancing the overall accuracy and robustness of emotion recognition systems. The performance of cross-attention and self-attention mechanisms in integrating multimodal data is compared in Rajan et al. (2022). This paper examines how attention mechanisms address discrepancies across modalities. While self-attention captures within-modal dependencies, cross-attention facilitates interactions between modalities, effectively aligning features from multiple sources. The paper demonstrates that cross-attention is more effective in overcoming modality gaps by enabling the model to learn cross-modal relationships.

Optimal fusion: Optimal multimodal fusion aims to integrate diverse data modalities to maximize emotion classification accuracy. Each modality provides unique emotional cues, but its reliability varies by context, and modalities may not always align. The goal is to identify the most effective stage, including early, intermediate, or later, in the processing pipeline for fusing these modalities to optimize system performance. Middya et al. (2022) proposes a model-level fusion approach that employs specialized feature extractor networks for audio and video data to derive relevant features before integration into a cohesive multimodal model. The study systematically evaluates various combinations of these extractors, analyzing their performance across diverse configurations to determine the optimal synergy between audio and video features. This methodical approach significantly improves emotion recognition accuracy. It highlights the significance of choosing the right combination of modalities and their features to improve overall classification results. Kim & Hong (2024) employs a shifted window transformer encoder combined with symmetric cross-attention mechanisms to model complex interactions among diverse data modalities. Physiological signals are transformed into 2D images, enabling effective extraction of spatial and temporal features. The model incorporates metadata, such as environmental conditions and personal traits, to enhance the emotional context. This integrative approach facilitates adaptation to individual variability, thereby improving classification accuracy. Guo et al. (2024) proposes a speaker-aware cognitive network with cross-modal attention to address optimal fusion in emotion recognition, effectively integrating multimodal data from text, audio, and video sources while incorporating speaker-specific information. The model employs a cross-modal attention fusion module to synergistically combine features across modalities, capturing their complementary information. Subsequently, a GRU-based cognitive network module simulates conversational dynamics and leverages speaker-specific data to extract richer emotional cues.

3.2.5 Usage of LLMs/FMs

The application of foundation models and LLMs to emotion recognition presents significant challenges due to the nuanced and diverse nature of emotional cues. Although these models excel at processing vast datasets and performing across multiple tasks, their generalization to diverse emotional expressions, particularly in complex social or cultural contexts, remains inconsistent. Unlike traditional emotion recognition models tailored to specific datasets or tasks, foundation models and LLMs are designed for broad applicability, which can compromise their precision in capturing subtle emotional nuances. Fine-tuning these models for emotion recognition is challenging, as it requires balancing general knowledge with context-specific adaptability to achieve high accuracy. To address this, techniques such as domain-specific pretraining, transfer learning with emotionally rich datasets, and the integration of multimodal data (e.g., text, audio, and visual cues) can enhance model performance. Additionally, incorporating cultural and social metadata into training pipelines may improve the models' ability to interpret diverse emotional expressions, thereby advancing their effectiveness in real-world applications.

Annotation cost The high cost of data annotation poses a significant challenge in applying foundation models(FMs) and LLMs to emotion recognition. These models require extensive labeled datasets that capture the diversity of emotional expressions across varied contexts, cultures, and modalities. Annotating such datasets is labor-intensive and costly, as it involves human experts labeling nuanced emotional cues in multimodal data, such as text, audio, and video. Moreover, maintaining and updating these datasets to reflect evolving emotional expressions and emerging contexts further escalates costs. This challenge limits the accessibility and scalability of foundation models for emotion recognition, hindering their widespread adoption. Schuller et al. (2024) uses foundation models in emotion recognition to reduce reliance on specialized, annotated affective data. Large multimodal FMs, pretrained on diverse, extensive datasets, achieve robust emotion recognition performance without requiring extensive labeled datasets. This approach significantly lowers the cost and effort associated with manually annotating emotional data for training affective computing models. Furthermore, FMs can generate synthetic emotional data, further decreasing dependency on human-annotated datasets. Li (2024) investigates whether LLMs can serve as a cost-effective alternative to traditional human-based crowdsourcing for data annotation. The study compares the quality of individual annotations from human crowd workers and LLMs, evaluating various label aggregation methods to enhance overall annotation quality. Findings indicate that integrating LLM-generated labels with crowdsourced data improves the quality of aggregated annotations while reducing costs. This approach offers a scalable solution for producing high-quality labeled datasets for tasks such as emotion recognition. EMER-Coarse dataset Lian et al. (2024) consists of large-scale, coarsely labeled data for emotion recognition tasks. To tackle the high costs of manual annotation, the framework replaces costly closed-source models with open-source alternatives, streamlining the pre-labeling process and eliminating the need for manual verification. The AffectGPT framework leverages this dataset in a dual-stage training approach. In the first stage, it learns general alignments between multimodal inputs and emotion-related descriptions using the coarsely labeled data. In the second stage, the model is fine-tuned with a smaller, finely labeled dataset to enhance accuracy. This approach significantly reduces annotation effort and costs while maintaining high-quality performance in emotion recognition.

Contextual understanding: Foundation models and LLMs face challenges in recognizing emotions within nuanced contexts. Emotional expressions are strongly influenced by factors such as speaker background, social environment, and cultural norms, which these models often inadequately represent. Despite their large-scale data processing strengths, LLMs may misinterpret emotional cues, leading to inaccurate recognition in complex situations. Amin et al. (2024) examines ChatGPT models on multiple affective computing tasks, such as sentiment analysis and opinion extraction, focusing on their ability to understand contextual factors in emotional expressions. To enhance this capability, the study introduces a novel prompting framework and reformulates regression tasks as pair-wise ranking classification tasks. This structured methodology enables a robust assessment of ChatGPT's performance in discerning subtle emotional cues in text, improving contextual understanding. Lu et al. (2024) demonstrates that GPT-4V struggles to accurately recognize certain facial expressions and emotions without sufficient contextual information, particularly in complex emotional states such as fear or subjective tasks. The lack of adequate contextual information often results in errors or omissions in emotion recognition, underscoring the importance of effective contextual modeling in affective computing. DialogueLLMZhang et al. (2023c) is an emotion and context knowledge-enhanced Large Language Model (LLM) designed explicitly for emotion recognition in conversations (ERC). The model is fine-tuned on multimodal datasets, including text and video, to leverage contextual cues and emotional relationships within conversations. This approach enhances DialogueLLM's ability to accurately detect emotions by integrating conversational context and visual cues, improving its performance across diverse scenarios. Kumar et al. (2024) proposes a unified feature representation approach to enhance contextual understanding in LLMs for emotion recognition. By converting audio and visual features into a standardized text format using rule-based systems before inputting them into the LLM, the model effectively captures the full emotional context across modalities. This unification ensures consistent processing of diverse emotional cues, reducing misinterpretations and improving the LLM's ability to recognize complex emotions in varied contexts.

Hallucinations: Hallucinations in LLMs pose a significant challenge in emotion recognition, as models may generate incorrect emotional labels or infer spurious emotional cues not present in the input data. For instance, an LLM might lose an emotion based on learned patterns that are irrelevant or contradictory to the actual emotional state. This issue stems from LLMs' tendency to draw on unrelated or erroneous associations within their extensive training data, leading to inaccurate predictions. Such hallucinations undermine the reliability of emotion recognition systems, particularly in sensitive applications like mental health assessment, where precision is paramount. Sahoo et al. (2024) provides a comprehensive analysis of hallucinations in LLMs, categorizing them into contextual disconnection, semantic distortion, and factual inaccuracies. The study proposes detection and mitigation strategies, including self-checking mechanisms, fact-checking techniques, and data augmentation, to reduce hallucination risks. It also explores fine-tuning, prompt engineering, and consistency checking to enhance the factual reliability of LLMs. By establishing a structured taxonomy of hallucination types and corresponding detection methods, the paper lays a foundation for developing robust approaches to mitigate hallucinations across diverse domains, including affective computing. Farquhar et al. (2024) introduces a novel semantic entropy-based approach to detect confabulations, a specific type of hallucination in LLMs. The method measures semantic uncertainty by clustering model responses based on their semantic meaning, rather than lexical variations, and computing entropy across these clusters. High entropy reveals potential hallucinations in LLMs. Entropy indicates greater uncertainty, correlating with unreliable outputs, such as confabulations. This approach enhances the accuracy and reliability of LLMs in question-answering systems by identifying potentially erroneous outputs, and its task-agnostic nature makes it versatile across various domains, including affective computing. Li et al. (2024a) systematically studies the sources, detection, and mitigation of factuality hallucinations. It introduces a comprehensive approach using the HaluEval 2.0 benchmark, which includes 8,770 questions from various domains, including biomedicine, finance, science, education, and open domain. The paper proposes a practical framework for detecting hallucinations by extracting factual statements from model responses and then evaluating their correctness. It explores factors contributing to hallucinations, such as pre-training data, fine-tuning, prompt design, and inference methods. To mitigate hallucinations, the paper tests techniques like reinforcement learning from human feedback (RLHF), retrieval augmentation, and self-reflection, finding that RLHF significantly reduces hallucinations, especially in open-domain and biomedicine contexts.

4 Affective Cognition

The Affective Theory of Mind extends beyond emotion recognition, requiring reasoning about emotions and responding appropriately. Achieving this requires not only emotional recognition but also internal modeling of affective strategies and context-appropriate expressive behavior Raggioli et al. (2025). Although current affective computing primarily addresses emotion recognition and expression synthesis, advancing artificial emotional intelligence depends on modeling both emotional elicitation and experience Zall & Kangavari (2022). Thus, integrating cognitive and affective theories of mind is essential to enable affective cognition. This process involves two central components: first, identification of events and cognitive states that elicit emotions, and second, recognition of behaviors and cognitive states shaped by those emotions. This paper

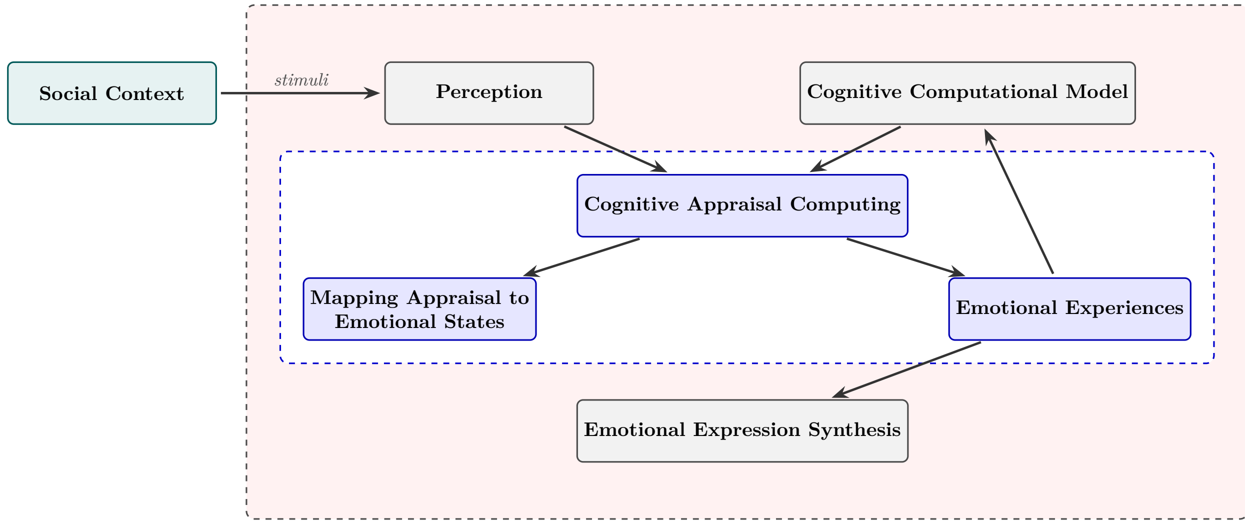

Figure 8: Block diagram of intelligent agent with affective cognition

examines affective cognition from varied perspectives, highlighting its challenges and discussing approaches for progressing artificial emotional intelligence.

4.1 Approaches

Overall, the proposed methods for modeling artificial emotional elicitation are categorized into two main groups: theory-driven Zall & Kangavari (2024) and data-driven approaches Liu et al. (2024). Theory-driven approaches often focus on cognitive appraisal theory, which is one of the most well-established psychological theories of emotional elicitation in humans. This theory outlines that emotions arise in response to both internal and external events. To elicit the appropriate emotion, it is necessary to evaluate the experienced event through various cognitive appraisals and personal experiences. These theories emphasize cognitive appraisal variables that assess an event from multiple perspectives. Some theory-driven approaches are based on a cognitive computational framework Jokinen & Oulasvirta (2025) that outlines essential cognitive information, such as concerns, goals, and needs, which are crucial for forming theory-based appraisals. The overall overview of these approaches is illustrated in Figure 8. As shown in this figure, the perceived stimuli are evaluated to calculate the cognitive appraisal variables. The mental states and experiences necessary for determining these appraisal variables are described within the cognitive computational model. Following this, the appraisal variables are mapped to specific emotional states. These resulting emotional states then influence emotional modulation, which in turn affects decision-making and emotional expression. Theory-based methods predict emotional states in future utterances by concentrating on cognitive processes. In contrast, data-driven methods analyze the conversation's history to identify specific events within the observed data and determine the emotions associated with those events for upcoming contributions.

4.2 Challenges

The modeling of emotional elicitation and experiences in intelligent agents is a crucial aspect of the affective theory of mind. A key component of this process involves developing a cognitive model that serves as a foundational framework for managing knowledge, information, experiences, and cognitive mental states related to emotions. One of the main challenges in this area is establishing reliable methods for computing cognitive appraisal variables based on underlying cognitive theories. Data-driven approaches also face significant obstacles, such as limited and ambiguous data, which impede the accurate capture and representation of the complex and multifaceted nature of human emotions. Additionally, technological and methodological barriers, including the need for robust models and the integration of LLMs, further complicate the task. Addressing these challenges is essential for advancing precise and effective models of emotional responses that are crucial for creating empathetic and socially intelligent systems. The primary challenges in this domain, along with their associated sub-challenges, are illustrated in Figure 9. Furthermore, Table 3 provides an overview of various studies and the solutions they propose for tackling these issues. In the following sections, we will explore these challenges in detail and review potential solutions to overcome them, to enhance the fidelity and applicability of emotion modeling in intelligent agents.

4.2.1 Data-Related Challenges

The field of emotion elicitation and experiences currently lacks a comprehensive and cohesive dataset. There is a clear need for data collections that explicitly clarify the relationship between various emotions and the appraisal variables and events that influence them. Integrating data-driven approaches with theoretical frameworks requires the availability of extensive and complete datasets. Although some efforts have been made in this area, a fully integrated and systematic dataset is still essential for advancing research. Developing such datasets would lead to a better understanding of emotional responses and their underlying mechanisms, ultimately contributing to the improvement of both theoretical models and practical applications in emotional analysis. Emerging evidence indicates that induction mechanisms are essential for eliciting emotions by simulating the sensory experiences needed for experimental paradigms. Over the years, many reviews have predominantly focused on passive elicitation methods, where individuals act as observers, often neglecting the importance of self-relevance in emotional experiences. To address this gap in the literature, Somarathna et al. (2022) explores the potential of Virtual Reality (VR) as an active mechanism for emotion induction. Furthermore, to ensure the effectiveness and reliability of research outcomes, VR environments must incorporate well-selected stimuli to successfully evoke specific emotional responses. Bayro & Jeong (2025) presents experimental protocols for collecting datasets with virtual reality.
Gandhi et al. (2024) proposes an automatic dataset generation that includes diverse scenarios for benchmarking affective cognition, specifically focusing on understanding and reasoning about human emotions in foundation models. It presents a pipeline for generating diverse and naturalistic stimuli that can systematically and scalability evaluate affective reasoning. The generated scenarios explore relationships between appraisals, emotions, expressions, and outcomes.

4.2.2 Model-related challenges

In this section, we review the challenges are related to learning model.

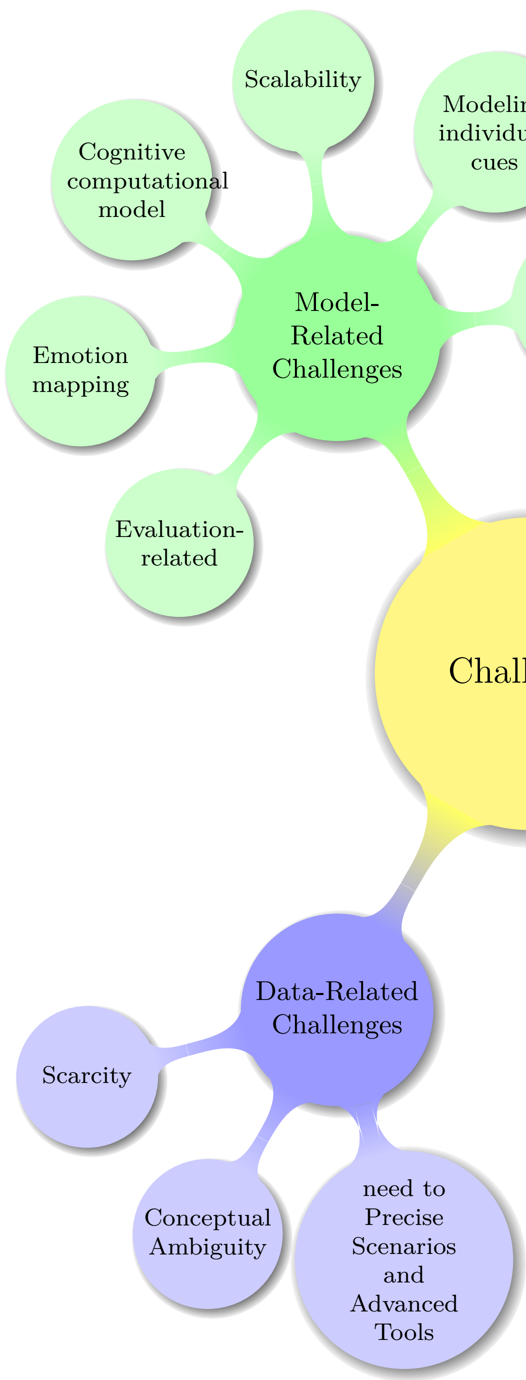

Figure 9: Challenges in affective cognition

Table 3: Summary of challenges and solutions in affective cognition

Challenge	Sub-challenge	Solution
Data-Related	Scarcity	Automatic dataset generation with diverse scenarios for benchmarking affective cognition (Gandhi et al., 2024)
	Conceptual Ambiguity	Experimental protocols for VR dataset collection (Bayro & Jeong, 2025)
	Precise Scenarios & Tools	VR as active mechanism for emotion induction (Somarathna et al., 2022)
Model-Related	Cognitive computational modeling	ACT-R architecture (emotional memory valuation) (Juvina et al., 2018); E-VOX SOAR + ALMA (real-time processing) Perez et al. (2017); EMA Automatic/deliberate appraisal Gratch & Marsella (2004); Emotional-BDI Resource management Pereira et al. (2005a); Fuzzy-based BDI Cultural/linguistic modeling Taverner et al. (2021); ABC-EBDI BDI + ABC theory Sanchez et al. (2019a); InFra Stimuli appraisal Rodriguez et al. (2016); Silicon Coppelia Ethics/aesthetics variables Hoorn et al. (2021); EIAEC Appraisal theories + memory Zall & Kangavari (2024)
	Scalability	Knowledge graphs + continuous learning Zall & Kangavari (2024)
	Emotion mapping	EIAEC Data-driven approach Zall & Kangavari (2024); EEGS Weighted mapping strategy Ojha et al. (2020b); CPM + VR data-driven Somarathna & Mohammadi (2024); Regret-based RL Soman et al. (2024)
	Emotion experiences modeling	SUSAN Inner speech modeling Corvaia et al. (2025); Neural network framework Hernández-Marcos & Ros (2024); EIAEC Context-dependent values Zall & Kangavari (2024); Memory valuations Juvina et al. (2018); Time-based annotations Tsfasman et al. (2025); RL + appraisal theory Zhang et al. (2024d); Value prediction Zhang et al. (2024b); Weber-Fechner + Q-learning Wu & Sun (2025); Emotional motivation Berto et al. (2025); Reward modulation Nikodemou & Christodoulou (2024); TD-based expressions Nijeholt & Broekens (2023)
	Evaluation	Scenario simulation & human comparison Sanchez et al. (2019b); Zall & Kangavari (2024); Multi-agent performance assessment (EBDI) Jiang et al. (2007); Emotional reasoning in critical situations Bourgais et al. (2016); Human feedback evaluation (E-VOX) Perez et al. (2017)
LLM-based	Emotional reasoning	Distinguishes self-attribution vs perception of emotions using appraisal theory Tak & Gratch (2024); Combines cognitive/affective reasoning with intrinsic motivation Raggioli et al. (2025); Explainable emotion alignment for Metaverse agents Khan et al. (2025)
	Contextual misunderstandings	Assesses affective cognition in LLMs (GPT-4, Claude-3, Gemini-1.5-Pro) Gandhi et al. (2024)

Cognitive computational modeling Theoretical approaches to emotional elicitation modeling primarily depend on cognitive models to calculate appraisal variables. These models provide the foundational information required to determine emotional states. Most methods not only specify the emotion to be expressed but also examine how emotions influence decision-making and other cognitive processes, such as inference. A significant challenge in emotion elicitation modeling is the effective implementation of cognitive models. Several computational cognitive models are available for this purpose, including cognitive architectures Anderson et al. (2004b); Laird (2019b), Belief-Desire-Intention (BDI) frameworks Pereira et al. (2005b); Sanchez et al. (2019b), and other specific designs Hoorn et al. (2021); Becker-Asano & Wachsmuth (2010). These computational models exhibit substantial differences, with each group demonstrating unique structural and functional characteristics. Cognitive architectures such as ACT-R Anderson et al. (2004a), SOAR Laird (2019a), and LIDA Franklin et al. (2007) provide structured frameworks for modeling the essential components required to produce intelligent behavior in artificial agents. By integrating diverse sources of knowledge, these architectures enable agents to address a wide range of complex problems. Typically, cognitive architectures incorporate specialized memory systems: procedural memory for action execution and planning, semantic memory for environmental knowledge, episodic memory for past experiences, and perceptual memory for object recognition and classification. In addition to memory systems, they include control mechanisms, processing modules, learning algorithms, structured data representations, and input/output interfaces. The approach described in Juvina et al. (2018) utilizes the ACT-R cognitive architecture, assigning emotional values (positive or negative) to memory elements. The development of modern conversational agents requires the seamless integration of emotional and cognitive models to achieve human-like interactions.

The E-VOX Perez et al. (2017) system exemplifies this by integrating emotional and cognitive models, combining a SOAR-based framework with an ALMA affective model Flavian-Blanco et al. (2011). It enables real-time emotional processing, ongoing mood tracking, and stable personality representation, leading to improvements in knowledge retrieval, adaptive learning, and emotional intelligence. The Emotion and Adaptation Model (EMA) Gratch & Marsella (2004)involves both fast (automatic) and slow (deliberate) appraisal processes. It systematically analyzes perceived events based on the virtual human's goals and beliefs, generating appraisal frames that help interpret environmental features. These frames connect appraisal variables to specific emotions, which in turn determine the agent's current emotional state.

Cognitive frameworks explore how beliefs, desires, and intentions interact to demonstrate the reasoning processes in agent architectures, as seen in the belief-desire-intention (BDI) model. While these frameworks primarily focus on high-level decision-making structures, cognitive architectures offer a more comprehensive range of cognitive capabilities. This includes reinforcement learning mechanisms, specialized memory systems (procedural, semantic, episodic, and perceptual), advanced data representations, processing components, and control mechanisms—features that cognitive frameworks often lack or implement only partially. Emotional-BDIPereira et al. (2005a) is a conceptual framework that merges emotional components with the traditional BDI cognitive architecture, focusing on the management of resources, capabilities, and emotional states. While it primarily emphasizes the emotion of fear, it also lays the groundwork for future research into emotion-driven behavior in agents. Fuzzy-based affective BDI Taverner et al. (2021) is a computational model that incorporates cultural and linguistic differences into emotion modeling. It consists of two primary components: the event appraisal process, which evaluates stimuli using fuzzy appraisal rules to generate fuzzy emotions, and the affect-adaptation process, which converts these fuzzy emotions into PAD (Pleasure, Arousal, Dominance) dimensions through defuzzification. ABC-EBDI Sanchez et al. (2019a;c) is an emotion model designed for BDI agents that connects beliefs to emotional and behavioral outcomes through the ABC theory. It takes into account mood and personality to produce more realistic behavior.

Some methods develop cognitive components selectively, based on their specific requirements. The Integrative Framework (InFra) Rodriguez et al. (2016) outlines the essential components required for generating emotions. It includes input and output interfaces that connect emotional processes with cognitive modules. InFra takes environmental information and processes it through perception, then conducts an emotional appraisal of stimuli. The resulting emotions subsequently influence behavior. Additionally, the framework considers the effects of personality and culture on the appraisal process, leading to more realistic emotional responses. Silicon Coppelia Hoorn et al. (2021) is designed to respond emotionally to users. It evaluates perceived features using appraisal variables such as ethics, aesthetics, epistemics, and affordances. Ethics guide the agent's moral behavior, while aesthetics relate to its social appearance. Epistemics assess the realism of perceptions, and affordances indicate possible actions in the current context. The agent determines the relevance and positivity or negativity of observations based on its beliefs. To handle multiple emotions occurring simultaneously, Silicon Coppelia employs fuzzy sets and operators. This approach allows it to manage ambiguous emotions and feelings, enabling it to make human-like decisions during interactions. The EIAECZall & Kangavari (2024) framework is designed to develop emotion-aware intelligent agents by utilizing appraisal theories to determine the agents' emotional states in different situations. It features an efficient episodic memory that stores events and their contexts, allowing the agent to recall relevant information for emotional expression and decision-making. The framework learns the affective values associated with events based on the agent's experiences in various contexts, extracting appraisal variables from memory metadata. It also employs a data-driven approach to map these variables to emotional states. Furthermore, a method is implemented to update action activation levels based on the agent's emotional states, effectively modeling how emotions influence decision-making.

Scalability Most existing emotion models are explicitly designed for specific scenarios or contexts Hernández-Marcos & Ros (2024). For example, the MAMID model Hudlicka (2004) is exclusively tailored to treatment settings, relying on specialized if–then rules that render its action evaluation and selection highly domain-dependent. Similarly, WASABI Becker-Asano (2008) has been primarily validated in virtual character gaming environments, which limits its applicability to broader use cases. More generally, many computational emotion models concentrate on only one or two components of the emotional process, depending on the target application. Due to limited scalability and difficulties in seamless integration, extending these models with additional components remains challenging. In contrast, the EIAEC framework Zall & Kangavari (2024) supports scalability across diverse scenarios and applications by leveraging multiple memory types enriched with knowledge graphs. Rather than relying on predefined rules or fixed relationships, this framework enables the continuous learning and refinement of emotional relationships and experiences over time. ***Emotion mapping modeling*** An emotion model implanted in appraisal theory aims to map event assessments (appraisals) into corresponding emotion intensity levels, as specified by the theoretical framework. However, most appraisal theories do not provide explicit, quantifiable rules for mapping these appraisals to emotion intensities. Consequently, many computational implementations adopt heuristic or ad-hoc methods to approximate this relationship, primarily to facilitate research and experimental validation Ojha et al. (2021). For instance, EEGS Ojha et al. (2020b) introduces a mathematical formula for calculating emotion intensities based on appraisal theory, employing a weighted mapping strategy that utilizes quantitative appraisal variables to derive implications for various emotions. Similarly, Somarathna & Mohammadi (2024) leverages the Component Process Model (CPM), which encompasses appraisal, expression, motivation, physiology, and feeling components, to explore their interrelations with emotion elicitation. This study adopts a data-driven approach using interactive virtual reality (VR) games and multi-modal data collection, including self-reports, physiological signals, and facial expressions, to establish a dataset that delineates the relationship between appraisal variables and emotional states. Furthermore, Zall & Kangavari (2024) presents an advanced mapping methodology grounded in datasets from Mohammadi & Vuilleumier (2022) and Somarathna et al. (2023), employing a data-driven framework to refine the mapping between appraisal processes and emotional responses.
Path planning (PP) in autonomous systems can significantly benefit from reinforcement learning (RL) due to its flexible decision-making capabilities. Soman et al. (2024) has introduced emotion-inspired mechanisms, particularly regret-based models, to enhance the performance of RL. These models incorporate two key concepts: experienced regret, which measures past suboptimal actions, and anticipated regret, which predicts future errors. By utilizing regret dynamics to adjust the epsilon decay in epsilon-greedy policies, agents can achieve a better balance between exploration and exploitation. This leads to improved learning efficiency.

Emotion experiences modeling Developing emotional intelligence in an intelligent agent involves more than merely modeling emotional elicitation or artificial emotion synthesis. A fundamental and critical challenge in this domain is modeling emotional experience, specifically understanding how emotions influence various cognitive processes, such as decision-making Hernández-Marcos & Ros (2024). The invention of personalized experiences in intelligent agents improves communication and fosters authentic connections with humans. When an intelligent agent experiences emotions during the process of emotional elicitation, these emotions have impacts on various cognitive components, including inference, reasoning, and decision-making. Self-dialogue Utility in Simulating Artificial Emotions (SUSAN) Corvaia et al. (2025)

proposes a novel perspective for modeling emotional experiences. Specifically, it aims to emulate human-like emotional experience through the use of inner speech modeling. By leveraging inner speech, the model infers and simulates underlying contextual and cognitive processes that contribute to emotional experiences. The framework is grounded in Damasio's theory, providing a structured basis for understanding and replicating emotional dynamics through internal cognitive mechanisms. Mature emotional responses that help regulate behavior offer a significant evolutionary advantage. However, a clear theoretical foundation in cognitive neuroscience is lacking, particularly regarding how emotions are elicited. Hernández-Marcos & Ros (2024) suggests that emotions correspond to specific temporal patterns in essential environmental variables and introduces a self-learning emotional framework for intelligent agents. An unsupervised neural network trained on unlabeled experiential data to identify eight basic emotional patterns that are contextually coherent and mimic natural emotional dynamics. EIAEC Zall & Kangavari (2024) learns context-dependent affective values by associating events with the agent's emotional experiences across different situations, storing these associations in episodic memory. Additionally, it dynamically updates activation values for actions in procedural memory based on the agent's emotional state, effectively modeling the influence of emotion on decision-making processes. The approach outlined in Juvina et al. (2018) learns to associate emotional valuations and arousal levels with each declarative memory chunk, taking into account their frequency of use and reward signals derived from core affect dynamics. Ultimately, the model assesses how emotional value and arousal levels affect memory retrieval and decision-making processes. The relationship between emotional experiences and their memorability has long been acknowledged, with highly emotional events generally considered to be more memorable. This relationship suggests that emotional annotations could potentially serve as proxies for assessing memorability. However, most existing emotion recognition systems depend heavily on third-party annotations, which may not accurately reflect individuals' firsthand emotional relevance and memory encoding. Recognizing this gap, Tsfasman et al. (2025) empirically explores the connection between perceived emotions and collective memorability within conversational interactions. It involves continuous, time-based annotations of both emotional states and memorability in dynamic, unstructured group settings, aiming to approximate real-world conversational AI environments.

Recent research has explored the integration of emotional and cognitive variables into reinforcement learning (RL) agents, frequently using these variables as reward to enhance decision-making processes Moerland et al. (2018); Alkam et al. (2025). Building upon this foundation, Zhang et al. (2024d) introduced a computational model of emotion that synthesizes RL with appraisal theory, establishing a formal link between reward processing, goal-directed learning, cognitive appraisal, and emotional experience. The model formalizes four appraisal variables, such as suddenness, goal relevance, goal conduciveness, and power, derived from the component process model (CPM), and operationalizes them via temporal difference learning updates. Remarkably, this framework is task-independent and applicable to any environment modeled as a Markov decision process (MDP), allowing the analysis and simulation of emotional responses within RL-based decision-making contexts. Complementing this approach, Zhang et al. (2024b) proposed a computational cognitive model that conceptualizes emotion as a dynamic, continuous process rather than a static state, particularly during interactive decision episodes. Their framework integrates cognitive theories of emotion with principles of computational rationality, using RL for value prediction to capture the evolving nature of emotional experiences during ongoing interactions. Agent-based multi-issue bilateral automated negotiation is gaining increased attention from researchers. However, the impact of emotional deception and multi-attribute preferences on negotiation outcomes is often neglected. Wu & Sun (2025) develops a negotiation model based on agents and employs reinforcement learning. It quantifies emotions and generates emotional deception using the Weber-Fechner law in automated negotiation. Furthermore, emotions are used to adjust the reward function of Q-learning to update opponents' attribute preferences. A time-dependent issue updating model is proposed, which integrates emotional deception, attribute preferences, and a fairness function. Berto et al. (2025) presents a cognitive architecture that simulates human cognitive functions, including perception, motivation, and decision-making. The intelligent agent operates based on motivations derived from internal states, which are influenced by affective states.

Self-control, a fundamental aspect of human decision-making, is often understood as the internal conflict between higher-order executive functions and lower limbic systems, especially in situations involving the dilemma between smaller, immediate rewards and larger, delayed rewards. Previous computational models have employed game-theoretic frameworks, such as the Iterated Prisoner's Dilemma, with learning mechanisms to explore decision dynamics related to self-control. However, these models generally do not explicitly

account for the influence of emotional states. Nikodemou & Christodoulou (2024) incorporates emotion as a modulatory factor by adjusting reinforcement rewards, thereby simulating temporal fluctuations in positive and negative emotional intensities. This approach captures the effects of emotions on self-control by modeling their dynamic influence, rather than treating emotions as static or isolated states.
Transparency in behavior is crucial for robots that interact with humans. When robots need to adapt to different users and tasks, they must optimize their behavior, and Reinforcement Learning offers a promising method for achieving this. However, the behaviors generated by RL are not inherently transparent due to the exploration/exploitation tradeoff involved in optimizing policies. Emotions serve as a natural way for humans to communicate intent and evaluate situational relevance. Incorporating emotional expressions into robots has been suggested as a way to enhance the transparency of their learning processes. In this regard, some studies Nijeholt & Broekens (2023) have implemented emotional expressions based on Temporal Differences (TD) to make the robot's learning trajectory more interpretable to human observers and instructors.

Evaluation-related A significant challenge in developing agents with emotional intelligence is evaluating their computational models of emotion. Typically, assessments rely on subjective measures within specified scenarios. In such cases, if the agent's behavior aligns with expected responses, the underlying model is regarded as effective. When evaluating the emotional behavior of cognitive agents, two fundamental criteria can be considered: believability and social acceptabilityOjha et al. (2020a). Believability assesses how natural and contextually appropriate the agent's actions appear, contributing to a sense of lifelikeness. While believable emotions are essential for an agent's acceptance, they alone may not guarantee social approval. An agent can appear lifelike but still fail to adhere to social norms. Therefore, social acceptability assesses whether the emotions expressed by the agent align with societal expectations and standards, ensuring that the agent's behavior is not only believable but also socially appropriate Zall & Kangavari (2022).
Each of the reviewed approaches has extended the cognitive components necessary for emotional elicitation from various perspectives. Consequently, a cognitive component based on a similar conceptual framework is not consistently evident across these models. This presents significant challenges for the development of a novel approach, particularly given that detailed implementation specifics of these architectures are often not readily accessible.
To assess the effectiveness of computational models of emotion, some methods involve simulating an emotional scenario and comparing the reactions of the virtual agent to those of a human in a similar situation Sanchez et al. (2019b); Zall & Kangavari (2024). EBDI Jiang et al. (2007) models a multi-agent environment capable of evaluating the performance of various agent architectures across different contexts by adjusting multiple parameters. It focuses on decision-making in agents with diverse emotional capabilities to analyze the influence of emotions on making appropriate choices. Similarly, Bourgais et al. (2016) simulates the behavior of multiple agents that incorporate emotions into their reasoning processes during critical situations, highlighting the role of emotion in adaptive decision-making. E-VOX Perez et al. (2017) is evaluated based on human feedback, where users interacting with the E-VOX-based agent assess its believability and acceptability by completing several questionnaires.

4.2.3 LLMs-Related Challenges

The emergence of LLMs has paved the way for developing emotionally intelligent agents that leverage these technologies. In response, the development of LLMs is revolutionizing agents, transforming them from mere instrumental tools into autonomous "Metaverse Citizens" equipped with decision-making and interaction abilities. The recent study Khan et al. (2025) demonstrates how intelligent agents can integrate emotions into their decision-making processes, resulting in behaviors that are both acceptable and satisfying to users. This research examines the role of LLMs in creating intelligent agents within the Metaverse service ecosystem, where digital avatars, digital twins, and digital natives play essential roles. To improve the reliability and social realism of these agents, the paper proposes an explainable emotion alignment framework that incorporates factual data into their decision-making processes. The paper Raggioli et al. (2025) introduces an integrated architecture that combines cognitive and affective reasoning using LLMs. In this framework, LLMs generate beliefs from natural language inputs. It employs an intrinsically motivated approach that

relies on an internal emotional state to influence internal rewards.
Tak & Gratch (2024) builds on recent studies of the emotional reasoning capabilities of LLMs. While prior work has assessed LLMs' general understanding of emotions, it has not distinguished between their predictions of self-attribution versus perception of others' emotions, based on appraisal theory, which reflects the fundamental mechanisms involved in human emotion elicitation. It reveals that GPT-4 excels in reasoning about stimuli designed to evoke inferred emotional attributions, closely mirroring human judgments, especially in idiosyncratic scenarios. Conversely, GPT-4's interpretations align more with perceptions of others' emotions than with self-assessments, supporting the idea that LLMs adopt an observer-like perspective. The paper Gandhi et al. (2024) assesses affective cognition in foundation models to specify the abilities of LLMs such as GPT-4, Claude-3, and Gemini-1.5-Pro in determining emotion elicitation in various scenarios. It manifests that these models constantly align with human judgments, occasionally exceeding average human accuracy, particularly with chain-of-thought reasoning. Recent studies have moved beyond surface-level emotion recognition to examine the cognitive mechanisms underlying emotional reasoning in large language models (LLMs). In particular, Bhattacharyya et al. (2025) introduces CoRE, a large-scale benchmark grounded in cognitive appraisal theory, to evaluate whether LLMs exhibit coherent and interpretable cognitive reasoning when processing emotionally charged stimuli. This paper indicates that while many models demonstrate broadly human-like appraisal structures, they struggle with nuanced emotions and exhibit inconsistencies across abstract cognitive dimensions, revealing implicit biases in emotion representation and limitations of current training paradigms.

5 Emotional Text Synthesis

ETS in NLP focuses on producing emotionally resonant text through advanced techniques such as cross-modal emotion generation and emotion-controlled language models. The process involves creating written or spoken content that conveys specific emotional states, enabling machines to produce emotionally resonant and contextually appropriate responses. It employs techniques such as style transfer, conditional generation, and fine-tuning of LLMs. This capability is a cornerstone of affective computing, which seeks to enhance HCI by imbuing systems with emotional intelligence Schuller & Schuller (2018). Emotion generation plays a pivotal role in applications such as virtual assistants Xue et al. (2024), mental health support systems Feng (2024), and personalized content creation Abilbekov et al. (2024), where the ability to express and evoke emotions significantly impacts user engagement and satisfaction Becker et al. (2024). For instance, virtual assistants equipped with emotion generation capabilities can provide more empathetic and context-aware responses, fostering a more natural and human-like interaction Zheng et al. (2021). In NLP, techniques like sentiment analysis and emotion detection lay the groundwork for generating emotionally charged text by identifying and interpreting emotional cues in existing data Truong (2024). Advances in AI, particularly in generative models and LLMs, have further propelled the field by enabling fine-grained control over emotional expression in generated text Becker et al. (2024). Here, we first briefly introduce approaches to generate emotional text. Then, we proceed with the challenges within this area and how different research studies have tried to overcome them.

5.1 Approaches

***Style Transfer*:** Style transfer alters the emotional tone of a text while maintaining its meaning, which is beneficial for personalized content creation. However, it often struggles with maintaining semantic coherence and emotional consistency in more complex texts. Innovations such as the lexicon-based attention mechanism and methods separating content from style Fu et al. (2018) have been developed to improve emotional nuance and address the lack of parallel corpora.
***Conditional Generation*:** Conditional generation controls emotional content through specific prompts, enhancing applications like virtual assistants. Models such as the co-attention neural network Li et al. (2018a) and Emotional Tacotron Lee et al. (2017) exemplify this approach, though they require high-quality datasets and struggle with multi-emotion scenarios. Affective Chatbot Jiang et al. (2022) and emotion prediction in TTS systems Yoon et al. (2022) highlight its versatility.
***Fine-tuning LLMs*:** Leveraging models like GPTs and BERT enables high-quality emotional text gener-

Figure 10: Challenges in ETS

ation with minimal training Singh et al. (2020). This method requires significant computational resources and careful bias management.

5.2 Challenges

In affective text generation, defining and representing emotions presents significant challenges, crucial for accurately conveying emotional states and their contextual expressions. Accurately capturing and generating emotions requires a deep understanding of emotional states and their contexts. Foundational models like Ekman's six basic emotions Ekman (1992) and Plutchik's wheel of eight emotions Plutchik (1980) provide structured frameworks for categorizing emotions, highlighting their interconnected nature Zhang et al. (2024a). This section explores these complexities, emphasizing their foundational role in emotion generation

Table 4: Summary of challenges and solutions in emotional text synthesis

Challenge	Sub-challenge	Solutions
Data-Related	Scarcity	Data augmentation (Firdaus et al., 2023); Transfer learning (Li et al., 2018b); SentiGAN (Wang & Wan, 2018); CS-GAN (Li et al., 2018b); LeakGAN (Guo et al., 2018)
	Limited Annotation	Desire annotation (Jia et al., 2022)
Model-Related	Emotional Consistency & Coherence	Appraisal theories (Xia & Ding, 2019; Singh et al., 2020; Resendiz & Klinger, 2023); Emotion embeddings (Tan et al., 2023); MSEG (Firdaus et al., 2023); SEPRG (Firdaus et al., 2021); RCVAE (Zhou & Wang, 2017); DecoupledESC (Zhang et al., 2025a); CARE (Zhu et al., 2025); COMPEER (Wang et al., 2025b)
	Handling Multi-Emotion	Emotion blending (Ghosh et al., 2017; Resendiz & Klinger, 2023); ECM (Zhou et al., 2018)
	Precise Emotion Control	FUDGE (Yang & Klein, 2021); MOAEP (Resendiz & Klinger, 2024); Prefix-tuning (Qian et al., 2022); DEXPERTS (Mahmood & Manning, 2023)
Cross-Modal	Alignment	Hierarchical attention (Zhang et al., 2017); Word-level alignment (Zhang et al., 2017)
	Fusion	Time-dependent fusion (Zhang et al., 2017); FaceChat (Alnuhait et al., 2023); Omni-perception Pre-Trainer (Kaur et al., 2024); MAGIC (Su et al., 2022); emoTTS (Luo et al., 2024); FIRES (Authors, 2025)
LLM-based	Controllability & Adaptability	Coda (Evuru et al., 2024); MOPO (Resendiz & Klinger, 2024); Emotion Vectors (Dong et al., 2025)
	Bias	Debiasing techniques (Sheng et al., 2021)
	Computational Cost	Affective prompt-tuning (Gu & He, 2024)
	Evaluation	EmoBench (Sabour et al., 2024); Kardia-R1 (Yuan et al., 2025)

and how different studies have tried to overcome these challenges. Main challenges in this area and their related sub-challenges are depicted in Figure 10. Moreover, Table 4 highlights different studies and their solutions to these issues.

5.2.1 Data-Related Challenges

The performance of affective text generation models is heavily dependent on the availability of large, high-quality, and culturally diverse datasets. The scarcity of annotated datasets limits a model's ability to replicate nuanced emotional states, which in turn affects the accuracy of the generated text Li et al. (2023a). This is particularly evident in the lack of annotations for emotions like desire Jia et al. (2022). To mitigate this, strategies such as data augmentation and transfer learning are used to enhance the diversity of training data and improve model generalization Li et al. (2018b); Firdaus et al. (2023). GANs have been used to address data scarcity. SentiGAN Wang & Wan (2018) and CS-GAN Li et al. (2018b), for example, enhance emotional diversity in generated text, and LeakGAN Guo et al. (2018) tackles challenges related to discrete outputs and provides continuous guidance.

5.2.2 Model-Related Challenges

Generating text that captures multiple emotions, emotional transitions, and maintains coherence is a complex task. Emotion blending and sequential emotion modeling are key approaches to integrate multiple emotions into a coherent expression Ghosh et al. (2017). Despite progress, achieving fine-grained control over emotional expressions remains a significant challenge Zhang et al. (2023b). This is particularly difficult for non-native speakers who often struggle with expressing mixed emotions, highlighting the need for models that can replicate these nuanced expressions Feng (2024). The Emotional Chatting Machine Zhou et al. (2018) addresses coherence with specified emotions through the use of emotion embeddings and attention mechanisms. For precise emotion control, techniques like FUDGE Yang & Klein (2021) have been developed to offer fine-grained control over emotional intensity and category. Advanced model architectures like prefix-tuning Qian et al. (2022) and DEXPERTS Mahmood & Manning (2023) have been proposed to address challenges like mode collapse. The SemEvalMohammad et al. (2025) shared task on bridging the gap in text-based emotion detection highlights the importance of developing culturally aware and context-sensitive models that can accurately interpret and generate emotions across different languages and cultures. Modifying generative models to incorporate new emotional attributes without compromising the integrity of the content is a major obstacle, exacerbated by the non-differentiable nature of discrete text Wang & Wan (2018). Furthermore, cultural and linguistic diversity introduces additional complexity, necessitating advanced frameworks that merge linguistic and psychological theories with machine learning techniques Jain et al. (2017). The absence of non-verbal cues in text poses another layer of difficulty, often leading to misinterpretations, especially among non-native speakers. Diffusion and flow-based models are also at the forefront of generating nuanced emotional text. Diffusion models, such as Diffusion-LM Li et al. (2022b), iteratively transform noise into structured data, allowing for precise emotional control. Flow-based models use invertible transformations for high-dimensional data generation, capturing emotional nuances while maintaining linguistic quality. These models can integrate affective parameters, enhancing the expressiveness of conversational language Ghosh et al. (2017). DVAE-interVA Chen et al. (2023a) is proposed to solve the emotion-semantic entanglement problem in continuous sentiment text generation via adversarial sentiment decoupling, continuous sentiment embeddings, denoising training, and interactive attention to maintain emotional consistency throughout generation. Another challenge is maintaining emotional consistency and coherence throughout a piece of text or a conversation. This requires models to sustain a consistent emotional tone and narrative. Integrating nuanced understanding, such as appraisal theories, is one approach to enable contextually relevant emotional expression Truong (2024); Xia & Ding (2019); Resendiz & Klinger (2023). However, the variability of language and the informal nature of social media continue to make emotional consistency a complex problem Tan et al. (2023). Context understanding is critical, as many chatbots struggle to maintain coherent conversations and lose track of the emotional context Shum et al. (2018). To address this, methods like MSEG Firdaus et al. (2023) have been developed to generate responses that align with identified emotions. The SEPRG model Firdaus et al. (2021), for instance, aims to maintain emotional connections while adhering to a consistent persona. Disentangled representations are vital for isolating emotion-specific features, enabling precise emotional expression while maintaining semantic coherence. This is particularly crucial for applications like virtual assistants and mental health support systems Truong (2024). The integration of these representations with LLMs, like the MOPO method, improves emotion control and prediction accuracy Resendiz & Klinger (2024). The RCVAE model demonstrates the effectiveness of disentangled representations in generating emotionally rich responses from tweets and emojis Zhou & Wang (2017). A key challenge in emotional support conversation is that supervised fine-tuning of large language models produces psychologically inconsistent responses, and applying direct preference optimization is hindered by the entanglement of psychological strategies and response content in existing emotional support conversation data, leading to ambiguous optimization objectives. DecoupledESC Zhang et al. (2025a) addresses this by decomposing the emotional support conversation task into two sequential subtasks, strategy planning and empathic response generation, inspired by Gross's Extended Process Model of Emotion Regulation. An Inferential Preference Mining method constructs high-quality preference pairs for each subtask, and each is independently trained via supervised fine-tuning followed by direct preference optimization. This decoupled approach reduces preference bias and improves both strategic soundness and emotional appropriateness of generated responses. Current emotional support conversation approaches often overlook the deeper cognitive reasoning processes that underpin effective emotional support, focusing instead on data augmentation and synthetic corpus construction. CARE Zhu et al. (2025) proposes a framework that strengthens cognitive reasoning in emotional support conversation without relying on large-scale synthetic data. It leverages the original emotional support conversation training set to guide models in generating logically coherent and supportive responses, then employs reinforcement learning to refine and reinforce the reasoning process. Experimental results demonstrate significant improvements in both the logical soundness and supportive quality of generated responses. Generating emotionally supportive text that is both empathetically grounded and non-repetitive remains challenging, as current models lack deep empathetic reasoning rooted in psychological principles, and reinforcement learning-based training often suffers from entropy collapse leading to repetitive outputs. COMPEER Wang et al. (2025b) proposes controllable empathetic reasoning that combines natural language reasoning with structured psychological steps. It constructs a fine-grained dataset annotated with reasoning correctness and response preferences, and employs reinforcement learning with a unified process-outcome reward model for precise feedback. To mitigate repetitiveness, COMPEER introduces personality-based dialogue rewriting and a redundancy-aware reward reweighting strategy.

5.2.3 Multimodal-Related Challenges

Cross-modal learning enhances emotional accuracy by integrating text with other modalities like speech and facial expressions. Since this type of challenge is associated with the following sections in emotional content synthesis, here some challenges are underscored briefly and meticulous information will be delivered there. A key challenge in this area is alignment, which involves synchronizing information from different sources, such as aligning visual cues with corresponding text segments. This is particularly difficult when the relationship between modalities is not explicit, as is often the case with text that only describes a video or audio track in general terms Zhang et al. (2017). Another significant challenge is fusion, which refers to the method of combining features from heterogeneous data sources. Many early models fused modalities only at an abstract level, which fails to capture the time-dependent interactions between them Zhang et al. (2017). Addressing these challenges is crucial for creating seamless and contextually aware affective experiences. Systems like FaceChat Alnuhait et al. (2023) have demonstrated real-time improvements in emotional accuracy by combining speech recognition, NLP, and facial analysis. The Omni-perception Pre-Trainer Kaur et al. (2024) integrates emotional cues from multiple modalities to enhance the quality of generated text. The MAGIC model highlights the potential of cross-modal learning to improve emotional accuracy and adapt text to diverse user needs Su et al. (2022). The emoTTS model Luo et al. (2024) is another example of a cross-modal application that enhances emotion control and generation accuracy. Chain-of-thought based emotional support conversation methods typically employ rigid, text-only reasoning, which limits adaptability in dynamic multimodal interactions and introduces reasoning noise that degrades support quality. FIRES Authors (2025) introduces “Flexible Thinking” for multimodal emotional support conversation, enabling models to adaptively select contextually relevant reasoning aspects, including visual scene, emotion, situation, and response strategy. The framework integrates supervised fine-tuning for initial learning with reinforcement learning for refinement, directly linking thinking processes to response quality via tailored rewards. Experiments on the MESC and EMOTyDA datasets demonstrate improved quality and generalizability of emotional support responses through this adaptive multimodal reasoning approach.

***LLM-Based Challenges*:** Recent research has also focused on leveraging LLMs and developing novel techniques to enhance emotional expression while maintaining semantic coherence. The high computational cost associated with fine-tuning large models for emotional text generation is of importance. APT-LM Gu & He (2024), a parameter-efficient solution, overcomes computational inefficiency in emotional text generation by freezing pre-trained language models and using minimal affective parameters through prompt-tuning, enhanced by affective decoding that systematically strengthens emotional expression at multiple linguistic levels while preserving fluency. Another key challenge is the evaluation of emotional intelligence in LLMs. While models are becoming more sophisticated, their ability to understand and apply emotional intelligence is still limited. To this end, benchmarks like EmoBench Sabour et al. (2024) have been developed to assess the emotional intelligence of LLMs through a series of hand-crafted multiple-choice questions that cover both emotional understanding and application. This benchmark has revealed that even the most advanced LLMs still have a considerable gap to bridge to reach human-level emotional intelligence. Multi-Objective Prompt Optimization (MOPO) Resendiz & Klinger (2024), a method that optimizes prompts for LLMs to generate text meeting multiple objectives simultaneously, specifically transforms the desired emotion and fits the stylistic requirements of different domains. MOPO employs a three-layer optimization process: Layer-1 generates initial text prompts, Layer-2 refines these by paraphrasing or combining them using genetic operations, and Layer-3 provides fixed prompts to guide the optimization and improves performance by up to 2% for individual objectives while providing balanced, flexible solutions, enhancing the adaptability and effectiveness of affective text generation across diverse applications. Generating affective text that conveys specific emotions with controlled intensity and topic relevance, while preserving grammatical accuracy, has been addressed by enhancing the GPT-2 model and integrating emotional priors and applying a gradient descent-based perturbation technique Singh et al. (2020). CoDa Evuru et al. (2024) introduces a training-free data augmentation framework for low-resource NLP that extracts simple heuristic-based constraints from limited training data and verbalizes them to prompt off-the-shelf instruction-tuned LLMs, achieving controlled generation that balances diversity and consistency while outperforming existing methods across 11 datasets. While large language models exhibit strong reasoning capabilities, they struggle to express emotions in a consistent, controllable, and contextually appropriate manner, and existing prompt-based or fine-tuning-based methods lack flexibility or require costly retraining. The Emotion Vector framework Dong et al. (2025) addresses this by extracting latent representations from internal activation shifts between neutral and emotion-conditioned responses. By injecting these vectors into the hidden states of pretrained large language models during inference, the method enables fine-grained, continuous modulation of emotional tone and intensity without any additional training or architectural modification, while preserving semantic fidelity and linguistic fluency. Existing emotional support conversation systems rely on situation-centric datasets that lack persistent user identity, which limits the capture of personalized affective nuances. Moreover, opaque and coarse reward signals hinder the development of verifiable empathetic reasoning in large language models. Kardia-R1 Yuan et al. (2025) introduces KardiaBench, a large-scale user-grounded benchmark comprising 178,080 question-answer pairs anchored to 671 real-world user profiles, and proposes Rubric-as-Judge Empathetic Reinforcement Learning. This group relative policy optimization-based method uses explainable, human-aligned rubric rewards that tightly couple user understanding, emotional inference, and supportive response generation, enabling interpretable stepwise empathetic cognition with consistent improvements in emotion accuracy, empathy, and persona consistency across multiple large language model backbones. Finally, another hurdle is modifying generative models to incorporate new emotional attributes without compromising the integrity of the content, a problem exacerbated by the non-differentiable nature of discrete text Wang & Wan (2018). The community is also addressing the ethical implications of affective computing, particularly the issue of bias. As models become more capable of generating emotional text, there is a growing need to ensure that they are used responsibly and do not perpetuate biases present in the training data Sheng et al. (2021). Integration of disentangled representations with LLMs improves emotion control and prediction accuracy.

6 Emotional Speech Synthesis

Emotional speech generation is a key area in the field of AI and HCI, aimed at enabling machines to produce emotionally natural, meaningful speech. This process involves transferring the intended emotions in the speech generation process. In recent years, advancements in machine learning, NLP, and signal processing have propelled speech generation technologies forward, enabling virtual assistants, chatbots, and social robots to engage with users in more natural and fluid ways. Emotional speech generation can be categorized into two primary subfields: Emotional voice Conversion and ESS. These areas, while related, serve different purposes and are implemented through distinct technologies and techniques. Even though TTS technology has achieved mature and reliable outcomes, ESS still faces significant challenges. Developing robust EVC systems capable of performing any-to-any emotion conversion could address some of the key issues related to ESS in a two-stage manner and pave the way for substantial progress in this field. Inspired by Triantafyllopoulos et al. (2023), we present a typical ESS workflow that leverages EVC. Initially, based on the stimuli and the given context, an appropriate response is generated, along with the recognition of the desired emotion to be imbued. A TTS system then generates the spectrum of the target speech in a neutral format. Finally, the EVC component processes the input speech and converts it to the desired emotional

Figure 11: Challenges in ESS

output. During the generation and utterance of emotional speech, capturing feedback and stimuli from the environment is crucial. This is because the state of the situation may change, requiring adjustments to both the emotion being expressed and the context of the speech. In the following, the first approaches for emotional voice generation are discussed. Then, we dive into the challenges in the realm and propose solutions for them.

6.1 Approaches

***ESS*:** ESS or Affective Speech Synthesis focuses on generating speech imbued with emotional expressivity, enabling machines to produce human-like, emotionally rich communicationTriantafyllopoulos et al. (2023). ESS primarily relies on text-to-speech (TTS) technology and plays a crucial role in enhancing HCI and

enriching audio broadcast scenariosLei et al. (2022). Conversational Emotional Speech Synthesis (CESS) builds upon ESS by incorporating the ability to maintain context from previous interactions. While ESS focuses on generating speech with emotional expressiveness, CESS enhances this by enabling the system to remember and reference past conversations. This memory capability distinguishes CESS from ESS, which typically lacks such continuity.
***Emotional Voice Conversion*:** Emotional voice conversion (EVC) involves transforming the emotional expression in speech from a source emotion to a target emotion while preserving both the linguistic content and the speaker identity Zhou et al. (2022b). In contrast, voice conversion (VC) focuses on modifying one's voice to sound like another's without changing the linguistic content Sisman et al. (2020). EVC involves altering specific acoustic features, such as pitch, timbre, and formant frequencies, to reflect the target emotion while keeping the original speech intact. This process can be used in various applications, including virtual assistantsElgaar et al. (2020), Human-Robot InteractionCrumpton & Bethel (2016), and HCIPittermann et al. (2010). Early approaches to EVC relied on statistical models such as Gaussian Mixture Models (GMMs)Aihara et al. (2012), which modeled the relationship between source and target features, and techniques like frequency warpingSheikhan et al. (2012), which directly modified spectral features to reflect emotional changes. While effective for their time, these methods were limited by their reliance on parallel data and their inability to capture complex, non-linear transformations. With advancements in machine learning, modern approaches have shifted towards deep learning techniquesWalczyna & Piotrowski (2023), which enable more flexible and robust emotional conversions. Encoder-decoder frameworks have been widely adopted to disentangle speaker identity, linguistic content, and emotional features, facilitating more effective style transferZhou et al. (2021a).

6.2 Challenges

In this section, we investigate the challenges associated with emotional speech generation from diverse standpoints. Consequently, we review the studies on these challenges and their proposed solutions. As illustrated in Figure 11, we categorize these challenges into four main domains: data-related challenges, model-related challenges, challenges inherent to emotions, and challenges related to LLMs. Table 5 illustrates these categories, detailing their associated sub-issues and highlighting studies attempting to mitigate these challenges. Following, we will discuss challenges and methods proposed to address them.

6.2.1 Data-Related Challenges

One of the main issues in developing emotional speech generation methods is related to speech datasets. Data-related challenges encompass issues such as the reliance on parallel dataLiu et al. (2020); Zhou et al. (2020b); Prabhu et al. (2024), data scarcitySchnell et al. (2021); Zhou et al. (2022b), monolingual datasetsZhou et al. (2022b), and the need for realistic (not acted), multi-speaker datasetsLotfian & Busso (2019).
There are two major types of EVC according to data: 1)Parallel and 2)Non-Parallel. In the former, methods utilize pairs of utterances that contain the same content from the same speaker but are expressed with different emotions. During training, the conversion model learns to map features from the source to the target emotion using these paired feature vectors. Most EVC systems are implemented in this manner cite zhou2022. In contrast, non-parallel EVC involves learning to map source speech features (e.g., spectral, prosodic) to target emotional features without relying on paired utterances Gao et al. (2018). Parallel data requires both input and output audio recordings for each utterance, which is resource-intensive and time-consuming to collect. Consequently, there is a growing preference for non-parallel data to alleviate the burdens associated with parallel datasets.
Adversarial Generative Network (GAN) models, such as CycleGANZhou et al. (2020a) and StarGANRizos et al. (2020), have proven particularly powerful in non-parallel EVC scenarios by learning mappings between source and target emotional styles without requiring paired data. CycleTransGANFu et al. (2021), a CycleGAN-based model enhanced with transformers for non-parallel EVC, uses transformers to capture temporal intra-relations over wider receptive fields. Also, it adopts curriculum learning, gradually increasing the frame length during training to improve feature comprehension, and incorporates fine-grained discriminators for detailed emotional mapping. In Meftah et al. (2023), the use of the StarGANv2-VC framework Li

et al. (2021) for EVC in English has been explored. Current systems often struggle with the ability to handle multiple speakers and emotions, especially when limited data is available. This research aims to address these gaps by evaluating the StarGANv2-VC model across various configurations, including speaker-dependent, gender-dependent, and gender-independent scenarios. CycleGAN Liu et al. (2020) is a non-parallel EVC model that incorporates two discriminators to differentiate between natural and converted speech, along with a classifier to identify the underlying emotion from both types of speech. Recently, diffusion models have emerged as a promising approach in emotional speech generation Ma et al. (2024), leveraging their ability to model complex data distributions and generate high-quality emotional transformations. EmoConv-Diff Prabhu et al. (2024) presents a diffusion-based model utilizing non-parallel and in-the-wild data. It effectively disentangles lexical content, speaker identity, and emotional information through a diffusion-based decoder, enabling precise emotion transformation via reverse stochastic differential equations.
Current emotional TTS systems struggle to authentically capture human emotions due to reliance on over-simplified emotional labels and single-modality inputs. To solve this, the authors propose UMETTS Li et al. (2025a), featuring (1) EP-Align that uses contrastive learning to align emotional features across text, audio, and visual modalities, and (2) EMI-TTS that integrates these aligned embeddings with state-of-the-art TTS models, resulting in significantly improved emotion accuracy and speech naturalness. EmoCat Schnell et al. (2021), a language-agnostic EVC model designed to convert neutral speech to emotional speech, incorporates gradient inverter Ganin & Lempitsky (2015) blocks to suppress emotional leakage and uses a VAE-based encoder-decoder structure inspired by the CopyCat Karlapati et al. (2020). EmoCat effectively reduces the emotional training data required in the target language by leveraging emotional data from another language. As a result, it achieves high-quality emotional conversion in German, using only 45 minutes of German emotional data while being supported by extensive emotional datasets in US English. EmoSphere++ Cho et al. (2025a) introduces an emotion-adaptive spherical vector (EASV) that represents emotional style through angular position and intensity through radial distance in a spherical coordinate system, enabling fine-grained emotion control without predefined labels. The framework employs a joint attribute style encoder with orthogonality loss to effectively disentangle speaker identity and emotional features, achieving high-quality zero-shot emotional speech synthesis for unseen speakers while eliminating the need for additional discriminators.

6.2.2 Model-Related Challenges

Challenges associated with model structures include difficulties in adapting frameworks for emotional speech generation, defining effective loss functions for various components Liu et al. (2020); Chou et al. (2024), preventing emotion leakage Schnell et al. (2021), and overcoming speaker dependency Zhou et al. (2020b); Choi & Hahn (2021) and achieving disentangled emotion and speaker representations. Moreover, models must be capable of processing long sequences of data and accurately aligning each segment with its corresponding emotional expression Fu et al. (2021). Although these challenges are significant, they are generally considered less critical than those related to data.
Using minimum generation loss as the objective function in speech generation can constrain the learning process, often leading to over-smoothing of the generated speech parameters. In contrast, employing a more sophisticated loss function that incorporates perceptual aspects, diversity, and emotional qualities can enhance the emotional quality of speech generation, resulting in outputs that resonate more authentically with human listeners. Liu et al. (2020) utilizes various losses, including adversarial, cycle consistency, and emotion classification, to effectively learn parameters in the training process. The adversarial loss measures how distinguishable the generated speech is from the true speech. Cycle consistency loss deposits that the input can keep its original form after passing through the two generators. To capture emotional aspects, this method employs an emotion classification loss. Chou et al. (2024) uses a disentangled loss in the diffusion EVC model, which reduces the correlation between different speech representations, particularly emotion information and speaker identity. Additionally, an expressive guidance mechanism enhances emotional expressiveness throughout the reverse diffusion process. Zhou et al. (2022c) leverages perceptual losses in the training process to enhance the intelligibility of the generated emotional speech. It uses a contrastive loss to ensure that the text and audio embeddings are similar, effectively disentangling linguistic and emotional elements. Additionally, by using pre-trained Speech Emotion Recognition (SER), this method

Table 5: Summary of challenges and solutions in emotional speech synthesis

Challenge	Sub-challenge	Solutions
Data-Related	Parallel Data	Non-parallel data with Adversarial Networks (Zhou et al., 2020a; Liu et al., 2020; Rizos et al., 2020); Transformer (Fu et al., 2021); MSP-Podcast dataset (Lotfian & Busso, 2019)
	Scarcity & Cross-linguality	Emotional datasets in other languages (Schnell et al., 2021); ESD dataset (Zhou et al., 2022b); EmoSphere++ EASV (Cho et al., 2025a); UMETTS contrastive learning (Li et al., 2025a)
	Mono Speaker	ESD dataset (Zhou et al., 2022b)
	Acted vs. Natural	MSP-Podcast dataset (Lotfian & Busso, 2019)
Model-Related	Loss Function	Disentangled loss (Chou et al., 2024); Perceptual losses (Zhou et al., 2022c); Mutual information minimization (Yang et al., 2025)
	Speaker Dependency	VAW-GAN encoder-decoder (Zhou et al., 2020b); Seq2seq with attention (Choi & Hahn, 2021); Self-supervised distillation (Cho et al., 2025b)
	Emotion Leakage	Gradient reversal layer (Schnell et al., 2021)
	Long Sequence Processing	Long-sequence processing (Fu et al., 2021); Emotion alignment (Fu et al., 2021)
Emotion-Related	Emotion Generalization	VAW-GAN (Zhou et al., 2021b); Two-stage training (Chen et al., 2023b); Ranking-based SVM (Zhou et al., 2022a); Dual-granularity diffusion (Su et al., 2025)
	Intensity & Control	Continuous arousal scales (Prabhu et al., 2024); Two-stage training (Chen et al., 2023b); Expressive guidance (Chou et al., 2024); Seq2seq attention (Choi & Hahn, 2021); Activation steering (Xie et al., 2025)
LLM-Related	Duration Flexibility	Expressive guidance for enhanced emotional diffusion (Oh et al., 2024)
	Noisy Output	Emotion/intensity encoders (Zhang et al., 2025c); Prosody adjustment (Zhang et al., 2025c)
	Human Expectations	ECoT prompting (Cai & Li, 2024)
	Computational Overhead	Efficient model design (Mishra et al., 2023)

predicts the emotion category of the generated speech and calculates the emotion classification loss at the utterance level.

Emotion leakage occurs when emotional intensity is weak during conversion, leading to a significant decline in signal quality and intelligibility. EmoCat Schnell et al. (2021) addresses this issue by incorporating a gradient reversal block before the emotion classifier during training Ganin & Lempitsky (2015). This approach reverses the gradients during backpropagation to eliminate any input activation that could aid the emotion classifier, thereby reducing emotional leakage.

(ProEmo) Zhang et al. (2025c) generates synthesized speech with refined emotional expressiveness and controlled intensity, which conventional TTS systems often lack. It extends the FastSpeech 2 framework Ren et al. (2020) by incorporating an emotion encoder and an intensity encoder. The emotion encoder employs a fine-tuned HuBERT transformer Hsu et al. (2021) with a classification head to derive robust emotion embeddings from the waveform, while the intensity encoder, also based on HuBERT and equipped with a regression head, estimates continuous emotion intensity through a speaker- and emotion-specific ranking function applied to acoustic features. During inference, GPT-4 OpenAI (2023) produces global and local scaling factors that modulate the FS2-predicted duration, energy, and pitch via multiplicative and additive adjustments, using a quadratic mapping function to yield speech with human-like expressiveness while preserving linguistic content.

There is a need to develop efficient emotional TTS systems that capture nuanced distinctions between emotions while generating expressive speech. Emo-DPO Gao et al. (2024) combines a text encoder, an emotion-aware decoder, and a flow-matching model with a vocoder, enabling the synthesis of nuanced, controllable emotional speech with improved fidelity and expressiveness. The proposed Emo-DPO framework leverages Direct Preference Optimization (DPO) to refine emotional outputs by optimizing for preferred emotions. Integrating an LLM-based TTS architecture (LLM-TTS), Emo-DPO combines instruction tuning and DPO fine-tuning with Jensen-Shannon regularization to enhance emotional control. Leveraging emotional representation from various perspectives and levels is critical in developing TTS systems that disentangle various emotions. MsEmoTTS Lei et al. (2022), a multi-scale framework, models emotions at global, utterance, and local levels. It employs a pre-trained BERT-based classifier for global emotion prediction, models intonation variations at the utterance level, and controls syllable-level strengths using a ranking-based method. Built on a Tacotron2 backbone with GMM-based attention, it unifies emotion transfer, prediction, and manual control, achieving superior performance in expressive speech synthesis.

Capturing the disentangled emotion and speaker representations in developing emotional speech generation is critical. Zhang, G. at Zhang et al. (2023a) proposes iEmoTTS, a TTS system for cross-speaker emotion transfer that disentangles prosody from timbre. This approach allows robust emotion transfer even when the speaker and emotion features are entangled. iEmoTTS also supports zero-shot emotion transfer to unseen speakers through a timbre encoder with an information bottleneck mechanism, which retains only speaker-specific timbre features while excluding prosodic information. The system is trained end-to-end in a semi-supervised framework, reducing reliance on extensive labeled data. Additionally, it incorporates pre-trained models for style encoding and SER to reduce training data dependency. It introduces perceptual loss functions to enhance emotion intelligibility and discrimination in converted speech.

According to independent emotion expression from the speaker style, Zhou et al. (2020b) proposed a VAW-GAN-based encoder-decoder structure to learn the spectrum and prosody mapping in a speaker-independent manner. DurFlex-EVCOh et al. (2024) model uses de-stylize and stylize transformers, which separate the source style from input features and apply the target emotion style.

Cross-speaker emotion transfer in text-to-speech synthesis requires extracting speaker-independent emotion embeddings, yet existing timbre compression methods such as gradient reversal layers and vector quantization fail to fully separate speaker and emotion characteristics, causing speaker leakage and degraded synthesis quality. DiEmo-TTS Cho et al. (2025b) addresses this through a self-supervised distillation approach that minimizes emotional information loss while preserving speaker identity. The method introduces cluster-driven sampling and information perturbation to retain emotional content while removing speaker-related factors, and proposes an emotion clustering and matching strategy using emotional attribute prediction and speaker embeddings to generalize to unlabeled data. A dual conditioning transformer integrates style features more effectively, and experimental results on the Emotional Speech Dataset confirm that the approach achieves state-of-the-art performance in learning speaker-irrelevant emotion embeddings, excelling in expressiveness, naturalness, and speaker identity preservation.

Current emotional text-to-speech and style transfer methods rely on reference encoders that compress the reference speech into a single global style or emotion vector, which fails to capture nuanced phoneme-level acoustic details and leaves timbre and emotion features entangled in the same representation. To overcome this limitation, Yang et al. (2025) proposes a novel emotional text-to-speech method built on the FastSpeech 2 backbone that predicts fine-grained phoneme-level emotion embeddings while disentangling them from global timbre information. The architecture employs two parallel feature extractors within a dedicated style encoder, a global timbre extractor and a phoneme-aware emotion extractor that aligns reference acoustics with target phonemes via multi-head cross-attention. Mutual Information Neural Estimation explicitly minimizes the mutual information between the two representations, ensuring that the timbre embedding retains only speaker-specific information while the emotion embeddings capture prosodic nuance.

Experimental results demonstrate that this combination of phoneme-level emotion modeling with principled feature disentanglement outperforms strong baselines in both naturalness and style similarity, producing well-separated emotion clusters and enabling more expressive and controllable emotional speech synthesis.
Emotional voice conversion faces persistent challenges due to the complexity of emotion features, which are deeply entangled with speaker identity and linguistic content characteristics, making it difficult to achieve high-quality any-to-any emotion conversion. DiffEmotionVC Su et al. (2025) proposes a diffusion-based framework for any-to-any emotional voice conversion that integrates a dual-granularity emotion encoder capturing both utterance-level emotional context and frame-level acoustic details. The framework employs an orthogonality-constrained condition encoder that disentangles emotion features through gated cross-attention while preserving feature independence with an orthogonal loss. Additionally, multi-objective diffusion training enhances both reconstruction fidelity and emotion discriminability via contrastive learning. Experimental results demonstrate the effectiveness of the framework in maintaining speech quality while optimizing emotional expression.

6.2.3 Challenges Inherent to Emotions

The third category pertains to the intrinsic nature of emotions. Emotions are not merely discrete states; rather, they are highly dynamic Prabhu et al. (2024); Zhou et al. (2021b); Chen et al. (2023b); Zhou et al. (2022a), varying in duration, intensity, valence, and strength Prabhu et al. (2024); Chen et al. (2023b); Chou et al. (2024); Choi & Hahn (2021); Oh et al. (2024). This variability introduces complexities that make it challenging to model and reproduce emotional expressions accurately. Among the three categories, this challenge ranks second in importance, underscoring its substantial impact on achieving human-like emotional speech generation.
EmoConv-Diff Prabhu et al. (2024) uses conditioning on continuous arousal dimensions, allowing for effective control over emotional intensity. Chou et al. (2024) utilized the generative power of diffusion models to tackle significant issues in previous deep learning approaches that use GANs and Autoencoders (AE), specifically concerning quality degradation and limited control over emotions. This method employs a stochastic differential equation-based diffusion process to progressively transform speech features, enabling precise emotion transformation while preserving speaker identity and linguistic content. A disentangled loss distinguishes speaker and emotion representations, while an expressive guidance mechanism enhances emotional expressiveness during reverse diffusion. Zhou et al. (2021b) employs a variational auto-encoding Wasserstein generative adversarial network (VAW-GAN) to transfer seen and unseen emotional style during training and run-time inference. Attention-based Interactive Disentangling Network (AINN)Chen et al. (2023b) uses a two-stage training process to transfer emotional attributes, such as emotional strength and category, from a reference speech to a source speech while preserving the source's content. Li et al. (2022a) addresses the challenges of speaker leakage, emotion strength control, and effective disentanglement of speaker and emotion features in cross-speaker emotion transfer for TTS. It proposes a modified Tacotron2-based framework incorporating an Emotion Disentangling Module (EDM), which uses emotion and speaker encoders with orthogonality constraints to ensure speaker-irrelevant and emotion-discriminative embeddings. Additionally, a scalar value is introduced to control emotion strength in synthetic speech, enabling flexible adjustments without dependency on manually labeled data. The challenges of mono-scale emotion modeling, limited flexibility in emotion transfer and prediction, and the lack of fine-grained emotion control in emotional speech synthesis are explored in Lei et al. (2022). To address the challenge of intensity variation, Emovox Zhou et al. (2022c), a Seq2Seq EVC framework, was developed. Emovox controls emotion intensity using relative attributes to capture fine-grained variations. Zhang, G. at Zhang et al. (2023a) proposes iEmoTTS, which uses a probability-based method for emotion intensity control, enabling a nuanced generation of emotional speech with varying strengths. Although prior frameworks mark a significant advancement in the quality and versatility of EVC systems Sisman et al. (2020), enabling applications like personalized speech synthesis and cross-lingual conversion, they process speech on a frame-by-frame basis, limiting their ability to modify speech duration and also intensity. The first issue was addressed in DurFlex-EVC Oh et al. (2024). The DurFlex-EVC Oh et al. (2024) model incorporates a style AE to disentangle emotional style from linguistic content. This is achieved using de-stylize and stylize transformers, which separate the source style from input features and apply the target emotion style. The

unit aligner further compresses the features to unit-level representations and predicts durations, creating an efficient and context-aware framework for emotional style transformation.

Most existing text-to-speech systems offer only coarse and rigid emotion control, typically relying on discrete emotion labels or carefully crafted emotional text prompts, which makes fine-grained emotion manipulation either inaccessible or unstable, and these models require extensive high-quality datasets for training. EmoSteer-TTS Xie et al. (2025) proposes a training-free approach to achieve fine-grained speech emotion control, including emotion conversion, interpolation, and erasure, through activation steering. The method builds on the empirical observation that modifying a subset of internal activations within a flow matching-based text-to-speech model can effectively alter the emotional tone of synthesized speech. It develops an efficient algorithm comprising activation extraction, emotional token searching, and inference-time steering that can be seamlessly integrated into a wide range of pretrained models. Extensive experiments demonstrate that this approach enables fine-grained, interpretable, and continuous control over speech emotion, outperforming the state of the art as the first method to achieve training-free and continuous fine-grained emotion control in text-to-speech synthesis.

Transferring a mix of primary emotions is an essential task in speech synthesis. Most existing methods focus on imitating a single emotion; however, to facilitate natural and engaging interactions between humans and agents, it is essential to incorporate mixed emotions into speech synthesis. Unfortunately, the development of such models is hindered by the lack of extensive multi-speaker corpora that contain mixed emotion labels Kreibig & Gross (2017). To tackle this issue, Zhou et al. (2022a) uses a ranking-based SVM to model emotional styles as attributes reflecting the relevance of different emotions. This approach enables the system to quantify relationships between emotion pairs and synthesize new emotional mixtures by manually defining these attributes during conversion. Its architecture uses a seq2seq emotional voice conversion framework that integrates these attributes for mixed emotion synthesis. The architecture integrates a text encoder, an emotion encoder for embeddings, and a decoder with bidirectional LSTMs and attention mechanisms to generate natural and expressive emotional speech. Moreover, Zhou et al. (2022d) uses a pre-trained model on massive speech corpora without emotional annotations, which is then fine-tuned using emotional speech data. It similarly uses a ranking function to determine the level of primary emotion according to variations between pairs of emotional speech samples. Bott et al. (2024) deals with controllable emotional prosody in TTS systems using natural language prompts. A FastSpeech 2-based architecture integrates prompt embeddings, derived from DistilRoBERTa, with speaker embeddings through a squeeze-and-excitation mechanism for accurate prosodic control. The system produces high-quality, emotionally expressive speech while preserving the speaker's identity by training with curriculum learning. Period VITS Shirahata et al. (2023) focuses on the challenge of unstable pitch contours and artifacts in end-to-end emotional TTS systems caused by prosodic diversity. The proposed Period VITS integrates a periodicity generator for explicit pitch modeling, producing sample-level sinusoidal sources to enhance pitch stability and waveform quality. A frame pitch predictor within the prior encoder estimates frame-level prosodic features while normalizing flows augment prior distributions for richer acoustic variation. The HiFi-GAN-based decoder aligns pitch signals with latent acoustic features through down-sampling layers. The model is optimized end-to-end using variational inference with combined loss functions to ensure stability and expressiveness. The challenge of incorporating fine-grained intonation control, particularly questioning intonation, into emotional speech synthesis was considered in Tang et al. (2023a). It claims existing TTS models can transfer emotions but struggle to model nuanced prosody like "angry question" versus "angry statement." The proposed QI-TTS builds on FastSpeech 2 and introduces a multi-style extractor to capture emotion at the sentence level and intonation at the final syllable level. By using relative attributes, it models intonation intensity in an unsupervised manner, enabling fine-grained control. A gradient reversal layer ensures content and style disentanglement to prevent interference.

Humans can experience roughly 34,000 distinct types of emotions. This includes eight basic emotions, along with secondary emotions that arise from combinations of these basics, as explained in the Theory of the Emotion Wheel Plutchik (2001). While secondary emotions are critical in social human interaction, synthesizing these emotions is often overlooked. EmoMix Tang et al. (2023b) addresses the challenge of synthesizing emotional speech with controllable intensity and the ability to express mixed emotions, a significant limitation of current text-to-speech (TTS) systems. EmoMix overcomes this issue using a diffusion probabilistic model conditioned on emotion embeddings from a pre-trained SER model. It achieves flexible emotion control by blending predicted noise for different emotions during the sampling process and mixing neutral noise with the target emotion's noise. The architecture of EmoMix incorporates several components: GradTTS, a U-Net model with linear attention, the SER model, the HiFi-GAN vocoder, and a style reconstruction loss that ensures emotional consistency and naturalness in the synthesized speech.

6.2.4 LLM Challenges

The application of LLMs has shown great potential in regulating emotional expression in synthesized speech, particularly through prompt-based techniques Guo et al. (2023); Sigurgeirsson & King (2024). This approach enhances the expressiveness and naturalness of generated speech while preserving clarity and quality Zhang et al. (2025c). While LLMs excel at producing diverse and contextually rich text, their output can sometimes be noisy and inconsistent when directly applied to emotional modifications in speech synthesis. In (ProEmo) Zhang et al. (2025c), researchers noted that relying solely on LLM outputs for emotion control could compromise expressiveness due to output noise. They addressed this by integrating specific emotion and intensity encoders to guide prosody adjustment in systems like FastSpeech2. Moreover, approaches that modify backbone architectures (e.g., FastSpeech2) by adding emotion and intensity encoders have been explored to bridge the gap of effective integration with TTS systems. Such systems leverage the linguistic expressiveness of LLMs while fine-tuning acoustic outputs to convey emotion, although this remains an active area of research with room for improvement. Emotional expression is inherently subjective, and LLMs may generate outputs that do not align with human emotional expectations or ethical guidelines. Li et al. (2024b) have proposed methods like the Emotional Chain-of-Thought (ECoT) prompting technique. This plug-and-play method guides LLMs through multiple reasoning steps to generate emotionally appropriate content, thereby improving human preference alignment in generated outputs.
Computational overhead and latency present significant challenges, particularly for real-time applications such as human-robot interaction. Mishra et al. (2023) explored the real-time use of LLMs for tasks like emotion prediction in dialogue systems. The study indicated that efficient model design and careful engineering are crucial for the practical deployment of these technologies.

7 Emotional Face Synthesis

Emotional facial expressions are fundamental nonverbal signals that enable accurate assessment of internal states in psychiatric applications and bolster the integration of machine learning in mental-health diagnostics Coda-Forno et al. (2023). In therapeutic contexts, these cues guide clinicians in recognizing and addressing patient emotions Iftikhar et al. (2024), while in everyday social exchanges, their contextual interpretation is critical for effective communication and rapport building Han et al. (2024). Beyond human-to-human interaction, facial expressions drive dynamic user engagement across interactive systems, enhancing empathetic response generation in dialogue agents Wang et al. (2025a) and underpinning social intelligence in conversational AIs Chen et al. (2024). The breadth of these roles underscores the centrality of facial affect in advancing emotion research and human-machine interaction Chen & Moscholios (2024). Emotional face synthesis builds on this foundation by enabling virtual avatars, social robots, and conversational agents to exhibit realistic emotional expressions, highlighting its transformative potential in interactive applications Tan et al. (2024). By integrating multimodal cues—visual, auditory, and textual—the emotional authenticity and contextual sensitivity of AI systems are greatly enhanced, which is critical for education, therapy, and collaborative AI Zhan et al. (2022). Moreover, precise lip synchronization combined with coordinated nonverbal signals (facial expressions, gaze, head pose) yields seamless, engaging animations for film, gaming, and HCI Liang & Lu (2024); Ma et al. (2024); Wang et al. (2024). These advancements not only meet ethical and equitable AI standards but also continue to drive innovation across virtual/augmented reality and teleconferencing environments. Here, we briefly outline three core approaches for generating realistic talking faces from audio inputs, static images, and their integration. We then analyze the primary challenges in this field and evaluate the various methods researchers have developed to overcome them. We highlight diverse application domains—virtual avatars, social robotics, sentiment analysis, and HCI—where advanced emotional synthesis drives deeper engagement and improved user experience Pandey & Vishwakarma (2024); Saunders & Namboodiri (2023); Xu et al. (2024a); Pataranutaporn et al. (2021), and discuss ethical considerations such as bias mitigation and responsible deployment of facial manipulation technologies.

Figure 12: Challenges in emotional face synthesis

7.1 Approaches

Photo-realistic facial animation: Photo-realistic facial animation techniques are advanced computational methods designed to generate digital facial animations that closely resemble real human expressions and movements in terms of both visual detail and emotional subtlety. These techniques, such as the Warp-Guided GANs introduced by Geng et al. Geng et al. (2018), advance visual fidelity and emotional nuance, significantly enhancing the realism of animated characters.
Audio-driven synthesis: Audio-driven synthesis systems, exemplified by EDTalk Tan et al. (2024), integrate auditory and visual inputs to create interactions imbued with emotional depth and realism. Unlike photo-realistic facial animation techniques, which prioritize high visual fidelity to achieve lifelike faces, these

methods do not require a highly realistic facial appearance. Instead, they focus on synchronizing audio and visual cues to produce expressive and engaging facial animations that enhance the quality of virtual interactions.
Multi-modal synthesis and editing: Multi-modal synthesis and editing techniques, encompassing GAN-based, autoregressive, diffusion, and NeRF approaches, enhance emotional authenticity and contextual awareness in facial animations Zhan et al. (2022).

7.2 Challenges

Emotional facial generation faces several critical challenges, including challenges related to data, models, emotions, multi-modal integration, and computational inefficiencies. Addressing these issues is essential for creating inclusive, robust systems capable of accurately representing the full spectrum of human emotions. These insights provide a foundation for advancing the field and overcoming current limitations. Figure 12 classifies these challenges into five major domains: data, learning models, emotion synthesis, multimodal, and computational issues. Table 6 summarizes these categories, outlining their respective sub-issues and highlighting studies that address these challenges. Next, we will discuss the challenges and the methods proposed to overcome them.

7.2.1 Data-Related Challenges

Dataset bias significantly impacts the generalization capability of affective face generation models, particularly in representing diverse demographics, emotional expressions, and cultural nuances. Limited training data diversity further restricts the generalizability of models to diverse emotional states Siddiqui (2022). Biased datasets often fail to capture the complexity of human emotions, leading to inequities in model performance and reduced inclusivity Washington et al. (2021). This issue is compounded by the difficulty of achieving localized control over facial muscle movements and generating nuanced expressions Varanka et al. (2024). Moreover, existing benchmarks inadequately capture the fluidity and diversity of emotional expressions, limiting the generalization capabilities of models in real-world scenarios Xu et al. (2024a). Current methods relying on detailed face modeling encounter challenges in unconstrained environments with variations in pose, lighting, and expression Wiles et al. (2018). The need for comprehensive annotation of facial attributes further restricts the manipulation of identity and attributes, impeding the development of robust models Bao et al. (2018). Multi-modal datasets also struggle with aligning features across modalities, resulting in limited high-resolution, contextually relevant outputs Zhan et al. (2022). The lack of diversity within and across identities in datasets exacerbates these issues, underscoring the importance of inclusive data collection practices Xu et al. (2024b). ComFace Akamatsu et al. (2025) is introduced to address the challenge of capturing intra-personal facial changes for health and emotion monitoring, hindered by the scarcity of temporally varying real-world face images. Its novel representation learning approach leverages StyleGAN-generated synthetic data to simultaneously learn both inter-personal facial differences and intra-personal facial changes within individuals. Remarkably, their method trained exclusively on synthetic data achieves comparable or superior performance to state-of-the-art approaches trained on real images across multiple facial change estimation tasks. Many approaches reconstructed 3D faces from 2D videos using parametric models (3DMMs), but these lacked precision for accurate lip-syncing. The authors solve this with EmoVOCA Nocentini et al. (2025), a data-driven framework that combines VOCAset (neutral 3D talking heads with speech) and Florence4D (expressive 3D faces without speech) using a Double Encoder/Shared Decoder architecture, where separate encoders learn speech and expression features while a shared decoder combines them to create synthetic emotional 3D talking heads that preserve both accurate lip movements and convincing expressions. Privacy concerns intersect with dataset bias, as existing methods often compromise either privacy or the quality of expression recognition Chen et al. (2018). Ethical implications arise from biased datasets, which hinder fairness and generalization across diverse populations Stahl et al. (2023). Addressing these concerns requires promoting diversity in datasets while maintaining identity consistency, as well as enhancing annotation processes to improve the representation of nuanced and localized expressions Xu et al. (2024b); Varanka et al. (2024). These efforts are critical for fostering fairness, equity, and trust in affective face generation technologies.

Table 6: Summary of challenges and solutions in emotional face synthesis

Challenge	Sub-challenge	Solutions
Data-Related	Scarcity & Diversity	Promoting diversity in datasets (Xu et al., 2024b); Enhanced annotation for expressions (Chen et al., 2018)
	Limited Training Data	Synthetic data generation (Xu et al., 2024c); StyleGAN-generated data (Akamatsu et al., 2025); EmoVOCA dataset combination (Nocentini et al., 2025)
	Occlusions	Robust model design (Geng et al., 2018); Specialized training data (Xu et al., 2024c)
Model-Related	Realism & Identity	Advanced GANs (StyleGAN) (Jiang et al., 2023); Diffusion Models (Yin et al., 2022); Identity-emotion disentanglement (Tan et al., 2025); Instruction-driven 3D generation (Vo et al., 2025)
	Generalization	Robust architectures (Xu et al., 2024c); Domain adaptation (Geng et al., 2018); Long-range autoregressive diffusion (Zhang et al., 2025b)
	Scalability	Unified frameworks (UniPortrait) (He et al., 2024b); Motion modeling (MotionGAN) (Otberdout et al., 2022); Temporal layers (Xu et al., 2024d); Sparse landmark methods (Tu et al., 2024)
	Training Stability	Improved GAN architectures (Jiang et al., 2023); Regularization techniques (Yin et al., 2022); Alternative loss functions (Jiang et al., 2023)
Emotion Synthesis	Nuanced Expressions	Fine-grained control (Action Units) (Yin et al., 2022); Continuous modeling (GC-GANs) (Qiao et al., 2018); Subtle cue analysis (Retsinas et al., 2024); Emotion-audio spatial attention (Ma et al., 2026); Continuous valence-arousal conditioning (Cha et al., 2025); FLAME-guided diffusion (Zhang et al., 2024c); Semantic expression parameters (Shen et al., 2025)
	Micro-expressions	High-res temporal modeling (Xu et al., 2024d)
	Localized Control	Advanced facial modeling (Retsinas et al., 2024); Anatomical models (Varanka et al., 2024)
Multi-Modal	Feature Alignment	Collaborative diffusion (Huang et al., 2023); Self-supervised learning (Pham et al., 2017); Audio-visual integration (C3D-DBN) (Nguyen et al., 2017); Text-to-expression synthesis (Cheng et al., 2024); Multi-modal emotion embedding (Yee et al., 2025)
	Contextual Relevance	Textual data integration (Cheng et al., 2024); Sentiment analysis (VECTN) (Pandey & Vishwakarma, 2024); Cross-reconstruction disentanglement (Liu et al., 2025)
Computational	Efficiency	Hybrid models (GANs+Diffusion) (Xu et al., 2024d); Compact representations (He et al., 2024b); Efficient alignment (Liu et al., 2018); Model quantization/pruning (Liu et al., 2018); Knowledge distillation (He et al., 2024b); Hardware acceleration (He et al., 2024b)

7.2.2 Model-Related Challenges

Three primary sub-challenges reside in model-related challenges: preserving identity, ensuring the realism of the model's outputs, and achieving generalization to real-world scenarios. Advanced generative methodologies for generating emotionally expressive facial animations, such as GANs and diffusion models, could address these issues, and these frameworks have significantly improved the quality of affective facial generation, enabling applications in human-computer interaction, virtual avatars, and sentiment analysis. GANs have been instrumental in advancing affective face generation, producing high-quality animations that capture nuanced emotional states. Methods like EDTalk Tan et al. (2024) enhance realism by disentangling facial features such as mouth shape, head pose, and emotional expression. Similarly, techniques integrating driving videos with single-image inputs, as demonstrated in Averbuch-Elor et al. (2017), expand the versatility of GANs in synthesizing dynamic animations. Despite the greatness of GANs, challenges like generalization to real-world faces persist Xu et al. (2024c). Diffusion models provide a robust alternative to GANs, leveraging iterative noise refinement to generate diverse emotional attributes while maintaining identity consistency. Conditional diffusion frameworks such as ID3 excel at preserving intra-class identity and generating emotionally coherent expressions Wang et al. (2023). Innovations like Stable Animator address temporal stability, ensuring smooth transitions in dynamic animations Tu et al. (2024). These models also bridge textual input and emotional synthesis, as demonstrated by frameworks like ContinuousText-to-Expression Generator (CTEG) and Globally-informed Gaussian Avatar (GiGA), which produce nuanced 3D avatars Xu et al. (2024a). These models excel at capturing subtle emotional cues and offer advantages in emotional synthesis Zhan et al. (2022). Innovations like READ Avatars, Emo3D Xu et al. (2024a) generation, and UniPortrait He et al. (2024b) leverage adversarial loss, autoregressive models, and unified frameworks to enable applications in hyper-realistic avatars, empathetic human-computer interaction, and nuanced sentiment analysis Saunders & Namboodiri (2023); Pataranutaporn et al. (2021). Hybrid approaches combining GANs and diffusion models aim to leverage the strengths of both techniques. For instance, UniPortrait He et al. (2024b) employs a dual-module architecture to enhance identity preservation and adaptability, balancing the speed of GANs with the nuanced output of diffusion models. Additionally, frameworks utilizing Action Units for fine-grained expression control demonstrate the potential for bridging the gap between realism and expressiveness in affective face synthesis Yin et al. (2022). By integrating the strengths of GANs and diffusion models, researchers are advancing the synthesis of nuanced, contextually appropriate facial expressions. These models enhance realism and emotional dynamics, enabling applications in virtual avatars, human-computer interaction, and personalized content creation Pumarola et al. (2018); Bouzid & Ballihi (2022); Siddiqui (2022); Huang & Khan (2017).

Existing diffusion-based talking head generation methods struggle to produce emotionally expressive portraits while preserving speaker identity, due to insufficient utilization of audio's inherent emotional cues, identity leakage in emotion representations, and isolated learning of emotion correlations. DICE-Talk Tan et al. (2025) addresses these limitations through a framework that disentangles identity from emotion and then cooperates emotions with similar characteristics. It develops a disentangled emotion embedder that jointly models audio-visual emotional cues through cross-modal attention, representing emotions as identity-agnostic Gaussian distributions. A correlation-enhanced emotion conditioning module with learnable Emotion Banks explicitly captures inter-emotion relationships through vector quantization and attention-based feature aggregation, while an emotion discrimination objective enforces affective consistency during the diffusion process through latent-space classification. Experiments on the MEAD and HDTF datasets demonstrate superiority in emotion accuracy while maintaining competitive lip synchronization performance. Audio-driven portrait animation methods typically emphasize lip synchronization and short-range visual fidelity in constrained speaking scenarios, but fail to capture nuanced, dynamically evolving emotions that flow coherently with the rhythm and content of speech over long temporal contexts. X-Actor Zhang et al. (2025b) presents an audio-driven portrait animation framework that generates lifelike, emotionally expressive talking head videos from a single reference image and an input audio clip, enabling actor-quality long-form portrait performance. Central to the approach is a two-stage decoupled generation pipeline: an audio-conditioned autoregressive diffusion model predicts expressive yet identity-agnostic facial motion latent tokens within a long temporal context window, followed by a diffusion-based video synthesis module that translates these motions into high-fidelity video animations. By operating in a compact facial motion latent space decoupled from visual and identity cues, the autoregressive diffusion model captures long-range correlations between audio and facial dynamics through a diffusion-forcing training paradigm, enabling infinite-length emotionally rich motion prediction without error accumulation. Generating 3D facial expressions from natural language instructions remains challenging because most existing methods rely on discrete emotion labels or predefined expression categories, which cannot capture the richness and specificity of textual descriptions for both static expressions and dynamic expression transitions. Vo et al. (2025) proposes an instruction-driven approach for 3D facial expression generation and transition that takes text instructions as input and produces corresponding 3D facial animations. The method leverages a language model to interpret free-form textual descriptions and maps them to parametric 3D face model expression parameters, enabling both the generation of target expressions and smooth transitions between emotional states. This text-based control paradigm offers greater flexibility and user accessibility compared to conventional label-driven or reference-driven approaches, broadening the applicability of 3D facial animation in interactive systems and content creation.

7.2.3 Multi-modal Integration Challenges

Multi-modal integration is pivotal for generating emotionally authentic facial expressions. By combining visual, auditory, and textual cues, multi-modal frameworks align generated expressions with intended emotional states, ensuring contextual relevance. Frameworks like EmotiveTalk Wang et al. (2024) and Emotion-LLaMA Cheng et al. (2024) use audio decoupling and self-supervised learning to align speech, lip movements, and expressions, producing realistic talking-head videos Liang & Lu (2024); Zhan et al. (2022). Collaborative Diffusion Huang et al. (2023) exemplifies this by integrating pre-trained uni-modal diffusion models, enhancing emotional synthesis through modality synergy. EDTalk Tan et al. (2024) exemplifies the importance of synthesizing multiple modalities for lifelike outputs. Audio-visual integration remains central to multi-modal systems; methods like C3D-DBN Nguyen et al. (2017) align auditory cues (e.g., tone, pitch) with visual signals for coherent emotional outputs. Emotion-LLaMA further incorporates textual data, enabling contextually appropriate expression synthesis Cheng et al. (2024). Techniques for compact multi-modal representations, like Liu et al. (2018) improve real-time processing and scalability, making these systems suitable for virtual avatars and conversational AI. Multi-modal frameworks like VECTN integrate textual data for sentiment analysis, aligning emotional synthesis with contextual information and capturing complex states like sarcasm Pandey & Vishwakarma (2024). By synthesizing facial expressions aligned with multi-modal cues, these systems enhance user engagement and emotional resonance in AI-driven interactions Zhan et al. (2022). Most existing emotion-aware talking face generation methods rely on a single modality, either audio or image, for emotion embedding, which limits their ability to capture nuanced affective cues, and conditioning on a single reference image restricts the representation of dynamic changes in actions or attributes across time. SynchroRaMa Yee et al. (2025) introduces a framework that integrates a multi-modal emotion embedding by combining emotional signals from text via sentiment analysis and audio via speech-based emotion recognition and audio-derived valence-arousal features, enabling the generation of talking face videos with richer and more authentic emotional expressiveness. To ensure natural head motion and accurate lip synchronization, the framework includes an audio-to-motion module that generates motion frames aligned with the input audio. Additionally, scene descriptions generated by a large language model serve as additional textual input, capturing dynamic actions and high-level semantic attributes that enhance temporal consistency and visual realism. Audio-driven emotional 3D facial animation typically relies on static and predefined emotion labels, which limits the diversity and naturalness of generated expressions and prevents fine-grained dynamic emotional control. MEDTalk Liu et al. (2025) proposes a framework for fine-grained and dynamic emotional talking head generation that first disentangles content and emotion embedding spaces from motion sequences using a carefully designed cross-reconstruction process, enabling independent control over lip movements and facial expressions. Beyond conventional audio-driven lip synchronization, the method integrates audio and speech text to predict frame-wise intensity variations and dynamically adjust static emotion features for realistic emotional expressions. Furthermore, multimodal inputs including text descriptions and reference expression images guide the generation of user-specified facial expressions, and the generated results are compatible with MetaHuman for integration into industrial production pipelines.

7.2.4 Computational Challenges

These systems often face computational inefficiencies and scalability challenges, particularly in synthesizing realistic facial expressions under diverse conditions. Advanced generative models, such as GANs and diffusion models, require substantial computational resources for training and inference, limiting their applicability in resource-constrained environments Zhan et al. (2022). Slow inference speeds and extensive training data requirements further hinder real-time applications and large-scale deployments. Existing methods, such as Warp-Guided GANs, struggle with generalizing to unseen data or scenarios with significant occlusions, highlighting the limitations of current frameworks Geng et al. (2018). Additionally, techniques that assume a neutral face as a starting point or rely on simplified representations restrict adaptability to dynamic environments Averbuch-Elor et al. (2017). Identity blending and the need for extensive fine-tuning also pose challenges, as frameworks like UniPortrait demonstrate the resource-intensive nature of achieving identity preservation and adaptability He et al. (2024b). The inability to effectively model dynamic and contextually adaptive facial animations further limits the realism of generated outputs. Many methods fail to capture intricate temporal dynamics, such as micro-expressions and head movements, which are critical for producing emotionally resonant animations Otberdout et al. (2020); Xu et al. (2024c). Multi-modal frameworks introduce additional complexity during training and inference, requiring significant resources to align features and synthesize coherent outputs Zhan et al. (2022). Efforts to address these challenges include optimizing algorithms and architectures to reduce resource consumption while maintaining high-quality synthesis. Addressing computational constraints and improving multi-modal integration will enhance the robustness and accessibility of affective face generation systems, facilitating their application in diverse domains Liu et al. (2018); Zhan et al. (2022).

7.2.5 Emotion Synthesis Challenges

Temporal consistency is a cornerstone of dynamic facial animation, preventing abrupt transitions that undermine emotional impact and is critical for video synthesis, ensuring smooth transitions and coherent animations Zhan et al. (2022); Tan et al. (2024). Techniques like those in Tan et al. (2024) disentangle temporal dynamics from other facial attributes, while methods such as driving-video-based animation ensure smooth transitions Averbuch-Elor et al. (2017). Techniques like MotionGAN model expression transitions on a hypersphere, minimizing motion artifacts and enhancing temporal fidelity Otberdout et al. (2022). Sparse landmark-based methods leverage anatomical priors for coherent facial deformations with reduced computational complexity Xu et al. (2024d). Disentangling temporal dynamics from attributes like expression and pose further enhances coherence. Frameworks integrating temporal layers mitigate interference with spatial priors, ensuring stable animations Averbuch-Elor et al. (2017); Tu et al. (2024); Xu et al. (2024d). Multi-modal cues, such as audio and textual signals, align facial dynamics with emotional intent, as seen in EmotiveTalk, which synchronizes lip movements and expressions with emotional audio cues Pham et al. (2017); Liang & Lu (2024); Wang et al. (2024). Innovations like Takin-ADA address challenges like expression leakage, while Action Unit-based techniques provide anatomically accurate representations of expressions. These advancements improve dynamic animations, broadening their applicability in teleconferencing, virtual reality, and digital media Vougioukas et al. (2020) Pumarola et al. (2018) Lin et al. (2024). Techniques such as SMIRK Retsinas et al. (2024) and Auxiliary Classifier GANs Siddiqui (2022) aim to address complex emotional nuances by enhancing the quality and diversity of generated expressions, yet challenges persist in capturing them. Occlusions often lead to assumptions of exaggerated expressions, compromising the fidelity of nuanced emotional synthesis Retsinas et al. (2024). Temporal modeling is critical for generating continuous and smooth facial expression videos, yet existing approaches often fail to simulate intricate transitions and micro-expressions effectively Otberdout et al. (2020); Xu et al. (2024c). Discrete methods for facial expression generation also cannot capture the continuity of emotional transitions, though advancements like Geometry Contrastive GANs (GC-GANs) demonstrate the potential for high-fidelity continuous expressions Qiao et al. (2018). Interactive generative adversarial networks (iGANs) and temporal behavioral biometrics illustrate the importance of incorporating temporal consistency and additional facial behaviors, such as head pose, to enhance emotional authenticity Nojavanasghari et al. (2018); Agarwal et al. (2020). To advance nuanced emotional synthesis, researchers must focus on robust temporal modeling, effective handling of occlusions, and seamless multimodal integration. Innovations such as GC-GANs and advanced frameworks for continuous, contextually adaptive expression generation are pivotal for achieving emotionally intelligent systems. These advancements are essential for applications requiring high levels of emotional realism and human-like interaction capabilities, driving progress in affective face generation. While some studies have addressed the generation of facial videos driven by emotional audio, efficiently generating high-quality talking head videos that integrate both emotional expressions and style features remains a significant challenge, as most current audio-driven facial animation research primarily focuses on generating videos with neutral emotions. ESGaussianFace Ma et al. (2026) proposes a framework for emotional and stylized audio-driven facial animation that leverages 3D Gaussian Splatting to reconstruct 3D scenes and render videos, ensuring efficient generation of 3D consistent results. The method introduces an emotion-audio-guided spatial attention mechanism that effectively integrates emotion features with audio content features, enabling more accurate reconstruction of facial details across different emotional states. Two 3D Gaussian deformation predictors achieve emotional and stylized deformations of the Gaussian points through emotion and style features, and a multi-stage training strategy enables step-by-step learning of the character's lip movements, emotional variations, and style features. 3D Gaussian splatting-based talking head synthesis has gained attention for its ability to render high-fidelity images with real-time inference speed, but since it is typically trained on only a short video that lacks diversity in facial emotions, the resultant talking heads struggle to represent a wide range of emotions. EmoTalkingGaussian Cha et al. (2025) addresses this by proposing a lip-aligned emotional face generator that trains a 3D Gaussian splatting model capable of manipulating facial emotions conditioned on continuous emotion values, specifically valence and arousal, while retaining synchronization of lip movements with input audio. To achieve accurate lip synchronization for in-the-wild audio, the method introduces a self-supervised learning approach that leverages a text-to-speech network and a visual-audio synchronization network. Experiments on publicly available videos demonstrate improvements over existing methods in image quality, emotion expression accuracy, and lip synchronization. While many existing approaches to audio-driven portrait animation focus on lip synchronization and video quality, few tackle the challenge of generating emotion-driven talking head videos with fine-grained control over both emotion categories and intensities. EMOdiffhead Zhang et al. (2024c) proposes a method for emotional talking head video generation that enables fine-grained control of emotion categories and intensities while supporting one-shot generation. Given the linearity of the FLAME 3D model in expression modeling, the method extracts expression vectors using the DECA approach and combines them with audio to guide a diffusion model in generating videos with precise lip synchronization and rich emotional expressiveness. This approach enables learning rich facial information from emotion-irrelevant data while facilitating the generation of emotional videos, effectively overcoming the limitations of emotional data such as the lack of diversity in facial and background information. Generating emotion-specific talking head videos from audio input is a complex challenge because emotion is a highly abstract concept with ambiguous boundaries, necessitating disentangled expression parameters to produce emotionally expressive results. EmoHead Shen et al. (2025) presents a method to synthesize talking head videos via semantic expression parameters. An audio-expression module that can be specified by an emotion tag predicts expression parameters for arbitrary audio input, enhancing the correlation from audio input across various emotions. The method leverages a pre-trained hyperplane to refine facial movements by probing along the vertical direction, and the refined expression parameters regularize neural radiance fields to facilitate emotion-consistent generation of talking head videos. Experimental results demonstrate that semantic expression parameters lead to improved reconstruction quality and controllability.

8 Discussion and Future Prospects

The rapid progress in affective computing highlights both remarkable achievements and unresolved challenges. To move closer to emotionally intelligent agents, future research must address key gaps in emotion understanding, affective cognition, and expression. This section outlines promising research directions, emphasizing the need for robust datasets, interpretable models, multimodal integration, and ethical frameworks to ensure trustworthy and human-centered development.

8.1 Future Prospects in Emotion Understanding

Despite significant advances in emotion recognition, the field continues to encounter persistent challenges that hinder practical deployment and broad generalization of affective computing systems. Current solutions often depend on limited or biased datasetsThakur & Gupta (2026), and models may struggle to interpret emotions that are complex, overlapping, or deeply shaped by cultural nuances. Although recent progress in advanced multimodal fusion strategies has led to improved performance, the optimal integration of diverse modalities, particularly in noisy, real-world environments, remains an open problem Nandini et al. (2025). Additional barriers, such as model interpretability, the lack of unified evaluation metrics, and limited cross-domain generalizability, require sustained research attention. As highlighted by Di Luzio et al. (2025), deep learning models, while highly accurate, frequently operate as "black boxes," making it difficult to justify or explain their predictions in sensitive applications. This fundamental limitation underscores the need for new interpretability frameworks in emotion recognition Di Luzio et al. (2025). To address the lack of explainability in existing multimodal aspect-based sentiment analysis methods, Wang et al. (2026) reformulates this as a generative task using multimodal large language models, enabling joint aspect-level sentiment prediction and natural language explanation generation. The proposed framework incorporates dependency-syntax-guided sentiment cues to enhance aspect-oriented reasoning and improve the faithfulness of generated explanations. Alharbi (2024) demonstrated that employing explainable feature selection methods in virtual reality environments not only increases user trust but also facilitates clinical adoption, though effectively communicating technical insights to end-users remains challenging. With the emergence of LLMs and foundation models, the landscape of affective computing is rapidly evolving, offering new opportunities such as zero-shot and few-shot emotion recognition. Nonetheless, these models introduce critical risks, including hallucinated or spurious emotional attributions, difficulties in context understanding, and high annotation costs Wognum et al. (2024). Future research should prioritize the creation of large-scale, diverse, and high-quality datasets that capture real-world variability across individuals and cultures. Further efforts are needed to develop robust and explainable emotion recognition models, along with standardized evaluation metrics tailored to the complexities of multimodal affective computing. Finally, realizing the full potential of LLMs and foundation models in this domain will require systematic solutions for mitigating hallucinations, improving context awareness, and reducing labeling costs—ensuring that advances in hybrid affective systems translate to trustworthy and generalizable real-world applications.

8.2 Future Prospects in Affective Cognition

Emotional intelligence has been extensively studied from psychological, neuroscientific, and technological perspectives, with prior work reviewing its theoretical foundations, neurobiological underpinnings, and measurement methods, as well as discussing its integration into emotion-aware and human-centered AI systems and related ethical considerationsEspinosa Gámez (2026). One of the critical tasks for advancing human-centered AI systems is the integration of emotion prediction and elicitation within complex interactive scenarios Zhang et al. (2024b). This involves developing models that can accurately forecast emotional responses and trigger appropriate emotional states in dynamic, real-world human-computer interaction (HCI) settings, thereby enhancing the adaptability and effectiveness of such systems. This could involve controlled experiments to elicit targeted emotions or the collection of large-scale naturalistic datasets capturing diverse human emotions. Additionally, replacing the simple moving window average with advanced time-dependent models could better capture the dynamic persistence of emotions. The implications of models include its potential integration into adaptive interactive systems that anticipate user emotional states, providing designers with insights into how task progression and user goals influence emotional outcomes. Future developments should focus on personalizing the model to individual user objectives and proficiency, thereby improving its alignment with cognitive states and enhancing prediction accuracy in affective computing applications.

While language models play a significant role in the advancement of modern affective computing, their contributions to emotional elicitation and emotional experiences are still limited. These models are mainly used for understanding emotions and generating emotional expressions. Most existing research has concentrated on assessing their ability to model emotional elicitation Khan et al. (2025). However, the development of intelligent agents based on language models, particularly those with specific cognitive skills such as decision-making in social contexts, has been relatively neglected. To create an intelligent agent with emotional intelligence using language models, it is essential to combine these models with a specific cognitive computational framework for the agent. This integration enables the knowledge and experiences to influence agent decision-making processes and emotional responses, ultimately enhancing its emotional intelligence and adaptability.

8.3 Future Prospects in ETS

The future of affective text synthesis is poised for significant advancements, driven by the integration of more sophisticated AI models and a deeper understanding of human emotion Picard (1997). As models become more adept at capturing the nuances of emotional expression, we can expect to see more personalized and empathetic human-computer interactions Singh et al. (2020). The development of large-scale, high-quality emotional datasets will be crucial in training models that can generate truly authentic and contextually appropriate affective text. Furthermore, the exploration of cross-modal emotion generation, where text is synthesized in conjunction with other modalities like speech and facial expressions, will open up new frontiers in creating immersive and emotionally resonant experiences McDarby et al. (2003). Ethical considerations will be a major focus, particularly in preventing the misuse of emotion generation for manipulation and ensuring fairness in algorithmic emotional expression. Methods like emotion embeddings and reinforcement learning can be used to improve emotional consistency through reward mechanisms. However, achieving a balance between emotional consistency and coherence, particularly for subtle emotions, remains a significant challenge Lu et al. (2022).

8.4 Future Prospects in ESS

As previously discussed, a primary challenge in speech generation is acquiring high-quality audio data. An optimal dataset for TTS should encompass diverse emotions and duration, multi-speaker, and multi-gender recordings, featuring authentic, unacted voices in parallel formats. Sources such as podcasts, films, and theater performances are valuable for collecting varied vocal expressions. Additionally, preserving the emotional nuances present in performed poetry can enhance the expressiveness of TTS outputs. A dataset similar to Emilia He et al. (2024a), but tailored for emotional speech generation (ESG), would be ideal. Models must be capable of converting and synthesizing various emotions with different durations, ensuring speaker independence and preventing emotion leakage. GANs and diffusion models have emerged as promising approaches to address these challenges. GANs, through their adversarial training mechanism, effectively generate complex, high-dimensional data, making them suitable for producing nuanced emotional speech Ma et al. (2024). Diffusion models, on the other hand, have been utilized to synthesize speech with mixed emotions or varying intensities, offering fine-grained control over emotional expression Tang et al. (2023b). Reinforcement learning (RL) also presents a valuable framework, particularly for conversational speech synthesis. In this context, an agent interacts with the environment, receiving feedback that guides the learning process. This interactive paradigm enables the model to learn and adapt to various emotional expressions, even from a limited set of basic emotions, by refining its performance through continuous interaction Liu et al. (2021). LLMs have demonstrated proficiency in generating TTS systems. Advancements in LLMs, like rising Deepseek Guo et al. (2025), will lead to more efficient and less computationally intensive emotional speech generation systems. In conclusion, ESG can be improved by synthesizing diverse emotions with accurate control and preventing emotion leakage. Techniques like GANs, diffusion models, and reinforcement learning show promise in achieving this. Advancements in efficient LLMs further enhance ESG systems, and the rise of speech-language models Cui et al. (2024) is a great opportunity to have efficient ESG systems. These innovations will lead to more natural and expressive emotional speech synthesis, enriching human-computer interaction.

8.5 Future Prospects in Emotional Face Synthesis

The future of emotional face synthesis presents exciting opportunities at the intersection of diffusion models, multimodal integration, and ethical AI development. Recent advancements in diffusion model-based approaches have demonstrated superior performance in generating realistic and emotionally expressive faces, with works like Xu et al. (2024b) achieving significant improvements in emotion similarity metrics. The field is rapidly moving toward more nuanced expression control, with emerging research focusing on micro-expression modeling and anatomically informed facial muscle simulation Retsinas et al. (2024); Varanka et al. (2024). Multimodal integration represents another promising direction, as researchers develop collaborative diffusion techniques that synchronize facial expressions with speech and text inputs Huang et al. (2023); Pham et al. (2017). Additionally, the development of efficient architectures and model compression techniques addresses computational challenges, making real-time emotional face synthesis increasingly feasible He et al. (2024b); Liu et al. (2018). However, recent research has highlighted significant ethical considerations that must be addressed as these technologies mature. Studies from 2024 have challenged the universality hypothesis of facial expressions, emphasizing that emotions are expressed and perceived differently across cultures and contexts Katirai (2024). This cultural variability raises concerns about potential biases in emotional face synthesis systems. Furthermore, as the global market for emotion recognition and synthesis technologies expands, issues of privacy, consent, and potential misuse in sensitive domains like employment, healthcare, and surveillance demand careful consideration Wu (2024). Future development will require robust ethical frameworks that address fairness, non-discrimination, and a defined scope of use, particularly as these technologies become more integrated into human-computer interaction systems Ballesteros et al. (2024). The responsible advancement of emotional face synthesis will depend on balancing technological innovation with ethical guidelines that protect individual privacy and prevent emotional manipulation, ensuring these powerful generative systems serve beneficial purposes.

9 Conclusion

This study has provided a comprehensive exploration of the integration of emotional intelligence into intelligent agents, highlighting the critical roles of emotion understanding, affective cognition, and emotion expression in fostering naturalistic and empathetic human-computer interactions. By systematically analyzing the challenges, ranging from dataset limitations and model interpretability in emotion understanding to contextual and cognitive complexities in affective cognition and multimodal synchronization in emotion expression, we have underscored the multifaceted hurdles that impede progress in affective computing. We outline and explore recent solutions for revealing promising pathways to overcome these obstacles. This work not only synthesizes current advancements but also proposes a roadmap for future research, paving the way for the development of emotionally intelligent agents capable of truly adaptive, empathetic, and human-like interactions. Our findings underscore the necessity of continued innovation in data collection, model design, and evaluation frameworks to overcome existing barriers and pave the way for emotionally intelligent agents that can foster trust, empathy, and more effective human-computer interaction.

References

H. Abdi. Multivariate analysis. *Encyclopedia for Research Methods for the Social Sciences*, pp. 699–702, 2003.

A. Abilbekov, S. Mussakhojayeva, R. Yeshpanov, and H. A. Varol. Kazemotts: A dataset for kazakh emotional text-to-speech synthesis. *arXiv preprint arXiv:2404.01033*, 2024.

S. Afzal, H. A. Khan, M. J. Piran, and J. W. Lee. A comprehensive survey on affective computing; challenges, trends, applications, and future directions. *IEEE Access*, 2024.

S. Agarwal, H. Farid, T. El-Gaaly, and S. N. Lim. Detecting deep-fake videos from appearance and behavior. In *2020 IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, pp. 1–6, 2020.

R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki. Gmm-based emotional voice conversion using spectrum and prosody features. *Amer. J. Signal Process.*, 2(5):134–138, 2012.

Y. Akamatsu, T. Umematsu, H. Imaoka, S. Gomi, and H. Tsurushima. Comface: Facial representation learning with synthetic data for comparing faces. In *2025 IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 5263–5273, 2025.

H. Alharbi. Explainable feature selection and deep learning based emotion recognition in virtual reality using eye tracker and physiological data. *Frontiers in Medicine*, 11:1438720, 2024.

Tursun Alkam, Andrew H. Van Benschoten, and Ebrahim Tarshizi. Reinforcement learning in artificial intelligence and neurobiology. *Neuroscience Informatics*, pp. 100220, 2025.

D. Alnuhait, Q. Wu, and Z. Yu. Facechat: An emotion-aware face-to-face dialogue framework. *arXiv preprint arXiv:2303.07316*, 2023.

H. F. T. Alsaadawi, B. Das, and R. Das. Tac-trimodal affective computing: Principles, integration process, affective detection, challenges, and solutions. *Displays*, pp. 102731, 2024.

M. M. Amin, R. Mao, E. Cambria, and B. W. Schuller. A wide evaluation of chatgpt on affective computing tasks. *IEEE Transactions on Affective Computing*, 2024.

J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin. An integrated theory of the mind. *Psychological Review*, 111(4):1036–1060, 2004a.

J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin. An integrated theory of the mind. *Psychological Review*, 111(4):1036, 2004b.

S. T. Aroyehun, L. Malik, H. Metzler, N. Haimerl, A. Di Natale, and D. Garcia. Leia: Linguistic embeddings for the identification of affect. *EPJ Data Science*, 12(1):52, 2023.

Authors. Flexible thinking for multimodal emotional support conversation via reinforcement learning. In *Findings of EMNLP 2025*, 2025.

H. Averbuch-Elor, D. Cohen-Or, J. Kopf, and M. F. Cohen. Bringing portraits to life. *ACM Trans. Graph. (TOG)*, 36(6):196:1–196:13, 2017.

J. A. Ballesteros, V. G. M. Ramirez, F. Moreira, A. Solano, and C. A. Pelaez. Facial emotion recognition through artificial intelligence. *Front. Comput. Sci.*, 6:1359471, 2024.

J. Bao, D. Chen, F. Wen, H. Li, and G. Hua. Towards open-set identity preserving face synthesis. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 6713–6722, 2018.

A. Bayro and H. Jeong. A systematic review of experimental protocols: Towards a uniform framework in virtual reality affective research. *IEEE Transactions on Affective Computing*, 2025.

J. Becker, J. P. Wahle, B. Gipp, and T. Ruas. Text generation: A systematic literature review of tasks, evaluation, and challenges. *arXiv preprint arXiv:2405.15604*, 2024.

C. Becker-Asano. *WASABI: Affect Simulation for Agents with Believable Interactivity*, volume 319. IOS Press, 2008.

C. Becker-Asano and I. Wachsmuth. Affective computing with primary and secondary emotions in a virtual human. *Autonomous Agents and Multi-Agent Systems*, 20(1):32–49, 2010.

L. Berto, A. Tanevska, A. Cirne, P. Costa, A. Simoes, R. Gudwin, F. Rea, E. Colombini, and A. Sciutti. Curiosity and affect-driven cognitive architecture for HRI. *IEEE Transactions on Affective Computing*, 2025. In Press.

Sree Bhattacharyya, Lucas Craig, Tharun Dilliraj, Jia Li, and James Z Wang. Do machines think emotionally? cognitive appraisal analysis of large language models. *arXiv preprint arXiv:2508.05880*, 2025.

S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7327–7347, 2021.

P. J. Bota, C. Wang, A. L. Fred, and H. P. Da Silva. A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals. *IEEE Access*, 7:140990–141020, 2019.

T. Bott, F. Lux, and N. T. Vu. Controlling emotion in text-to-speech with natural language prompts. *arXiv preprint arXiv:2406.06406*, 2024.

M. Bourgais, P. Taillandier, and L. Vercouter. An agent architecture coupling cognition and emotions for simulation of complex systems. In *Proceedings of the 12th Social Simulation Conference*, 2016.

H. Bouzid and L. Ballihi. Facial expression video generation based-on spatio-temporal convolutional gan: Fev-gan. *Intell. Syst. Appl.*, 16:200139, 2022.

A. Braylan, O. Alonso, and M. Lease. Measuring annotator agreement generally across complex structured, multi-object, and free-text annotation tasks. In *Proceedings of the ACM Web Conference 2022*, pp. 1720–1730, 2022.

Z. Cai and M. Li. Invertible voice conversion with parallel data. In *ICASSP 2024*, pp. 10041–10045, 2024.

E. Cambria, X. Zhang, R. Mao, M. Chen, and K. Kwok. Senticnet 8: Fusing emotion ai and commonsense ai for interpretable, trustworthy, and explainable affective computing. In *International Conference on Human-Computer Interaction*, pp. 197–216. Springer, 2024.

F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, and A. C. Sobieranski. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, 582:593–617, 2022.

Junuk Cha, Seongro Yoon, Valeriya Strizhkova, Francois Bremond, and Seungryul Baek. Emotalkinggaussian: Continuous emotion-conditioned talking head synthesis. *arXiv preprint arXiv:2502.00654*, 2025.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

H. Chen, H. Chen, M. Yan, W. Xu, X. Gao, W. Shen, et al. Socialbench: Sociality evaluation of role-playing conversational agents. *arXiv preprint arXiv:2403.13679*, 2024.

J. Chen, J. Konrad, and P. Ishwar. Vgan-based image representation learning for privacy-preserving facial expression recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, pp. 1570–1579, 2018.

R. Chen, J. Wang, L.-C. Yu, and X. Zhang. Decoupled variational autoencoder with interactive attention for affective text generation. *Engineering Applications of Artificial Intelligence*, 123:106447, 2023a.

Y. Chen, L. Yang, Q. Chen, J. H. Lai, and X. Xie. Attention-based interactive disentangling network for instance-level emotional voice conversion. *arXiv preprint arXiv:2312.17508*, 2023b.

Z. Chen and S. Moscholios. Using prompts to guide large language models in imitating a real person's language style. *arXiv preprint arXiv:2410.03848*, 2024.

Z. Cheng, Z. Q. Cheng, J. Y. He, K. Wang, Y. Lin, Z. Lian, et al. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. In *Adv. Neural Inf. Process. Syst.*, volume 37, pp. 110805–110853, 2024.

D.-H. Cho, H.-S. Oh, S.-B. Kim, and S.-W. Lee. Emosphere++: Emotion-controllable zero-shot text-to-speech via emotion-adaptive spherical vector. *IEEE Trans. Affect. Comput.*, 2025a.

Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, and Seong-Whan Lee. Diemo-tts: Disentangled emotion representations via self-supervised distillation for cross-speaker emotion transfer in text-to-speech. *arXiv preprint arXiv:2505.19687*, 2025b.

H. Choi and M. Hahn. Sequence-to-sequence emotional voice conversion with strength control. *IEEE Access*, 9:42674–42687, 2021.

H. H. Chou, Y. S. Lin, C. C. Sung, Y. Tsao, and C. C. Lee. Toward any-to-any emotion voice conversion using disentangled diffusion framework. *arXiv preprint arXiv:2409.03636*, 2024.

M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth. Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Computing and Applications*, 35(32):23311–23328, 2023.

J. Coda-Forno, K. Witte, A. K. Jagadish, M. Binz, Z. Akata, and E. Schulz. Inducing anxiety in large language models can induce bias. *arXiv preprint arXiv:2304.11111*, 2023.

K. Cortiñas-Lorenzo and G. Lacey. Toward explainable affective computing: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

S. Corvaia, A. Pipitone, and A. Chella. Inner speech and damasio's theory for modelling robot's emotions. *IEEE Transactions on Affective Computing*, 2025.

J. Crumpton and C. L. Bethel. A survey of using vocal prosody to convey emotion in robot speech. *Int. J. Soc. Robot.*, 8:271–285, 2016.

W. Cui, D. Yu, X. Jiao, Z. Meng, G. Zhang, Q. Wang, Y. Guo, and I. King. Recent advances in speech language models: A survey. *arXiv preprint arXiv:2410.03751*, 2024.

D. Dadebayev, W. W. Goh, and E. X. Tan. Eeg-based emotion recognition: Review of commercial eeg devices and machine learning techniques. *Journal of King Saud University-Computer and Information Sciences*, 34(7):4385–4401, 2022.

W. Dai, S. Cahyawijaya, Z. Liu, and P. Fung. Multimodal end-to-end sparse model for emotion recognition. *arXiv preprint arXiv:2103.09666*, 2021.

J. Deng and F. Ren. A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*, 14(1):49–67, 2021.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

F. Di Luzio, A. Rosato, and M. Panella. An explainable fast deep neural network for emotion recognition. *Biomedical Signal Processing and Control*, 100:107177, 2025.

Yurui Dong et al. From rational answers to emotional resonance: The role of controllable emotion generation in language models. *arXiv preprint arXiv:2502.04075*, 2025.

Q. Du, S. Labat, T. Demeester, and V. Hoste. Unic: A dataset for emotion analysis of videos with multimodal and unimodal labels, 2024.

R.-N. Duan, J.-Y. Zhu, and B.-L. Lu. Differential entropy feature for eeg-based emotion classification. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 81–84. IEEE, 2013.

P. Ekman. An argument for basic emotions. *Cognit. Emotion*, 6(3-4):169–200, 1992.

M. Elgaar, J. Park, and S. W. Lee. Multi-speaker and multi-domain emotional voice conversion using factorized hierarchical variational autoencoder. In *ICASSP 2020*, pp. 7769–7773, 2020.

Mario Espinosa Gámez. Emotional intelligence: A human-centred ai perspective. In *Handbook of Human-Centered Artificial Intelligence*, pp. 1–65. Springer, 2026.

S. Evuru, Y. Chen, and Y. Liu. Coda: Constrained generation-based data augmentation for low-resource text classification. *arXiv preprint arXiv:2402.10958*, 2024.

K. Ezzameli and H. Mahersia. Emotion recognition from unimodal to multimodal analysis: A review. *Information Fusion*, 99:101847, 2023.

S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Z. Feng. *Bridging emotional gaps in textual interactions: A study on the role of emotion analysis services.* PhD thesis, Purdue Univ., 2024.

M. Firdaus, U. Jain, A. Ekbal, and P. Bhattacharyya. Seprg: Sentiment aware emotion controlled personalized response generation. In *Proc. 14th Int. Conf. Nat. Lang. Gener.*, pp. 353–363, 2021.

M. Firdaus, G. Singh, A. Ekbal, and P. Bhattacharyya. Multi-step prompting for few-shot emotion-grounded conversations. In *Proc. 32nd ACM Int. Conf. Inf. Knowl. Manage.*, pp. 3886–3891, 2023.

C. Flavian-Blanco, R. Gurrea-Sarasa, and C. Orus-Sanclemente. Analyzing the emotional outcomes of the online search behavior with search engines. *Computers in Human Behavior*, 27(1):540–551, 2011.

S. Franklin, U. Ramamurthy, S. K. D'Mello, L. McCauley, A. Negatu, R. L. Silva, and V. Datla. Lida: A computational model of global workspace theory and developmental learning. In *Proceedings of the AAAI Fall Symposium on AI and Consciousness*, 2007.

C. Fu, C. Liu, C. T. Ishi, and H. Ishiguro. Cycletransgan-evc: A cyclegan-based emotional voice conversion model with transformer. *arXiv preprint arXiv:2111.15159*, 2021.

Z. Fu, X. Tan, N. Peng, D. Zhao, and R. Yan. Style transfer in text: Exploration and evaluation. In *Proc. AAAI Conf. Artif. Intell.*, volume 32, 2018.

K. Gandhi, Z. Lynch, J. P. Fränken, K. Patterson, S. Wambu, T. Gerstenberg, D. C. Ong, and N. D. Goodman. Human-like affective cognition in foundation models. *arXiv preprint arXiv:2409.11733*, 2024.

Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Int. Conf. Mach. Learn.*, pp. 1180–1189, 2015.

J. Gao, D. Chakraborty, H. Tembine, and O. Olaleye. Nonparallel emotional speech conversion. *arXiv preprint arXiv:1811.01174*, 2018.

X. Gao, C. Zhang, Y. Chen, H. Zhang, and N. F. Chen. Emo-dpo: Controllable emotional speech synthesis through direct preference optimization. *arXiv preprint arXiv:2409.10157*, 2024.

A. Geetha, T. Mala, D. Priyanka, and E. Uma. Multimodal emotion recognition with deep learning: advancements, challenges, and future directions. *Information Fusion*, 105:102218, 2024.

J. Geng, T. Shao, Y. Zheng, Y. Weng, and K. Zhou. Warp-guided gans for single-photo facial animation. *ACM Trans. Graph. (ToG)*, 37(6):1–12, 2018.

S. Ghafourian, R. Sharifi, and A. Baniasadi. Facial emotion recognition in imbalanced datasets. *Computer Science and Information Technology*, 2022.

S. Ghosh, M. Chollet, E. Laksana, L.-P. Morency, and S. Scherer. Affect-lm: A neural language model for customizable affective text generation. *arXiv preprint arXiv:1704.06851*, 2017.

J. Gratch and S. Marsella. A domain-independent framework for modeling emotion. *Cognitive Systems Research*, 5(4):269–306, 2004.

Z. Gu and K. He. Affective prompt-tuning-based language model for semantic-based emotional text generation. *Int. J. Semantic Web Inf. Syst.*, 20(1):1–19, 2024.

D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

J. Guo, S. Lu, H. Cai, W. Zhang, Y. Yu, and J. Wang. Long text generation via adversarial training with leaked information. In *Proc. AAAI Conf. Artif. Intell.*, volume 32, 2018.

L. Guo, Y. Song, and S. Ding. Speaker-aware cognitive network with cross-modal attention for multimodal emotion recognition in conversation. *Knowledge-Based Systems*, 296:111969, 2024.

Z. Guo, Y. Leng, Y. Wu, S. Zhao, and X. Tan. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023*, pp. 1–5, 2023.

N. Hajarolasvadi, M. A. Ramirez, W. Beccaro, and H. Demirel. Generative adversarial networks in human emotion synthesis: A review. *IEEE Access*, 8:218499–218529, 2020.

B. Han, C. Yau, S. Lei, and J. Gratch. Knowledge-based emotion recognition using large language models. *arXiv preprint arXiv:2408.04123*, 2024.

H. He, Z. Shang, C. Wang, X. Li, Y. Gu, H. Hua, et al. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Lang. Technol. Workshop (SLT)*, pp. 885–890, 2024a.

J. He, Y. Geng, and L. Bo. Uniportrait: A unified framework for identity-preserving single-and multi-human image personalization. *arXiv preprint arXiv:2408.05939*, 2024b.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, 1998.

A. Hernández-Marcos and E. Ros. A generic self-learning emotional framework for machines. *Scientific Reports*, 14(1):25858, 2024.

J. F. Hoorn, T. Baier, J. A. N. van Maanen, and J. Wester. Silicon coppelia and the formalization of the affective process. *IEEE Transactions on Affective Computing*, 2021.

M. Hou, Z. Zhang, C. Liu, and G. Lu. Semantic alignment network for multi-modal emotion recognition. *IEEE Transactions on Circuits and Systems for for Video Technology*, 33(9):5318–5329, 2023.

W. N. Hsu, B. Bolte, Y. H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 29:3451–3460, 2021.

G. Hu, Y. Xin, W. Lyu, H. Huang, C. Sun, Z. Zhu, L. Gui, R. Cai, E. Cambria, and H. Seifi. Recent trends of multimodal affective computing: A survey from nlp perspective. *arXiv preprint arXiv:2409.07388*, 2024.

H.-Y. Hu, L.-M. Zhao, Y.-Z. Liu, H.-L. Li, and B.-L. Lu. A novel experiment setting for cross-subject emotion recognition. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 6416–6419. IEEE, 2021.

Y. Huang and S. M. Khan. Dyadgan: Generating facial expressions in dyadic interactions. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, pp. 11–18, 2017.

Z. Huang, W. Xu, and K. Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

Z. Huang, K. C. Chan, Y. Jiang, and Z. Liu. Collaborative diffusion for multi-modal face generation and editing. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 6080–6090, 2023.

E. Hudlicka. Beyond cognition: Modeling emotion in cognitive architectures. In *Proceedings of the 6th International Conference on Cognitive Modeling (ICCM)*, pp. 118–123, 2004.

Z. Iftikhar, S. Ransom, A. Xiao, and J. Huang. Therapy as an nlp task: Psychologists' comparison of llms and human peers in cbt. *arXiv preprint arXiv:2409.02244*, 2024.

A. Iranmehr, H. Masnadi-Shirazi, and N. Vasconcelos. Cost-sensitive support vector machines. *Neurocomputing*, 343:50–64, 2019.

V. K. Jain, S. Kumar, and S. L. Fernandes. Extraction of emotions from multilingual text using intelligent text processing and computational linguistics. *J. Comput. Sci.*, 21:316–326, 2017.

M. Jeon. Emotions and affect in human factors and human–computer interaction: taxonomy, theories, approaches, and methods. In *Emotions and Affect in Human Factors and Human-Computer Interaction*, pp. 3–26. Elsevier, 2017.

A. Jia, Y. He, Y. Zhang, S. Uprety, D. Song, and C. Lioma. Beyond emotion: A multi-modal dataset for human desire understanding. In *Proc. 2022 Conf. North Am. Chapter Assoc. Comput. Linguist.*, pp. 1512–1522, 2022.

C. Jiang, C. Zhang, Y. Ji, Z. Hu, Z. Zhan, and G. Yang. An affective chatbot with controlled specific emotion expression. *Sci. China Inf. Sci.*, 65(10):202102, 2022.

D. Jiang, D. Song, R. Tong, and M. Tang. Styleipsb: Identity-preserving semantic basis of stylegan for high fidelity face swapping. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 352–361, 2023.

H. Jiang, J. M. Vidal, and M. N. Huhns. Ebdi: An architecture for emotional agents. In *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 1–3. ACM, 2007.

Jussi PP Jokinen and Andrew Oulasvirta, Antti andand Howes. Introduction to computational cognitive modeling. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–2, 2025.

I. Juvina, O. Larue, and A. Hough. Modeling valuation and core affect in a cognitive architecture: The impact of valence and arousal on memory and decision-making. *Cognitive Systems Research*, 48:4–24, 2018.

S. Kamran, R. Zall, S. Hosseini, M. R. Kangavari, S. Rahmani, and W. Hua. Emodnn: understanding emotions from short texts through a deep neural network ensemble. *Neural Computing and Applications*, 35(18):13565–13582, 2023.

S. Karlapati, A. Moinet, A. Joly, V. Klimkov, D. Saez-Trigueros, and T. Drugman. Copycat: Many-to-many fine-grained prosody transfer for neural text-to-speech. *arXiv preprint arXiv:2004.14617*, 2020.

A. Katirai. Ethical considerations in emotion recognition technologies: a review of the literature. *AI Ethics*, 4:927–948, 2024.

P. Kaur, G. S. Kashyap, A. Kumar, M. T. Nafis, S. Kumar, and V. Shokeen. From text to transformation: A comprehensive review of large language models' versatility. *arXiv preprint arXiv:2402.16142*, 2024.

L. U. Khan, M. Guizani, S. Muhaidat, and C. S. Hong. Large language models-empowered wireless networks: Fundamentals, architecture, and challenges. *arXiv preprint arXiv:2506.10651*, 2025.

U. A. Khan, Q. Xu, Y. Liu, A. Lagstedt, A. Alamäki, and J. Kauttonen. Exploring contactless techniques in multimodal emotion recognition: insights into diverse applications, challenges, solutions, and prospects. *Multimedia Systems*, 30(3):115, 2024.

H. Kim and T. Hong. Enhancing emotion recognition using multimodal fusion of physiological, environmental, personal data. *Expert Systems with Applications*, 249:123723, 2024.

A. Koufakou, D. Grisales, O. Fox, et al. Data augmentation for emotion detection in small imbalanced text data. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pp. 1508–1513. IEEE, 2023.

S. D. Kreibig and J. J. Gross. Understanding mixed emotions: paradigms and measures. *Curr. Opin. Behav. Sci.*, 15:62–71, 2017.

C. O. Kumar, N. Gowtham, M. Zakariah, and A. Almazyad. Multimodal emotion recognition using feature fusion: An llm-based approach. *IEEE Access*, 2024.

J. E. Laird. *The Soar Cognitive Architecture*. MIT Press, 2019a.

J. E. Laird. *The Soar cognitive architecture*. MIT Press, 2019b.

A. D. L. Languré and M. Zareei. Improving text emotion detection through comprehensive dataset quality analysis. *IEEE Access*, 2024.

J. Lee, P. Sattigeri, and G. Wornell. Learning new tricks from old dogs: Multi-source transfer learning from pre-trained networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Y. Lee, A. Rabiee, and S.-Y. Lee. Emotional end-to-end neural speech synthesizer. *arXiv preprint arXiv:1711.05447*, 2017.

Y. Lei, S. Yang, X. Wang, and L. Xie. Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 30:853–864, 2022.

J. Li. A comparative study on annotation quality of crowdsourcing and llm via label aggregation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6525–6529. IEEE, 2024.

J. Li, Y. Li, S. Li, M. Sun, Y. Zhou, J.-C. Zhang, and Z. Ren. Emotion-aware dialogue systems: A survey. *arXiv preprint arXiv:2305.01258*, 2023a.

J. Li, X. Wang, G. Lv, and Z. Zeng. Graphmft: A graph network based multimodal fusion technique for emotion recognition in conversation. *Neurocomputing*, 550:126427, 2023b.

J. Li, J. Chen, R. Ren, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen. The dawn after the dark: An empirical study on factuality hallucination in large language models. *arXiv preprint arXiv:2401.03205*, 2024a.

S. Li and W. Deng. Real world expression recognition: A highly imbalanced detection problem. In *2016 International Conference on Biometrics (ICB)*, pp. 1–6. IEEE, 2016.

T. Li, X. Wang, Q. Xie, Z. Wang, and L. Xie. Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 30:1448–1460, 2022a.

X. Li, K. Song, S. Feng, D. Wang, and Y. Zhang. A co-attention neural network model for emotion cause analysis with emotional context awareness. In *Proc. 2018 Conf. Empir. Methods Nat. Lang. Process.*, pp. 4752–4757, 2018a.

X. Li, J. Thickstun, I. Gulrajani, P. Liang, and T. Hashimoto. Diffusion-lm improves controllable text generation. *arXiv preprint arXiv:2205.14217*, 2022b.

X. Li, Z.-Q. Cheng, J.-Y. He, J. Chen, X. Fan, X. Peng, and A. G. Hauptmann. Umetts: A unified framework for emotional text-to-speech synthesis with multimodal prompts. In *ICASSP 2025*, pp. 1–5, 2025a.

Y. Li, Q. Pan, S. Wang, T. Yang, and E. Cambria. A generative model for category text generation. *Inf. Sci.*, 450:301–315, 2018b.

Y. Li, Q. Sun, M. Schlicher, Y. W. Lim, and B. W. Schuller. Artificial emotion: A survey of theories and debates on realising emotion in artificial intelligence. *arXiv preprint arXiv:2508.10286*, 2025b.

Y. A. Li, A. Zare, and N. Mesgarani. Starganv2-vc: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion. *arXiv preprint arXiv:2107.10394*, 2021.

Z. Li, G. Chen, R. Shao, D. Jiang, and L. Nie. Enhancing the emotional generation capability of large language models via emotional chain-of-thought. *arXiv preprint arXiv:2401.06836*, 2024b.

Z. Lian, H. Sun, L. Sun, J. Yi, B. Liu, and J. Tao. Affectgpt: Dataset and framework for explainable multimodal emotion recognition. *arXiv preprint arXiv:2407.07653*, 2024.

J. Liang and F. Lu. Emotional conversation: Empowering talking faces with cohesive expression, gaze and pose generation. *arXiv preprint arXiv:2406.07895*, 2024.

B. Lin, Y. Yu, J. Ye, R. Lv, Y. Yang, R. Xie, et al. Takin-ada: Emotion controllable audio-driven animation with canonical and landmark loss optimization. *arXiv preprint arXiv:2410.14283*, 2024.

Chang Liu, Ye Pan, Chenyang Ding, Susanto Rahardja, and Xiaokang Yang. Medtalk: Multimodal controlled 3d facial animation with dynamic emotions by disentangled embedding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 7538–7547, 2025.

R. Liu, B. Sisman, and H. Li. Reinforcement learning for emotional text-to-speech synthesis with improved emotion discriminability. *arXiv preprint arXiv:2104.01408*, 2021.

Rui Liu, Yifan Hu, Yi Ren, Xiang Yin, and Haizhou Li. Emotion rendering for conversational speech synthesis with heterogeneous graph-based context modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18698–18706, 2024.

S. Liu, Y. Cao, and H. Meng. Emotional voice conversion with cycle-consistent adversarial network. *arXiv preprint arXiv:2004.03781*, 2020.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L. P. Morency. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv iv preprint arXiv:1806.00064*, 2018.

L. Lo, B.-K. Ruan, H.-H. Shuai, and W.-H. Cheng. Modeling uncertainty for low-resolution facial expression recognition. *IEEE Transactions on Affective Computing*, 15(1):198–209, 2023.

R. Lotfian and C. Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Trans. Affect. Comput.*, 10(4):471–483, 2019.

H. Lu, X. Niu, J. Wang, Y. Wang, Q. Hu, J. Tang, Y. Zhang, K. Yuan, B. Huang, Z. Yu, et al. Gpt as psychologist? preliminary evaluations for gpt-4v on visual affective computing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 322–331, 2024.

X. Lu, S. Welleck, J. Hessel, L. Jiang, L. Qin, P. West, K. Cho, and Y. Choi. What makes dialogue representations learnable? *arXiv preprint arXiv:2210.16227*, 2022.

X. Luo, S. Takamichi, Y. Saito, T. Koriyama, H. Saruwatari, et al. Emotion-controllable speech synthesis using emotion soft label, utterance-level prosodic factors, and word-level prominence. *APSIPA Trans. Signal Inf. Process.*, 13(1), 2024.

Chuhang Ma, Shuai Tan, Ye Pan, Jiaolong Yang, and Xin Tong. Esgaussianface: Emotional and stylized audio-driven facial animation via 3d gaussian splatting. *IEEE Transactions on Visualization and Computer Graphics*, 2026.

F. Ma, Y. Li, Y. Xie, Y. He, Y. Zhang, H. Ren, et al. A review of human emotion synthesis based on generative technology. *arXiv preprint arXiv:2412.07116*, 2024.

F. Ma, Y. Xie, Y. Li, Y. He, Y. Zhang, H. Ren, Z. Liu, et al. A review of human emotion synthesis based on generative technology. *IEEE Transactions on Affective Computing*, 2025.

A. F. Mahmood and C. D. Manning. Dexperts: Decoding-time controlled text generation with experts and anti-experts. *arXiv preprint arXiv:2305.01258*, 2023.

G. McDarby, J. Condron, D. Hughes, and N. Augenblick. Affective feedback. In *Proc. 2003 Conf. Universal Usability*, pp. 136–142, 2003.

A. H. Meftah, A. A. Alashban, Y. A. Alotaibi, and S. A. Selouani. English emotional voice conversion using stargan model. *IEEE Access*, 11:67835–67849, 2023.

T. Meng, Y. Shou, W. Ai, N. Yin, and K. Li. Deep imbalanced learning for multimodal emotion recognition in conversations. *IEEE Transactions on Artificial Intelligence*, 2024.

A. I. Middya, B. Nag, and S. Roy. Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. *Knowledge-Based Systems*, 244:108580, 2022.

C. Mishra, R. Verdonschot, P. Hagoort, and G. Skantze. Real-time emotion generation in human-robot dialogue using large language models. *Front. Robot. AI*, 10:1271610, 2023.

T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 1359–1367, 2020.

R. Mobbs, D. Makris, and V. Argyriou. Emotion recognition and generation: A comprehensive review of face, speech, and text modalities. *arXiv preprint arXiv:2502.06803*, 2025.

T. M. Moerland, J. Broekens, and C. M. Jonker. Emotion in reinforcement learning agents and robots: A survey. *Machine Learning*, 107(2):443–480, 2018.

S. M. Mohammad et al. Semeval-2025 task 11: Bridging the gap in text-based emotion detection. In *Proc. 2024 Conf. Empir. Methods Nat. Lang. Process.*, pp. 20939–20962, 2025.

G. Mohammadi and P. Vuilleumier. A multi-componential approach to emotion recognition and the effect of personality. *IEEE Transactions on Affective Computing*, 13(3):1127–1139, 2022.

D. Nandini, J. Yadav, V. Singh, V. Mohan, and S. Agarwal. An ensemble deep learning framework for emotion recognition through wearable devices multi-modal physiological signals. *Scientific Reports*, May 2025.

S. Nayak, B. Nagesh, A. Routray, and M. Sarma. A human–computer interaction framework for emotion recognition through time-series thermal video sequences. *Computers & Electrical Engineering*, 93:107280, 2021.

D. Nguyen, K. Nguyen, S. Sridharan, A. Ghasemi, D. Dean, and C. Fookes. Deep spatio-temporal features for multimodal emotion recognition. In *2017 IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 1215–1223, 2017.

F. L. Nijeholt and J. Broekens. The role of simulated emotions in reinforcement learning: Insights from a human-robot interaction experiment. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 1–7, Sep 2023.

Andriani Nikodemou and Chris Christodoulou. Deconstructing emotions in self-control through computational modeling. *Cognitive Systems Research*, 88:101294, 2024.

F. Nocentini, C. Ferrari, and S. Berretti. Emovoca: Speech-driven emotional 3d talking heads. In *2025 IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 2859–2868, 2025.

B. Nojavanasghari, Y. Huang, and S. Khan. Interactive generative adversarial networks for facial expression generation in dyadic interactions. *arXiv preprint arXiv:1801.09092*, 2018.

H. S. Oh, S. H. Lee, D. H. Cho, and S. W. Lee. Durflex-evc: Duration-flexible emotional voice conversion with parallel generation. *arXiv preprint arXiv:2401.08095*, 2024.

S. Oh and D.-K. Kim. Noise-robust deep learning model for emotion classification using facial expressions. *IEEE Access*, 2024.

S. Ojha, J. Vitale, and M. A. Williams. Computational emotion models: A thematic review. *International Journal of Social Robotics*, pp. 1–27, 2020a.

S. Ojha, J. Vitale, and M. A. Williams. Eegs: A transparent model of emotions. *arXiv preprint arXiv:2011.02573*, 2020b.

S. Ojha, J. Vitale, and M. A. Williams. Computational emotion models: A thematic review. *International Journal of Social Robotics*, 13(6):1253–1279, 2021.

OpenAI. Gpt-4 technical report. Technical report, 2023. [Online]. Available: https://openai.com/research/gpt-4.

K. O'Shea and R. Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

N. Otberdout, M. Daoudi, A. Kacem, L. Ballihi, and S. Berretti. Dynamic facial expression generation on hilbert hypersphere with conditional wasserstein generative adversarial nets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):848–863, 2020.

N. Otberdout, C. Ferrari, M. Daoudi, S. Berretti, and A. Del Bimbo. Sparse to dense dynamic 3d facial expression generation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 20385–20394, 2022.

A. Pandey and D. K. Vishwakarma. Target-dependent multimodal sentiment analysis via employing visual-to emotional-caption translation network using visual-caption pairs. *arXiv preprint arXiv:2408.10248*, 2024.

P. Pataranutaporn, V. Danry, J. Leong, P. Punpongsanon, D. Novy, P. Maes, and M. Sra. Ai-generated characters for supporting personalized learning and well-being. *Nat. Mach. Intell.*, 3(12):1013–1022, 2021.

D. Pereira, E. Oliveira, N. Moreira, and L. Sarmento. Towards an architecture for emotional bdi agents. In *2005 Portuguese Conference on Artificial Intelligence*, pp. 40–46. IEEE, 2005a.

D. Pereira, E. Oliveira, N. Moreira, and L. Sarmento. Towards an architecture for emotional bdi agents. In *2005 Portuguese Conference on Artificial Intelligence*, pp. 40–46. IEEE, 2005b.

J. Perez, Y. Sanchez, F. J. Seron, and E. Cerezo. Interacting with a semantic affective eca. In *International Conference on Intelligent Virtual Agents*, pp. 374–384. Springer, 2017.

H. X. Pham, S. Cheung, and V. Pavlovic. Speech-driven 3d facial animation with implicit emotional awareness: A deep learning approach. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, pp. 80–88, 2017.

R. W. Picard. *Affective computing*. MIT press, 1997.

R. W. Picard. *Affective Computing*. MIT Press, 2000.

J. Pittermann, A. Pittermann, and W. Minker. Handling emotions in human-computer dialogues. In *Handling Emotions in Human-Computer Dialogues*, pp. 19–42. 2010.

R. Plutchik. A general psychoevolutionary theory of emotion. In *Theories of Emotion*, pp. 3–33. 1980.

R. Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *Amer. Sci.*, 89(4):344–350, 2001.

N. R. Prabhu, B. Lay, S. Welker, N. Lehmann-Willenbrock, and T. Gerkmann. Emoconv-diff: Diffusion-based speech emotion conversion for non-parallel and in-the-wild data. In *ICASSP 2024*, pp. 11651–11655, 2024.

A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 818–833, 2018.

J. Qian, L. Dong, Y. Shen, F. Wei, and W. Chen. Controllable natural language generation with contrastive prefixes. *arXiv preprint arXiv:2202.13257*, 2022.

F. Qiao, N. Yao, Z. Jiao, Z. Li, H. Chen, and H. Wang. Geometry-contrastive gan for facial expression transfer. *arXiv preprint arXiv:1802.01822*, 2018.

L. Raggioli, A. Rossi, and S. Rossi. Computational models of cognitive and affective theory of mind. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, pp. 136–143. ACM, June 2025.

S. Rahmani, S. Hosseini, R. Zall, M. R. Kangavari, S. Kamran, and W. Hua. Transfer-based adaptive tree for multimodal sentiment analysis based on user latent aspects. *Knowledge-Based Systems*, 261:110219, 2023.

V. Rajan, A. Brutti, and A. Cavallaro. Is cross-attention preferable to self-attention for multi-modal emotion recognition? In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4693–4697. IEEE, 2022.

Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Y. Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

Y. M. Resendiz and R. Klinger. Affective natural language generation of event descriptions through fine-grained appraisal conditions. 2023.

Y. M. Resendiz and R. Klinger. Mopo: Multi-objective prompt optimization for affective text generation. *arXiv preprint arXiv:2412.12948*, 2024.

G. Retsinas, P. P. Filntisis, R. Danecek, V. F. Abrevaya, A. Roussos, T. Bolkart, and P. Maragos. 3d facial expressions through analysis-by-neural-synthesis. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 2490–2501, 2024.

G. Rizos, A. Baird, M. Elliott, and B. Schuller. Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition. In *ICASSP 2020*, pp. 3502–3506, 2020.

L. F. Rodriguez, J. O. Gutierrez-Garcia, and F. Ramos. Modeling the interaction of emotion and cognition in autonomous agents. *Biologically Inspired Cognitive Architectures*, 17:57–70, 2016.

K. Ross, P. Hungler, and A. Etemad. Unsupervised multi-modal representation learning for affective computing with multi-corpus wearable data. *Journal of Ambient Intelligence and Humanized Computing*, 14 (4):3199–3224, 2023.

S. Sabour, S. Liu, Z. Zhang, J. Liu, J. Zhou, A. Sunaryo, T. Lee, R. Mihalcea, and M. Huang. Emobench: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071*, 2024.

P. Sahoo, P. Meharia, A. Ghosh, S. Saha, V. Jain, and A. Chadha. A comprehensive survey of hallucination in large language, image, video and audio foundation models. *arXiv preprint arXiv:2405.09589*, 2024.

Y. Sanchez, T. Coma, A. Aguelo, and E. Cerezo. Abc-ebdi: An affective framework for bdi agents. *Cognitive Systems Research*, 58:195–216, 2019a.

Y. Sanchez, T. Coma, A. Aguelo, and E. Cerezo. Abc-ebdi: An affective framework for bdi agents. *Cognitive Systems Research*, 58:195–216, 2019b.

Y. Sanchez, T. Coma, A. Aguelo, and E. Cerezo. Applying a psychotherapeutic theory to the modeling of affective intelligent agents. *IEEE Transactions on Cognitive and Developmental Systems*, 12(2):285–299, 2019c.

A. Sarkar, P. R. Behera, and J. Shukla. Multi-source transfer learning for facial emotion recognition using multivariate correlation analysis. *Scientific Reports*, 13(1):21004, 2023.

J. Saunders and V. Namboodiri. Read avatars: Realistic emotion-controllable audio driven avatars. *arXiv preprint arXiv:2303.00744*, 2023.

R. M. Schmidt. Recurrent neural networks (rnns): A gentle introduction and overview. *arXiv preprint arXiv:1912.05911*, 2019.

B. Schnell, G. Huybrechts, B. Perz, T. Drugman, and J. Lorenzo-Trueba. Emocat: Language-agnostic emotional voice conversion. In *11th ISCA Speech Synthesis Workshop*, 2021.

B. Schuller, A. Mallol-Ragolta, A. P. Almansa, I. Tsangko, M. M. Amin, A. Semertzidou, L. Christ, and S. Amiriparian. Affective computing has changed: The foundation model disruption. *arXiv preprint arXiv:2409.08907*, 2024.

D. Schuller and B. W. Schuller. The age of artificial emotional intelligence. *Computer*, 51(9):38–46, 2018.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

S. Sharma, S. Ramaneswaran, M. S. Akhtar, and T. Chakraborty. Emotion-aware multimodal fusion for meme emotion detection. *IEEE Transactions on Affective Computing*, 2024.

M. Sheikhan, D. Gharavian, and F. Ashoftedel. Using dtw neural–based mfcc warping to improve emotional speech recognition. *Neural Comput. Appl.*, 21:1765–1773, 2012.

Xuli Shen, Hua Cai, Dingding Yu, Weilin Shen, Qing Xu, and Xiangyang Xue. Emohead: Emotional talking head via manipulating semantic expression parameters. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025.

E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng. Societal biases in language generation: Progress and challenges. *AI Mag.*, 42(3):49–61, 2021.

Y. Shirahata, R. Yamamoto, E. Song, R. Terashima, J. M. Kim, and K. Tachibana. Period vits: Variational inference with explicit pitch modeling for end-to-end emotional speech synthesis. In *ICASSP 2023*, pp. 1–5, 2023.

H.-Y. Shum, X.-D. He, and D. Li. From eliza to xiaoice: Challenges and opportunities with social chatbots. *Front. Inf. Technol. Electron. Eng.*, 19:10–26, 2018.

R. J. Siddiqui. Explore the expression: Facial expression generation using auxiliary classifier generative adversarial network. *arXiv e-prints*, arXiv:2201, 2022.

A. Sigurgeirsson and S. King. Controllable speaking styles using a large language model. In *ICASSP 2024*, pp. 10851–10855, 2024.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

I. Singh, A. Barkati, T. Goswamy, and A. Modi. Adapting a language model for controlled affective text generation. 2020.

B. Sisman, J. Yamagishi, S. King, and H. Li. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 29:132–157, 2020.

G. Soman, M. V. Judy, and S. Madria. Regret emotion based reinforcement learning for path planning in autonomous agents. In *2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 266–274. IEEE, 2024.

R. Somarathna and G. Mohammadi. Exploring emotions in multi-componential space using interactive vr games. *arXiv preprint arXiv:2404.03239*, 2024.

R. Somarathna, T. Bednarz, and G. Mohammadi. Virtual reality for emotion elicitation–a review. *IEEE Transactions on Affective Computing*, 14(4):2626–2645, 2022.

R. Somarathna, P. Vuilleumier, and G. Mohammadi. Emostim: A database of emotional film clips with discrete and componential assessment. *IEEE Transactions on Affective Computing*, 15(3):1202–1212, 2023.

B. C. Stahl, D. Schroeder, and R. Rodrigues. *Ethics of artificial intelligence: Case studies and options for addressing ethical challenges*. Springer Nature, 2023.

Xiaosu Su, Bowen Yang, Xiaowei Yi, and Yun Cao. Diffemotionvc: A dual-granularity disentangled diffusion framework for any-to-any emotional voice conversion. *Proceedings of the Interspeech, Rotterdam, The Netherlands*, pp. 17–21, 2025.

Y. Su, T. Lan, Y. Liu, F. Liu, D. Yogatama, Y. Wang, L. Kong, and N. Collier. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*, 2022.

A. N. Tak and J. Gratch. Gpt-4 emulates average-human emotional cognition from a third-person perspective. *ArXiv*, 2024.

K. L. Tan, C. P. Lee, and K. M. Lim. A survey of sentiment analysis: Approaches, datasets, and future research. *Appl. Sci.*, 13(7):4550, 2023.

S. Tan, B. Ji, M. Bi, and Y. Pan. Edtalk: Efficient disentanglement for emotional talking head synthesis. In *Eur. Conf. Comput. Vis.*, pp. 398–416, 2024.

Weipeng Tan, Chuming Lin, Chengming Xu, FeiFan Xu, Xiaobin Hu, Xiaozhong Ji, Junwei Zhu, Chengjie Wang, and Yanwei Fu. Disentangle identity, cooperate emotion: Correlation-aware emotional talking portrait generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9987–9995, 2025.

H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao. Qi-tts: Questioning intonation control for emotional speech synthesis. In *ICASSP 2023*, pp. 1–5, 2023a.

H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao. Emomix: Emotion mixing via diffusion models for emotional speech synthesis. *arXiv preprint arXiv:2306.00648*, 2023b.

J. Taverner, E. Vivancos, and V. Botti. A fuzzy appraisal model for affective agents adapted to cultural environments using the pleasure and arousal dimensions. *Information Sciences*, 546:74–86, 2021.

M. K. Tellamekala, S. Amiriparian, B. W. Schuller, E. André, T. Giesbrecht, and M. Valstar. Cold fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Anjali Thakur and Gaurav Gupta. Emotion ai: Challenges and future directions. In *Emotion and Facial Recognition in Artificial Intelligence: Sustainable Multidisciplinary Perspectives and Applications*, pp. 349–364. Springer, 2026.

A. Triantafyllopoulos, B. W. Schuller, G. ?ymen, M. Sezgin, X. He, Z. Yang, et al. An overview of affective speech synthesis and conversion in the deep learning era. *Proc. IEEE*, 111(10):1355–1381, 2023.

V. Truong. Textual emotion detection–a systematic literature review. 2024.

M. Tsfasman, R. Ghorbani, C. M. Jonker, and B. Dudzik. The emotion-memory link: Do memorability annotations matter for intelligent systems? *arXiv preprint arXiv:2507.14084*, 2025.

S. Tu, Z. Xing, X. Han, Z. Q. Cheng, Q. Dai, C. Luo, and Z. Wu. Stableanimator: High-quality identity-preserving human image animation. *arXiv preprint arXiv:2411.17697*, 2024.

M. Umair, N. Rashid, U. S. Khan, A. Hamza, and J. Iqbal. Emotion fusion-sense (emo fu-sense)–a novel multimodal emotion classification technique. *Biomedical Signal Processing and Control*, 94:106224, 2024.

H. Uyanık, S. T. A. Ozcelik, Z. B. Duranay, A. Sengur, and U. R. Acharya. Use of differential entropy for automated emotion recognition in a virtual reality environment with eeg signals. *Diagnostics*, 12(10):2508, 2022.

T. Varanka, H. Q. Khor, Y. Li, M. Wei, H. Kung, N. Sebe, and G. Zhao. Towards localized fine-grained control for facial expression generation. *arXiv preprint arXiv:2407.20175*, 2024.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Anh H Vo, Tae-Seok Kim, Hulin Jin, Soo-Mi Choi, and Yong-Guk Kim. Instruction-driven 3d facial expression generation and transition. *IEEE Transactions on Multimedia*, 2025.

K. Vougioukas, S. Petridis, and M. Pantic. Realistic speech-driven facial animation with gans. *Int. J. Comput. Vis.*, 128(5):1398–1413, 2020.

T. Walczyna and Z. Piotrowski. Overview of voice conversion methods based on deep learning. *Appl. Sci.*, 13(5):3100, 2023.

H. Wang, X. Du, J. Li, R. A. Yeh, and G. Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 12619–12629, 2023.

H. Wang, Y. Weng, Y. Li, Z. Guo, J. Du, S. Niu, et al. Emotivetalk: Expressive talking head generation through audio information decoupling and emotional video diffusion. *arXiv preprint arXiv:2411.16726*, 2024.

K. Wang and X. Wan. Sentigan: Generating sentimental texts via mixture adversarial networks. In *IJCAI*, pp. 4446–4452, 2018.

L. Wang, J. Li, C. Yang, Z. Lin, H. Tang, H. Liu, et al. Sibyl: Empowering empathetic dialogue generation in large language models via sensible and visionary commonsense inference. In *Proc. 31st Int. Conf. Comput. Linguist.*, pp. 123–140, 2025a.

Yunxiao Wang et al. Compeer: Controllable empathetic reinforcement reasoning for emotional support conversation. *arXiv preprint arXiv:2508.09521*, 2025b.

Zhongzheng Wang, Yuanhe Tian, Hongzhi Wang, and Yan Song. Explainable multimodal aspect-based sentiment analysis with dependency-guided large language model. *arXiv preprint arXiv:2601.06848*, 2026.

T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah. A comprehensive review of speech emotion recognition systems. *IEEE Access*, 9:47795–47814, 2021.

P. Washington, H. Kalantarian, J. Kent, A. Husic, A. Kline, E. Leblanc, et al. Training affective computer vision models by crowdsourcing soft-target labels. *Cognit. Comput.*, 13:1363–1373, 2021.

J. Wei and K. Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

Q. Wei, X. Huang, and Y. Zhang. Fv2es: A fully end2end multimodal system for fast yet effective video emotion recognition inference. *IEEE Transactions on Broadcasting*, 69(1):10–20, 2022.

O. Wiles, A. Koepke, and A. Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 670–686, 2018.

C. Wognum, J. R. Ash, M. Aldeghi, R. Rodriguez-Perez, C. Fang, A. C. Cheng, D. J. Price, D.-A. Clevert, O. Engkvist, and W. P. Walters. A call for an industry-led initiative to critically assess machine learning for real-world drug discovery. *Nature Machine Intelligence*, 6:1120–1121, October 2024.

J. Wu. Social and ethical impact of emotional ai advancement: the rise of pseudo-intimacy relationships and challenges in human interactions. *Front. Psychol.*, 15:1410462, 2024.

J. Wu and Y. Sun. An automated negotiation model based on agents' attribute preference with emotional deception. *Expert Systems with Applications*, 278:127448, 2025.

Y. Wu, S. Zhang, and P. Li. Multi-modal emotion recognition in conversation based on prompt learning with text-audio fusion features. *Scientific Reports*, 15(1):8855, 2025.

R. Xia and Z. Ding. Emotion-cause pair extraction: A new task to emotion analysis in texts. *arXiv preprint arXiv:1906.01267*, 2019.

Tianxin Xie, Shan Yang, Chenxing Li, Dong Yu, and Li Liu. Emosteer-tts: Fine-grained and training-free emotion-controllable text-to-speech via activation steering. *arXiv preprint arXiv:2508.03543*, 2025.

H. Xu, M. Zhang, H. Ju, Z. Zheng, H. Zhu, E. Cambria, et al. Towards rich emotions in 3d avatars: A text-to-3d avatar generation benchmark. *arXiv preprint arXiv:2412.02508*, 2024a.

J. Xu, S. Li, J. Wu, M. Xiong, A. Deng, J. Ji, et al. ID^3: Identity-preserving-yet-diversified diffusion models for synthetic face recognition. In *Adv. Neural Inf. Process. Syst.*, volume 37, pp. 77777–77798, 2024b.

S. Xu, G. Chen, Y. X. Guo, J. Yang, C. Li, Z. Zang, et al. Vasa-1: Lifelike audio-driven talking faces generated in real time. In *Adv. Neural Inf. Process. Syst.*, volume 37, pp. 660–684, 2024c.

Z. Xu, J. Zhang, J. H. Liew, H. Yan, J. W. Liu, C. Zhang, et al. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 1481–1490, 2024d.

H. Xue, Y. Liang, B. Mu, S. Zhang, M. Chen, Q. Chen, and L. Xie. E-chat: Emotion-sensitive spoken dialogue system with large language models. In *Proc. IEEE 14th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, pp. 586–590, 2024.

L. Xue, N. Constant, A. Roberts, M. Kale, A. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

D. Yang, Z. Chen, Y. Wang, S. Wang, M. Li, S. Liu, X. Zhao, S. Huang, Z. Dong, P. Zhai, et al. Context de-confounded emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19005–19015, 2023.

Jianing Yang, Sheng Li, Takahiro Shinozaki, Yuki Saito, and Hiroshi Saruwatari. Emotional text-to-speech based on mutual-information-guided emotion-timbre disentanglement. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 567–572. IEEE, 2025.

K. Yang and D. Klein. Fudge: Controlled text generation with future discriminators. *arXiv preprint arXiv:2104.05218*, 2021.

W. Ye, Z. Zhang, F. Teng, M. Zhang, J. Wang, D. Ni, F. Li, P. Xu, and Z. Liang. Semi-supervised dual-stream self-attentive adversarial graph contrastive learning for cross-subject eeg-based emotion recognition. *IEEE Transactions on Affective Computing*, 2024.

Phyo Thet Yee, Dimitrios Kollias, Sudeepta Mishra, and Abhinav Dhall. Synchrorama: Lip-synchronized and emotion-aware talking face generation via multi-modal emotion embedding. *arXiv preprint arXiv:2509.19965*, 2025.

F. Yin, Y. Zhang, X. Cun, M. Cao, Y. Fan, X. Wang, et al. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *Eur. Conf. Comput. Vis.*, pp. 85–101, 2022.

H.-W. Yoon, O. Kwon, H. Lee, R. Yamamoto, E. Song, J.-M. Kim, and M.-J. Hwang. Language model-based emotion prediction methods for emotional speech synthesis systems. *arXiv preprint arXiv:2206.15067*, 2022.

Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5812–5823, 2020.

E. M. Younis, S. M. Zaki, E. Kanjo, and E. H. Houssein. Evaluating ensemble learning methods for multi-modal emotion recognition using sensor data fusion. *Sensors*, 22(15):5611, 2022.

Jiahao Yuan, Zhiqing Cui, Hanqing Wang, Yuansheng Gao, Yucheng Zhou, and Usman Naseem. Kardia-r1: Unleashing llms to reason toward understanding and empathy for emotional support via rubric-as-judge reinforcement learning. *arXiv preprint arXiv:2512.01282*, 2025.

Raziyeh Zall and Mohammad Reza Kangavari. Comparative analytical survey on cognitive agents with emotional intelligence. *Cognitive Computation*, 14(4):1223–1246, 2022.

Raziyeh Zall and Mohammad Reza Kangavari. Towards emotion-aware intelligent agents by utilizing knowledge graphs of experiences. *Cognitive Systems Research*, 88:101285, 2024.

S. Zanwar, D. Wiechmann, Y. Qiao, and E. Kerz. Improving the generalizability of text-based emotion detection by leveraging transformers with psycholinguistic features. *arXiv preprint arXiv:2212.09465*, 2022.

F. Zhan, Y. Yu, R. Wu, J. Zhang, S. Lu, L. Liu, et al. Multimodal image synthesis and editing: A survey. *arXiv preprint arXiv:2112.13592*, 2022.

Chao Zhang, Xin Shi, Xueqiao Zhang, Yifan Zhu, Yi Yang, and Yawei Luo. Decoupledesc: Enhancing emotional support generation via strategy-response decoupled preference optimization. *arXiv preprint arXiv:2505.16995*, 2025a.

Chenxu Zhang, Zenan Li, Hongyi Xu, You Xie, Xiaochen Zhao, Tianpei Gu, Guoxian Song, Xin Chen, Chao Liang, Jianwen Jiang, et al. X-actor: Emotional and expressive long-range portrait acting from audio. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025b.

F. Zhang, J. Chen, Q. Tang, and Y. Tian. Evaluation of emotion classification schemes in social media text: an annotation-based approach. *BMC Psychol.*, 12(1):503, 2024a.

G. Zhang, Y. Qin, W. Zhang, J. Wu, M. Li, Y. Gai, et al. iemotts: Toward robust cross-speaker emotion transfer and control for speech synthesis based on disentanglement between prosody and timbre. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 31:1693–1705, 2023a.

H. Zhang, H. Song, S. Li, M. Zhou, and D. Song. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Comput. Surv.*, 56(3):1–37, 2023b.

J. E. Zhang, B. Hilpert, J. Broekens, and J. P. P. Jokinen. Simulating emotions with an integrated computational model of appraisal and reinforcement learning. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2024b.

Jian Zhang, Weijian Mai, and Zhijun Zhang. Emodiffhead: continuously emotional control in talking head generation via diffusion. *arXiv preprint arXiv:2409.07255*, 2024c.

Jiayi Eurus Zhang, Joost Broekens, and Jussi Jokinen. Modeling cognitive-affective processes with appraisal and reinforcement learning. *IEEE Transactions on Affective Computing*, 2024d.

S. Zhang, A. Mehrish, Y. Li, and S. Poria. Proemo: Prompt-driven text-to-speech synthesis based on emotion and intensity control. *arXiv preprint arXiv:2501.06276*, 2025c.

T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In *Proc. 2017 ACM Multimedia Conf.*, pp. 223–231, 2017.

Y. Zhang, M. Wang, Y. Wu, P. Tiwari, Q. Li, B. Wang, and J. Qin. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. *arXiv preprint arXiv:2310.11374*, 2023c.

Y. Zhang, X. Yang, X. Xu, Z. Gao, Y. Huang, S. Mu, S. Feng, et al. Affective computing in the era of large language models: A survey from the nlp perspective. *arXiv preprint arXiv:2408.04638*, 2024e.

Q. Zhao, Y. Xia, Y. Long, G. Xu, and J. Wang. Leveraging sensory knowledge into text-to-text transfer transformer for enhanced emotion analysis. *Information Processing & Management*, 62(1):103876, 2025.

Y. Zheng, Y. Wang, P. Ke, Z. Yang, and M. Huang. Semantic-enhanced explainable finetuning for open-domain dialogues. *arXiv preprint arXiv:2106.03065*, 2021.

H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proc. AAAI Conf. Artif. Intell.*, volume 32, 2018.

K. Zhou, B. Sisman, and H. Li. Transforming spectrum and prosody for emotional voice conversion with non-parallel training data. *arXiv preprint arXiv:2002.00198*, 2020a.

K. Zhou, B. Sisman, M. Zhang, and H. Li. Converting anyone's emotion: Towards speaker-independent emotional voice conversion. *arXiv preprint arXiv:2005.07025*, 2020b.

K. Zhou, B. Sisman, and H. Li. Vaw-gan for disentanglement and recomposition of emotional elements in speech. In *2021 IEEE Spoken Lang. Technol. Workshop (SLT)*, pp. 415–422, 2021a.

K. Zhou, B. Sisman, R. Liu, and H. Li. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021*, pp. 920–924, 2021b.

K. Zhou, B. Sisman, C. Busso, and H. Li. Mixed emotion modelling for emotional voice conversion. *Computer*, 6(7), 2022a.

K. Zhou, B. Sisman, R. Liu, and H. Li. Emotional voice conversion: Theory, databases and esd. *Speech Commun.*, 137:1–18, 2022b.

K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li. Emotion intensity and its control for emotional voice conversion. *IEEE Trans. Affect. Comput.*, 14(1):31–48, 2022c.

K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li. Speech synthesis with mixed emotions. *IEEE Trans. Affect. Comput.*, 14(4):3120–3134, 2022d.

X. Zhou and W. Y. Wang. Mojitalk: Generating emotional responses at scale. *arXiv preprint arXiv:1711.04090*, 2017.

Jie Zhu et al. Care: Cognitive-reasoning augmented reinforcement for emotional support conversation. *arXiv preprint arXiv:2510.05122*, 2025.