Beyond the Face: A Taxonomy of Situation Context Cues in Audio-Visual Emotion Perception

Teodor Neagoe
Supervisor(s): Bernd Dudzik, Sayak Mukherjee
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Teodor Neagoe
Final project course: CSE3000 Research Project
Thesis committee: Bernd Dudzik, Sayak Mukherjee, Stephanie Tan

Abstract

Emotion recognition (ER) systems frequently fail in real-world settings because they are designed to be context-blind, relying solely on isolated facial or vocal expressions. While incorporating "context" is widely recognized as essential, the field lacks a structured, unified definition, often conflating situational factors with the target's body language or the observer's internal biases. This paper addresses this gap by establishing a comprehensive, non-overlapping taxonomy of situation context. Through a systematic literature review of empirical human studies, we disentangle situation context from observer and actor confounds, categorizing it into three distinct pillars: (1) the Objective Environment (physical and spatiotemporal surroundings), (2) the Social Array (the presence and actions of others), and (3) Semantic/Narrative Information (textual or verbal event descriptions). We map how cues within these categories systematically bias emotion judgments—such as shifting categorical labels or altering reaction times. By structuring the chaotic landscape of context, this taxonomy provides a concrete framework for designing context-aware, ecologically valid affective computing models.

1 Introduction

Consider a person crying: in isolation, we naturally interpret this facial expression as sadness, and assume a trembling, high-pitched voice signals fear. However, when the view expands to reveal a wedding altar or a lottery ticket, the tears and shaky voice are suddenly understood as intense joy. Conversely, if the context reveals a funeral, our initial interpretation of sadness is confirmed. These immediate environmental and narrative factors represent the *situation context*—which we define strictly as the external, dynamic circumstances embedding the expresser, structured across spatiotemporal, social, and narrative dimensions. Put simply, this context is composed of where and when the event takes place (spatiotemporal), who else is present and how they are behaving (social), and the story or sequence of events that led up to this moment (narrative). Current computational Emotion Recognition (ER) systems operate under a form of "context-blindness," attempting to classify human affect solely from isolated facial movements or vocal features [1]. Because behavioral signals are inherently ambiguous in isolation [2, 3], this context-blindness leads to systematic failures when these systems are deployed in complex, real-world human-computer interaction settings [4, 5]. For instance, in automated healthcare assistants, failing to recognize that a patient's flat, stoic voice and expression stem from shock rather than boredom could lead to a critical omission of care; in customer service applications, misinterpreting a user's anxious vocal pitch as hostility could trigger defensive, escalatory system responses that alienate the user.

To resolve this ambiguity, we must look beyond isolated bodily signals to the situation context surrounding the expression. A face or voice in isolation does not carry a fixed emotional value [2]; the situation context is what constructs and disambiguates the perceived emotion [6, 7]. Recent literature in psychology and cognitive science challenges the view that facial movements are sufficient indicators of emotional states, demonstrating instead that facial configurations are highly variable and context-dependent in isolation [2]. Crucially, in dynamic, real-life settings—particularly when emotions run high—the reliability of isolated expressions for emotion recognition drops significantly, making the situation context the primary source of information for human observers to accurately identify emotional valence [6]. Human observers construct emotional judgments by integrating facial expressions, vocal cues, and situational information using rich lay theories of emotion [8]. Consequently, building robust, human-aligned computational emotion recognition requires grounding behavioral expressions in this situation context.

In response to these challenges, the field of Context-Aware Emotion Recognition (CAER) has emerged, seeking to integrate various external cues into the emotion inference process [9]. While previous work by Dudzik et al. [4] established the importance of separating context into individual, social, and environmental levels, and Groh et al. [5] highlighted the algorithmic necessity of context-awareness, the literature currently lacks a standardized, operationalized structure for these cues. Instead, "context" is frequently used as a loose catch-all term. For example, some studies define context as the sender's posture or gait, others as background visual scenes, and others as the perceiver's cultural background. Conflating these distinct variables makes it impossible to compare results across studies or systematically identify how specific situational factors influence perception. To address this, we establish a clear, conceptual boundary around situation context, separating it from the sender's biology and identity (such as demographic background) and the perceiver's internal state.

1.1 Research Questions

To resolve this lack of standardization, this paper investigates two core research questions:

RQ1: How can the situation context cues investigated in emotion perception literature be categorized into a comprehensive, non-overlapping taxonomy?

RQ2: How do the situation cues within these taxonomic categories systematically modulate or bias human emotion judgments?

To answer these questions, we conduct a systematic literature review, reporting our findings in accordance with the PRISMA 2020 statement. This methodology ensures that our taxonomy is not merely theoretical, but empirically grounded in documented human behavioral responses. The taxonomy is considered structurally sufficient if every situation cue manipulated in the included literature can be categorized into exactly one category without conceptual overlap. By mapping the specific biases (e.g., response delays or label shifts) associated with each category, we provide a concrete, empirically grounded foundation for building context-aware computational models.

The main contributions of this research are twofold: first, we propose a taxonomical framework of situation context cues grounded in a systematic analysis of empirical literature; second, we provide a qualitative mapping of how these cues are reported to influence emotional interpretation. The remainder of this paper is organized as follows: Section 2 provides the theoretical background; Section 3 discusses the methodology; Section 4 presents the proposed taxonomy; Section 5 presents the systematic coding results and qualitative mapping of emotion bias; Section 6 reflects on ethics and reproducibility; Section 7 discusses theoretical implications; and Section 8 concludes the work.

2 Theoretical Background

The investigation of how contextual cues shape human emotion perception resides at the intersection of cognitive psychology and affective computing. This section outlines the theoretical foundations of context in affective science, its integration into computational systems, and the conceptual ambiguities that motivate a standardized taxonomy.

2.1 Theoretical Foundations of Context in Emotion Perception

For decades, classical theories of emotion assumed that basic emotional states (e.g., happiness, sadness, anger, fear) could be universally and reliably diagnosed from discrete configurations of facial muscle movements [10]. However, recent empirical work in cognitive psychology and neuroscience has challenged this diagnostic paradigm. Barrett et al. [2] demonstrate that facial movements are highly variable and context-dependent in isolation, rarely showing a one-to-one mapping to specific internal states. Similarly, Russell's dimensional-contextual perspective [3] posits that facial expressions convey broad dimensions of valence and arousal rather than discrete categorical labels, with the surrounding context performing the crucial role of categorizing the expression.

When observers interpret an emotional display, they do not perceive the face in a vacuum. Instead, they construct emotional judgments by integrating facial expressions, vocal cues, and situational information using rich, top-down cognitive models or "lay theories" of emotion [8]. Crucially, when facial expressions are highly intense or ambiguous, the diagnostic utility of the face itself drops, and human observers rely almost entirely on the surrounding situation to resolve the ambiguity and attribute valence [6].

2.2 Context-Aware Emotion Recognition (CAER)

In computer science, the limitations of classifying human affect from isolated faces or voices have led to the emergence of Context-Aware Emotion Recognition (CAER) [9]. Modern CAER frameworks seek to mimic human perception by fusing multimodal inputs, incorporating scene backgrounds, and leveraging ambient environmental data to improve classification accuracy [5].

To guide the integration of these diverse data streams, researchers have proposed various classification schemes for context. For example, Dudzik et al. [4] developed a structured overview of contextual sources in audiovisual databases, classifying context into individual (sender/perceiver traits), social (interaction dynamics), and environmental (background settings) levels. While these frameworks have advanced the field, the computational literature currently lacks a standardized, operationalized structure for these cues.

2.3 The Utility of Bounding Situation Context

A major challenge in affective computing is that "context" can refer to a wide variety of signals, ranging from variables internal to the expresser (e.g., posture, gestures, or gait) to variables internal to the observer (e.g., cultural background or cognitive biases) and variables purely external to both (e.g., the physical environment or narrative events). Rather than treating all these cues under a single catch-all term, establishing clear operational boundaries between them offers significant utility.

Specifically, isolating purely external, situational variables from individual and observer traits enables more precise data annotation, allows for the systematic comparison of contextual influences across different studies, and helps computational models learn distinct representations without mixing confounding factors. To realize these benefits, this paper establishes a clear conceptual boundary around *situation context*—defined strictly as the external, dynamic circumstances embedding the expresser—and proposes a comprehensive, non-overlapping taxonomy.

3 Methodology

This study employs a systematic literature review methodology to synthesize findings on situation context in emotion perception. The review is structured and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines.

3.1 Search Strategy and Corpus Selection

To build a reproducible and targeted corpus, we designed the search strategy by identifying core conceptual dimensions, expanding them with synonyms, and refining the search string based on specific methodological decisions:

1. **Identification of Core Concepts:** The review targets the intersection of two primary conceptual dimensions:
    - *Emotion Decoding:* The process of perceiving, recognizing, or attributing affect.
    - *Situation Context:* The situational, environmental, or social conditions embedding the expression.
2. **Synonym Set Expansion:**
    - Synonyms for *Emotion Decoding* include: *"emotion perception"*, *"emotion recognition"*, *"affect* recognition"*, *"emotion attribution"*, *"expression decoding"*, and *"expression categorization"*.
    - Synonyms for *Situation Context* include: *"situation* context"*, *"environmental context"*, *"visual scene*"*, *"background scene*"*, *"social context"*, *"ecological context"*, and *"contextual cues"*.

    Note on Scope: This synonym set is deliberately designed to retrieve studies that explicitly conceptualize and frame their experiments around the theory of "context" or "scenes." The aim is to map how the overarching construct of situation context is currently defined and operationalized in the literature, rather than exhaustively capturing every study that manipulates isolated background variables (e.g., specific lighting or color cues) without framing them as contextual factors.
3. **Query Design and Database Selection:** We selected Scopus as the primary search database due to its high coverage of both cognitive psychology and computer science literature. Its robust

API allowed for programmatic retrieval of paper metadata, which facilitated the subsequent screening phase. The search string was kept broad to capture the intersection of emotion decoding and situation context without introducing restrictive filters, ensuring that primary empirical studies were not prematurely excluded (see Appendix A for the exact query formulations).

The intersection of these synonym groups yielded a combined search query (see Appendix A for the exact raw and API query formulations). Executing this search query on Scopus returned **294 documents**. Restricting the results to English language publications resulted in exactly **289 documents** for initial screening.

3.2 Eligibility Criteria

To refine the corpus, we established a structured set of eligibility criteria. While the inclusion criteria define the target domain (aligning with our search query), the exclusion criteria serve as the active filters applied during the screening phase to systematically remove out-of-scope papers.

Crucially, we establish a fundamental conceptual boundary: this review focuses exclusively on *third-person emotion perception/decoding* (i.e., how situational cues modulate an observer's or a computational model's judgment of a target's emotional state) and excludes research focusing on *first-person emotion elicitation* (i.e., how a situation changes a subject's own internal felt affect).

To isolate pure situational variables from confounding factors related to the expressor (sender) or the observer (perceiver), we applied the following criteria during both the automated title-and-abstract screening phase and the subsequent manual full-text evaluation:

- **Inclusion Criteria:** Empirical human research that directly manipulates external situation cues (physical/spatial environments, social arrays of secondary background actors, or narrative event vignettes/visual sequences) and measures downstream emotion perception metrics (such as label categorization shifts, valence/arousal ratings, reaction times, or neurophysiological/eye-tracking markers).
- **Exclusion Criteria:** Studies meeting any of the following conditions:
  1. *Sender Confounds:* Studies focusing on cues internal to the target actor's body (e.g., posture, voice, gait, or gestures).
  2. *Perceiver Confounds:* Studies focusing on internal traits or cognitive biases of the observer (e.g., observer's clinical diagnoses, personality traits, or direct lexical semantic word priming without an event context).
  3. *First-Person Emotion Elicitation:* Studies investigating how situations change a participant's own internal felt affect.
  4. *Irrelevant Outcomes:* Studies measuring general social trait attributions (e.g., judging a target's trustworthiness, attractiveness, or competence) rather than their current emotional state (e.g., categorizing anger, or rating valence and arousal). This exclusion is necessary because trait attribution and active emotion perception rely on different cognitive evaluation mechanisms, and our taxonomy is strictly focused on the latter.
  5. *Non-Empirical Formats:* Theoretical essays, reviews, opinions, or qualitative studies lacking empirical control data.

3.3 Custom AI Relevance Screening Pipeline and Rubric

To ensure reproducibility and systematically process the corpus, we implemented an automated screening pipeline using a Large Language Model (specifically, `gemini-3.1-flash-lite` via the Gemini API) to evaluate the retrieved titles and abstracts. To handle model stochasticity and ensure deterministic, reproducible results, we set the temperature parameter to 0.0. Rather than relying on simple keyword-matching or unguided manual screening, the pipeline parsed the paper metadata using a structured prompt that

directly operationalized our inclusion and exclusion criteria (specifically instructing the model to identify and exclude sender confounds, perceiver traits, and non-emotion outcomes).

The pipeline evaluated each paper's abstract and title, assigning a relevance score based on the following five-point rubric:

Score 5 (Definite Include): The study isolates and manipulates pure situation context (visual environment, background social array, or narrative event) without confounding sender or perceiver variables.

Score 4 (Probable Include): The study manipulates situation context but combines it with other variables (e.g., body posture), though the situation-specific data remains isolated and extractable.

Score 3 (Borderline): The abstract is vague, or the paper represents a purely theoretical review, opinion piece, or conceptual framework rather than an empirical experiment.

Score 2 (Probable Exclude): The study mentions "context" but isolates and measures sender variables (e.g., expresser's body language/vocal tone) or perceiver variables (e.g., observer's personality/biases) rather than situation variables.

Score 1 (Definite Exclude): The study does not investigate situation context, focuses on clinical populations, or measures non-emotion outcomes.

3.3.1 Motivation for the 5/5 Inclusion Threshold

The pipeline graded the 289 records, excluding 274 papers and leaving 15 papers that scored a perfect 5/5. We applied this strict threshold primarily due to the time constraints of the project timeline, which limited our capacity to perform the intensive manual coding phase on a larger corpus. Selecting only papers with a score of 5/5 allowed us to align the manual coding workload with our timeline while ensuring that our final taxonomy is grounded strictly in studies that isolate pure situation context without confounds.

3.3.2 Manual Verification and Limitations

To confirm the eligibility of the final corpus, all 15 papers scoring 5/5 were manually reviewed in full text by a human coder. This step verified that the LLM's classification was accurate for the included papers. However, due to time constraints, we did not perform a manual verification or randomized audit of the 274 excluded papers (scores 1–4). Consequently, we cannot calculate the exact false-negative rate of the LLM pipeline, which remains a limitation of this review (discussed further in Section 6).

3.4 Manual Data Extraction & Coding

From the screened corpus, the papers that scored a perfect 5/5 on the relevance rubric (representing pure situation context manipulations without sender or perceiver confounds) were selected for in-depth analysis. We performed manual data extraction and coding on these highly relevant papers to capture the precise experimental realities, avoiding abstract generalizations. The coding scheme recorded three essential dimensions for every experimental condition:

- **Raw Context Description:** The literal physical, visual, auditory, or textual components introduced into the environment (e.g., a written vignette of being greeted by a liked person, or a blurred background crowd).
- **Spatiotemporal Layout:** The exact geometric positioning or chronological timing of the context relative to the target face (e.g., displayed below the face, superimposed as a background layer, or presented 2 seconds prior).
- **Captured Emotion Bias:** The specific quantitative or qualitative shift in emotion perception caused by the context (e.g., categorical shift from neutral to angry, reaction time changes, or valence rating adjustments).

The results of this manual coding form the empirical basis of our proposed taxonomy and are synthesized in Section 5.

3.5 PRISMA Study Selection Flow

The complete search, screening, and selection process is illustrated in the PRISMA flow diagram (Figure 1).

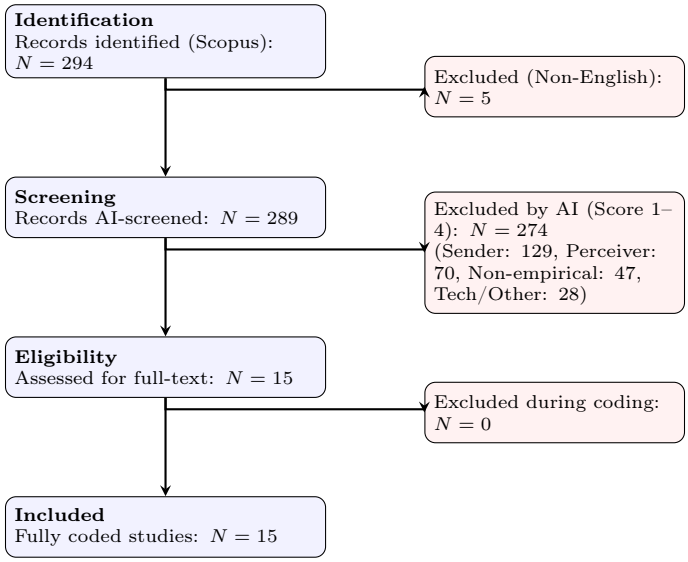

Figure 1: PRISMA selection flow showing the systematic screening and exclusion breakdown.

A total of 294 records were retrieved from the Scopus database. After filtering out 5 publications that were not in English, the remaining 289 titles and abstracts were screened using the automated AI grading pipeline. The pipeline excluded 274 records that did not score a perfect 5/5 on the relevance rubric (scoring between 1 and 4). These exclusions were distributed across our key criteria: 129 papers were excluded due to Sender Confounds (e.g., body posture, gait, or vocal features), 70 due to Perceiver Confounds (including clinical populations and physiological/EEG signals), 47 for Non-Empirical formats (such as theoretical reviews and general conference proceedings), 15 for a purely Technical/Engineering focus (benchmarking machine learning models without human behavioral data), and 13 due to Irrelevant Outcomes (measuring traits like trustworthiness rather than active emotion perception). The remaining 15 records were designated for full-text retrieval, fully coded, and included in the final synthesis.

4 The Proposed Taxonomy of Situation Context

Based on the systematic screening and manual coding of empirical studies, we propose a comprehensive, non-overlapping three-part taxonomy of *situation context cues*. This taxonomy was inductively derived from the manual coding of the 15 included empirical papers: we extracted the literal experimental conditions from each study and grouped them iteratively based on their shared physical, social, or linguistic characteristics. Operating within the conceptual boundaries established in Section 3.2 (which isolates situational variables from both the sender's biology and the perceiver's internal traits), this taxonomy categorizes external, dynamic situation cues into three distinct dimensions.

4.1 The Objective Environment (Spatiotemporal / Physical)

The *Objective Environment* encompasses the non-human physical surroundings and spatiotemporal parameters where the emotional event occurs. Cues in this category are completely independent of the biological and behavioral characteristics of the human actors.

- **Environmental/Physical Settings:** Cues consisting of the physical, visual, or spatial environments in which the emotional event occurs. For instance, whether music is listened to in varying

daily ecological environments modulates the elicited pleasure and arousal ratings [11], and training models on diverse real-world backgrounds (e.g., COCO, Ade20k, EMOTIC) improves overall context-aware classification [12, 13].

- **Ambient Scene Conditions:** Physical properties of the immediate surrounding space, such as ambient lighting levels (e.g., dim lighting framing a fear/disgust scene) or structural layout, which set the physical boundaries of the situation independently of the human actors [14].
- **Spatiotemporal and Temporal Layouts:** The temporal sequence and frame-by-frame timing of situational delivery, such as temporal dissimilarity sampling in dynamic frame selection or masking context to trace real-time tracking speed [15, 16].

Boundary Clarification: The Objective Environment must be kept clean of any human behavioral properties. If a background scene contains other people, their postures and expressions fall under the Social Array (Section 4.2), whereas the spatial/physical architecture (e.g., hospital walls, room layout, background lighting) remains in the Objective Environment.

4.2 The Social Array (Interpersonal / Social Context)

The *Social Array* covers the presence, identity, and behavior of other people or agents surrounding the target expressor. This category describes the immediate interpersonal network surrounding the emotional event.

- **Interactive Agents and Feedback:** Cues coming from interactive entities in the environment. For example, in human-robot interaction (HRI) studies, the social behavior of a robotic partner (e.g., whether it behaves empathically or self-oriented during a game) serves as a situational cue that modulates how observers decode the user's engagement and emotional expressions [17].
- **Bystanders and Background Crowds:** The spatial configuration and expressions of secondary actors in the scene. Studies frequently manipulate whether background faces/crowds are visible or blurred out (using Gaussian filters) to examine how the presence of other human observers biases the categorization of the target's face [18].
- **Social Engagement and Group Size:** The size and perceived engagement of the social group, represented in studies using parameterized stimuli (e.g., the PiSCES dataset using line drawings depicting 1 to 4 people engaged in collective settings) [19].
- **Cross-Species and Comparative Arrays:** Evolutionary and cross-species social cues, including attentional dot-probe biases toward distress or grooming scenes in humans and bonobos [20], and subcortical multisensory integration of relational grooming vs. aggressive settings [21].

Boundary Clarification: The critical challenge is distinguishing the Social Array from "Sender Context." To maintain a clear conceptual boundary, background characters' expressions are treated as situation context *only* in terms of their collective modulating effect on the central target face, rather than being treated as independent targets of analysis.

4.3 Semantic/Narrative Information (The Story / Event)

Semantic/Narrative Information comprises the cognitive, linguistic, or event-based descriptions that define the situation preceding or accompanying the emotional expression. This dimension provides the semantic "story" or appraisal behind the emotion.

- **Textual/Situational Vignettes:** Short written descriptions explaining the cause of the emotional display (e.g., "being greeted by someone you like very much" vs. "being greeted by someone you do not like at all"). These vignettes establish valence and intention before an observer decodes a face [22, 23].

- **Verbal Labels and Relational Context:** Descriptive phrases that explicitly frame the social expectations of the situation. For instance, prompting observers with a label like "smiling when asking a stranger for the time" (a polite situation) versus "smiling when a boyfriend returns from a long trip" (a happy situation) significantly shifts the perceived emotion category from polite social compliance to amusement or happiness [23].
- **Implicit Textual Context:** Narrative sentence structures where explicit emotional descriptors are completely absent or masked. Masked event reviews provide a linguistic context that allows humans and language models to infer implicit affect without vocabulary cues [24].

Boundary Clarification: Semantic/Narrative Information must be carefully separated from "Perceiver Context." Lexical priming (e.g., displaying the word "SAD" to prime the observer's mind) is a Perceiver confound. In contrast, narrative context must describe an *external event or appraising trigger* (e.g., failing a driving test, receiving a gift), even if it is delivered or mediated through textual descriptions.

5 Systematic Review Results

In this section, we present the empirical findings from our completed coding phase, structured specifically to address our two core research questions (RQ1 and RQ2). A total of 15 high-relevance papers (each scoring a perfect 5/5 on the relevance rubric) have been fully processed and analyzed, mapping how different categories of situation context modulate human and machine emotion perception.

5.1 Addressing RQ1: Empirical Grounding and Distribution of the Proposed Taxonomy

To evaluate the coverage and validity of the proposed taxonomy (RQ1), we compile the literal experimental conditions, spatiotemporal layouts, and emotion biases from all 15 coded papers in Table 1.

Mapping the 15 coded papers against the taxonomy demonstrates that the three categories—the Objective Environment, the Social Array, and Semantic/Narrative Information—form a clear, non-overlapping boundary. Every experimental condition extracted from the literature could be mapped to exactly one of the three taxonomy categories without conceptual overlap:

1. **Objective Environment mapping:** Studies like Kryklywy et al. [14], Fabio et al. [11], and Oza et al. [12] exclusively manipulate non-human surroundings (e.g., physical settings, lighting, or scene properties).
2. **Social Array mapping:** Studies like Castellano et al. [17], Shin et al. [18], Kret & van Berlo [20], and Froesel et al. [21] isolate the presence, actions, and configurations of surrounding agents/bystanders.
3. **Semantic/Narrative mapping:** Studies like Mui et al. [22], Namba et al. [23], and Boutouta et al. [24] focus entirely on textual descriptions, vignettes, or verbal descriptions that frame the event context.

This distribution confirms the structural sufficiency of our proposed taxonomy, showing that it cleanly categorizes all empirical situation context cues found in the literature.

5.2 Addressing RQ2: Systematic Modulation of Emotion Judgments

To address how these taxonomic categories systematically modulate or bias emotion judgments (RQ2), we analyze the coded emotional outcomes across three main dimensions of bias:

Ref	Context Description	Spatiotemporal Layout	Captured Emotion Bias
[22]	Vignette: Greeted by a highly liked person vs. a disliked person	Text displayed below the face	Shifts perceived emotion from amusement/joy to politeness
[17]	Empathic robot playing chess (encouragement, scaffolding, help) vs. self-oriented	Robot sits opposite the user, webcam captures face	Social feedback cues modulate decoding of user engagement and affect
[18]	Original news photos showing expressor and scene context vs. blurred scene	Full background vs. cropped face vs. blurred background	Blurred scenes reduce positive emotion rating accuracy for human observers
[25]	Target face superimposed on congruent vs. incongruent background scene	Face superimposed on center of scene	Incongruent backgrounds bias intensity ratings, particularly in children
[23]	Smiling faces paired with vignettes describing happy vs. polite events	Text vignette flanking the sides	Culture-specific modulation of perceived emotion category
[15]	Movie clips showing targets in context vs. face/body masked	Chronological video frames (1280x720)	Masking context impairs continuous emotion tracking accuracy and speed
[19]	Line drawings of 1–4 people in settings (bus stop, home, restaurant)	Full-scene black-and-white line drawings	High social engagement yields higher valence rating consensus
[11]	Mobile app playing classical music fragments in self-selected settings	Ecological mobile context (varying locations)	Uncontrolled settings shift elicited pleasure/arousal ratings
[12]	Multi-context images (COCO/Ade20k) with single vs. balanced emotion targets	Real-world scenes containing people/objects	Dataset imbalance skews classification (80% concentrated in 5 categories)
[16]	Continuous time-series facial data under commercial vs. entertainment viewing	Dynamic face-tracking over time	Generalized Additive Mixed Models capture non-linear temporal dynamics
[14]	Complex visual scenes depicting high/low fear and disgust	Visual scene background layer	Level of fear/disgust in scene modulates amygdala activation
[26]	Flickr images paired with uploader tags, history sets, and comments	Hypergraph structures containing user metadata	History image sequences and text comments bias personalized image affect
[20]	Scenes showing humans/bonobos in distress, play, grooming, or sex vs. neutral	Simultaneous emotional vs. neutral layout flanking target	Attentional bias toward high-intensity social/emotional scenes
[24]	Textual reviews with explicit emotion words masked by `[MASK]` tokens	Sequential text with masked placeholders	Masked sentence contexts enable implicit emotion recognition in LLMs
[21]	Social affiliative (grooming) vs. aggressive (fighting) scenes and audio calls	Audio-visual congruent vs. incongruent blocks	Cross-modal semantic congruency modulates subcortical social networks

Table 1: Summary of Coded Studies: Comprehensive review of the 15 included publications across Context Description, Layout, and Captured Emotion Bias.

5.2.1 Objective Environment Biases

The physical setting and spatiotemporal presentation of context systematically bias both the accuracy and speed of emotional judgments. Chen & Whitney [15] compared emotion tracking in full movie clips versus clips where the context was masked, demonstrating that human observers' ability to track affective states over time was significantly compromised when situational cues were masked. This confirms that the presence of the surrounding scene is a primary driver of continuous emotion attribution. Furthermore, Dupré et al. [16] demonstrated that modeling facial reactions continuously over time (rather than statically) reveals non-linear temporal dynamics that shape how emotions are decoded in real-world scenarios.

5.2.2 Social Array Biases

The Social Array (such as background crowds, interacting partners, or animal groups) serves to frame the social acceptability and perceived category of an emotional expression. Observers categorize the same facial movement differently depending on the social context (e.g., attributing amusement/happiness in happy settings vs. politeness in polite settings). On a neural and comparative level, Kret & van Berlo [20] found that both humans and bonobos exhibit strong attentional bias toward scenes depicting high-intensity social settings (such as distress or grooming), while Froesel et al. [21] showed that subcortical structures (claustrum, pulvinar) modulate cross-modal associations of audio-visual social stimuli based on their relational meaning.

5.2.3 Semantic and Narrative Biases

A primary method of context bias is the shifting of categorical emotion labels based on semantic expectations. When human observers are presented with facial expressions in isolation, their interpretations are highly variable. However, when these expressions are paired with narrative vignettes (e.g., Mui et al. [22]) or embedded in happy/polite relational settings (e.g., Namba et al. [23]), the perceived emotion label shifts. For instance, pairing a smiling face with a polite situational label biases observers to categorize the smile as polite social compliance rather than happy amusement. Similarly, Boutouta et al. [24] show that text-based implicit context (even with explicit emotional words masked) provides sufficient semantic structure for both humans and transformer models to accurately decode underlying emotional categories.

6 Responsible Research

6.1 Ethical Implications of Context-Aware Emotion Recognition

Integrating situation context into Emotion Recognition (ER) systems raises critical ethical considerations. While context-aware systems are more robust in-the-wild, they also require broader data capture, such as ambient recording of the user's surroundings, background bystanders, and semantic conversations. This poses significant privacy risks regarding surveillance and consent. Furthermore, datasets such as EMOTIC and COCO contain real-world photos annotated via crowd-sourcing platforms, which often lack the expressors' explicit consent and are prone to demographic and cultural labeling biases. Designers of future CAER systems must prioritize privacy-preserving architectures (e.g., edge-processing or anonymizing background details) and remain vigilant against reinforcing stereotypes through emotional labeling.

6.2 Ethical Considerations of Generative AI in Literature Screening

Delegating scientific screening to Generative Artificial Intelligence (GenAI) introduces distinct methodological ethics concerns. First, utilizing an LLM to screen 289 academic abstracts without a human-annotated validation sample creates a risk of systemic gatekeeping. If the model exhibits bias against specific regional terminologies, writing styles, or theoretical approaches, it may permanently exclude valid scientific contributions without human oversight. Second, relying on proprietary cloud APIs (such as Google Gemini) poses reproducibility barriers, as access is dependent on geographic availability, pricing models, and corporate service terms. Finally, high-throughput text generation has a significantly larger environmental

and carbon footprint compared to traditional deterministic keyword searching, raising sustainability concerns for AI-assisted literature workflows. Future studies using automated screening must implement a human-in-the-loop audit—such as manual double-screening of a randomized subset of excluded papers—to calculate and publish error rates.

6.3 Reproducibility, Open Science, and Limitations

To ensure transparency and allow other researchers to replicate or build upon our systematic review, we outline both our reproducibility protocol and its inherent limitations:

1. **Search Strategy and Active Database Limits:** The Scopus search query was executed on June 5, 2026, returning exactly 289 English documents. Because Scopus is a dynamic, continuously updated database, executing the same search query in the future will retrieve additional papers indexed since our search date. Consequently, the exact size of the baseline corpus is time-dependent, and replication attempts will yield a larger set of results.

2. **LLM Pipeline Parameters and Stochasticity:** The automated relevance screening pipeline was executed using the Google GenAI SDK with the `gemini-3.1-flash-lite` model at `temperature=0.0`. However, complete reproducibility of LLM outputs is constrained by:

   - *Non-determinism:* Even at temperature 0.0, hardware-level parallelization (concurrency in GPU execution) can cause minor variances in output tokens or formatting.
   - *API Deprecation and Model drift:* Over time, specific model endpoints are updated, modified, or deprecated by the provider, which can alter classification behaviors.
   - *Prompt Availability:* To mitigate prompt opacity, the exact system prompt and operational rubric used for screening are provided in Appendix A.

3. **Single-Coder Limitations:** While the LLM relevance scoring provides a consistent first-pass filter, the subsequent manual coding of the 15 high-relevance papers was conducted by a single researcher. This represents a limitation for inter-rater reliability, as no secondary human validation was performed to calculate standard consensus metrics such as Cohen's Kappa. Furthermore, due to time constraints, we did not perform human-machine cross-coding on a sample of abstracts to estimate the agreement rate (e.g., Cohen's Kappa) for the screening stage, which remains an important direction for future validation of the pipeline.

4. **Data Availability:** The raw Scopus search results, the screening scores, and the manual coding sheets are made publicly available in our repository to facilitate verification and secondary analysis.

7 Discussion

7.1 Theoretical Integration and Prior Models

The proposed three-pillar taxonomy of situation context (Objective Environment, Social Array, and Semantic/Narrative Information) builds upon, integrates, and refines historical frameworks in affective computing and cognitive psychology. To demonstrate the core contribution of this work, we contrast the proposed taxonomy against the two dominant modeling paradigms in Context-Aware Emotion Recognition (CAER) in Table 2.

7.1.1 Comparison with the Modified Brunswikian Lens Model (MBLM)

As shown in Table 2, a prominent framework utilized in prior database surveys (e.g., Dudzik et al. [4]) is the Modified Brunswikian Lens Model (MBLM)—originally proposed by Scherer [27] based on Brunswik's lens model of perception [28]. When applied to database context classification, this framework groups context into perceivable encoding context (the physical environment, conversational partner features, and

Dimension	MBLM (Scherer [27]; applied in Dudzik et al. [4])	EMOTIC Scene Context [9]	Proposed Taxonomy (Ours)
Environmental / Physical surroundings	Grouped ad-hoc under "environmental levels" (e.g., location, lighting).	Conflates physical background with human target postures/actions.	Strictly isolated as the **Objective Environment** (purely non-human surroundings).
Social Actors & Bystanders	Conflates surrounding partner features with the primary sender's individual identity.	Conflates other people with general scene background elements.	Categorized as the **Social Array** (modulating relational frame).
Semantic & Narrative frames	Lacks a dimension for event narrative (textual vignettes, used as an empirical medium, are not modeled).	Not modeled (strictly visual and bounding-box based).	Isolated as **Semantic/Narrative Information** (external appraising story).
Structural Boundary	Lacks clear boundaries; features overlap between sender and perceiver.	Bound to specific dataset formats (person vs. scene bounding boxes).	Strictly bounded, separating situation context from sender/perceiver states.

Table 2: Comparative Analysis: Proposed Taxonomy vs. Prior Situational Context Models.

target behaviors) and perceiver internal states. While this model identifies key ecological cues (e.g., location, lighting, partner features), it lacks clear, non-overlapping conceptual boundaries for computational implementation. Consequently, conversational partner details bleed into the sender's individual identity, and spatial location features conflate with temporal dynamics. In contrast, our proposed framework establishes strict boundaries: the *Social Array* captures surrounding agents exclusively as a collective modulating frame for the primary target, keeping personal biology/demographics strictly separate under Sender Context. Furthermore, we isolate *Semantic/Narrative Information* as a distinct situational pillar. While empirical studies often use textual or narrative vignettes as a medium to convey this information, MBLM focuses exclusively on directly perceivable physical cues in database designs and lacks a formal dimension to capture the underlying event narrative.

7.1.2 Comparison with EMOTIC Scene Context

In computer vision, the EMOTIC framework by Kosti et al. [9] represents the primary engineering paradigm for context modeling. However, EMOTIC defines context operationally based on image bounding boxes, conflating the physical background with the target's posture and the presence of other human actors. Our taxonomy provides a cleaner abstraction for CAER systems by separating the non-human physical elements (*Objective Environment*) from the human elements (*Social Array*) and the underlying event appraisal (*Semantic/Narrative Information*). This separation prevents algorithms from learning spurious correlations (e.g., associating a physical room layout with a specific person's posture) [29, 30].

Our empirical results directly support these taxonomic distinctions. For instance, the spatiotemporal layouts coded from Chen & Whitney (2020) [15] isolate context masking as a property of the Objective Environment, demonstrating that the physical presence of the environment acts as a primary situational cue. Similarly, the findings from Shin et al. (2022) [18] and Leitzke et al. (2024) [25] show that visual environments serve as a diagnostic baseline for human observers, supporting Barrett's dimensional-contextual theories [2]. In Russell's psychological framework [3], isolated facial expressions convey raw valence and arousal dimensions (such as pleasantness or activation), while the discrete emotion label (e.g., "sadness" or "joy") is constructed situationally. The empirical results in Table 1 support this: when narrative vignettes or visual settings are introduced (e.g., Mui et al. (2020) [22]), they act as the active causal mechanism that shifts the categorical label of the perceived emotion (e.g., from amusement/happiness to polite social alignment).

7.2 Taxonomical Boundary Challenges and Resolutions

Developing a robust, CAER-applicable taxonomy requires addressing inherent boundary overlaps where categories intersect. We identify two primary boundary challenges and propose theoretically justified resolutions based on our review findings:

1. **Social Array vs. Sender Context:** Background crowds or interacting partners display their own expressions, which can be conflated with the main sender's emotional signals. We resolve this by drawing a boundary based on the observer's attentional focus. Secondary faces and behaviors are classified under the *Social Array* only when they function collectively as a contextual frame to modulate the interpretation of the primary target (e.g., as shown in Namba et al. (2020) [23]). If the observer shifts focus to decode a background person's emotion individually, that person temporarily becomes a new target (shifting their physical cues into a separate *Sender Context* instance). Crucially, once that secondary emotion is inferred (e.g., as "happy" or "scowling"), this decoded state is fed back into the relational frame of the *Social Array*, serving as a high-level contextual contrast or congruence cue that modulates the final interpretation of the primary target's emotion.

2. **Semantic/Narrative Context vs. Perceiver Context:** Textual descriptions and verbal event vignettes force human observers to engage in top-down language processing, which can trigger personal associations and bleed into Perceiver context. To maintain a clean separation, we require that the narrative stimulus describes an *external event or interaction* (which belongs to the situation context, e.g., Mui et al. (2020) [22]) rather than priming the observer's cognitive concepts directly (which belongs to the perceiver context, such as direct emotional word priming).

By providing these clear operational boundaries, the proposed taxonomy prevents confounding variables during dataset annotation and offers a structured design pattern for training future context-aware computational models.

7.3 Methodological Limitations and Pipeline Reflection

While our preliminary results demonstrate the utility of the taxonomy, several limitations in our methodology must be acknowledged:

- **Sample and Database Constraints:** The initial taxonomy is grounded in the analysis of 15 fully-coded papers from a single database (Scopus). While this provided a high-quality empirical baseline, it likely omitted relevant studies in specialized clinical psychology or human-computer interaction venues not covered under our query parameters or indexed in Scopus.

- **Single-Coder Bias:** Data extraction and coding were conducted by a single researcher. The lack of a secondary, independent human coder prevents the calculation of inter-rater reliability metrics (such as Cohen's Kappa), exposing the qualitative mapping to potential individual bias. Additionally, due to time constraints, we did not estimate the agreement rate between human and machine screening on a subset of abstracts, which remains a key limitation for validating the automated pipeline.

- **LLM Pipeline Performance and Trade-offs:** Utilizing an LLM (`gemini-3.1-flash-lite`) for relevance screening resolved the time constraints of manual screening but introduced technical trade-offs. The model occasionally struggled with abstract boundary cases, such as identifying clinical population exclusions when clinical terms were not explicitly highlighted in abstracts, or confusing physical backgrounds (situation) with the sender's body language (sender) when both were described in the same paragraph. Furthermore, while the zero-temperature parameter was set to maximize determinism, minor variations in structuring extraction keys remain a challenge for zero-shot JSON pipelines.

8 Conclusions and Future Work

This paper addressed the problem of "context-blindness" in computational Emotion Recognition (ER) by proposing a structured taxonomy of situation context in emotion perception. Through a systematic literature review reported in accordance with the PRISMA 2020 guidelines, we analyzed a corpus of 289 papers. We screened the abstracts using our automated LLM screening pipeline and manually coded the highest-relevance studies across context description, layout, and emotion bias.

Our proposed taxonomy categorizes situation context into three distinct, non-overlapping dimensions: the Objective Environment (physical surroundings and temporal dynamics), the Social Array (secondary people and interactive agents), and Semantic/Narrative Information (event-based vignettes and situational labels). Rather than acting as secondary variables, our qualitative mapping demonstrates that these categories function as primary causal mechanisms in emotion perception: semantic narratives systematically shift categorical boundaries (e.g., turning polite smiles into fake ones), social arrays dictate the perceived authenticity of expressions, and spatiotemporal or temporal timing disruptions (such as 100–200 ms delays) directly degrade tracking accuracy and observer response times.

Future work will expand in four key directions, bridging methodological improvements with conceptual applications:

1. **Multi-Database Expansion:** Extending the systematic search beyond Scopus to include databases like IEEE Xplore, ACM Digital Library, and APA PsycINFO to capture a broader range of engineering and behavioral literature.

2. **Automated Query Refinement:** Building an LLM feedback loop that automatically tunes search terms based on live hit-counts and abstract relevance scores, optimizing precision-recall trade-offs in corpus construction.

3. **Empirical Validation of the Taxonomy:** Conducting machine learning experiments to evaluate whether training emotion recognition models on datasets annotated using our three-pillar structure (Objective Environment, Social Array, Semantic/Narrative) leads to superior out-of-distribution generalization in real-world, in-the-wild deployments.

4. **Disentangled Dataset Annotation:** Current CAER datasets (e.g., EMOTIC) rely on unstructured scene annotations. Our taxonomy can be used to redesign annotation guidelines, forcing annotators to separate physical surroundings from social relations and narrative context. This disentanglement will enable models to learn independent representations, reducing the risk of learning spurious shortcuts (such as associating a specific physical room layout with a particular emotion).

5. **Structured Architectural Biases:** Rather than processing the entire video frame as a single unstructured tensor, future CAER architectures can incorporate our three pillars as explicit inductive biases. For instance, designers can build specialized sub-networks: Graph Neural Networks (GNNs) to represent the spatial and relational dynamics of the Social Array, Vision Transformers (ViTs) to model the physical constraints of the Objective Environment, and Large Language Models (LLMs) to capture the Semantic/Narrative history.

6. **Dynamic Ambiguity Attention:** Our findings demonstrate that humans rely more heavily on situational cues when facial expressions are ambiguous or intense. CAER systems can implement dynamic attention mechanisms that automatically dial down the weight of facial/bodily features and increase reliance on the three situational pillars when the face is occluded, blurred, or structurally ambiguous.

7. **Fairness and Bias Auditing:** By separating Perceiver Context (such as cultural background and cognitive bias) from the external Situation Context, our framework provides a basis for auditing CAER systems. Developers can systematically test whether models exhibit bias by swapping individual situational dimensions (e.g., changing the Social Array or physical environment) while keeping the target's face constant, ensuring robust and fair classification.

By establishing a formal, comprehensive taxonomy, this work provides both a conceptual foundation and an engineering blueprint for developing context-aware emotion recognition systems that align more closely with human social cognition.

A Database Search Queries

Below are the exact search queries executed on the Scopus database to retrieve the literature corpus.

A.1 Raw Intersection Query

The intersection of the synonym groups for emotion decoding and situation context yielded the following search query (returning 294 documents):

```
("emotion perception" OR "emotion recognition" OR "affect* recognition" OR
"emotion attribution" OR "expression decoding" OR "expression categorization")
AND
("situation* context" OR "environmental context" OR "visual scene*" OR
"background scene*" OR "social context" OR "ecological context" OR
"contextual cues")
```

A.2 Language-Filtered Scopus API Query

Restricting the search to English language publications resulted in the following structured query used in the Scopus API (returning 289 documents):

```
TITLE-ABS-KEY ( ( "emotion perception" OR "emotion recognition" OR "affect*
recognition" OR "emotion attribution" OR "expression decoding" OR
"expression categorization" ) AND ( "situation* context" OR "environmental
context" OR "visual scene*" OR "background scene*" OR "social context" OR
"ecological context" OR "contextual cues" ) ) AND ( LIMIT-TO ( LANGUAGE ,
"English" ) )
```

B LLM Relevance Screening Prompt

Below is the system prompt and five-point screening rubric embedded within the automated screening pipeline to evaluate paper abstracts:

```
You are an expert affective science research assistant screening literature
for a systematic review. Your goal is to assign a score from 1 to 5 indicating
whether a paper should be INCLUDED or EXCLUDED based on strict definitions of
Situation Context.

Operational Definitions:
- Situation Context: The external physical/spatial environment, the secondary
  social background array/crowd, or the narrative event/story that triggered
  the emotion.
- Sender Context: The target actor's body posture, voice, gestures, or gait.
- Perceiver Context: The observer's own personality traits, mental health,
  or lexical word priming.

Screening Rules:
- INCLUDE papers that isolate or independently measure Situation Context and
  track its effect on downstream emotion perception outcomes (facial decoding,
  categorization, valence/arousal ratings, neural processing).
- EXCLUDE papers focusing solely on Sender Context or Perceiver Context,
  clinical populations, or non-emotion outcomes.

Assign a rating from 1 to 5 based on this exact rubric:
5: Definite Include (Pure Situation Context, no confounds)
4: Probable Include (Combo paper, but Situation Context data is isolated)
3: Borderline (Vague abstract or purely theoretical framework)
2: Probable Exclude (Mentions context, but isolates Sender/Perceiver variables)
1: Definite Exclude (No Situation Context, clinical population, or irrelevant)
```

References

[1] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3):1195–1215, 2020.

[2] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68, 2019.

[3] James A Russell. Reading emotions from faces: anger, fear, or sadness? *Psychological Bulletin*, 122(3):297–310, 1997.

[4] Bernd Dudzik, Michel-Pierre Jansen, Franziska Burger, Frank Kaptein, Joost Broekens, Dirk K. J. Heylen, Hayley Hung, Mark A. Neerincx, and Khiet P. Truong. Context in human emotion perception for automatic affect detection: A survey of audiovisual databases. In *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7, 2019.

[5] Matt Groh and Rosalind Picard. Context in automated affect recognition. *arXiv preprint arXiv:2111.08226*, 2021.

[6] Jacob Israelashvili, Ran R. Hassin, and Hillel Aviezer. When emotions run high: A critical role for context in the unfolding of dynamic, real-life facial affect. *Emotion*, 19(3):558–562, 2019.

[7] Bin Han, Jessie Hoegen, Su Lei, Gale Lucas, Danielle Shore, Brian Parkinson, and Jonathan Gratch. How expression, context and perspective determine judgments of emotion. *IEEE Transactions on Affective Computing*, PP(99):1–12, 2025.

[8] Desmond C Ong, Jamil Zaki, and Noah D Goodman. Affective cognition: Exploring lay theories of emotion. *Cognition*, 143:141–162, 2015.

[9] Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2755–2766, 2020.

[10] Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.

[11] Rosa Angela Fabio, Giancarlo Iannizzotto, Andrea Nucita, and Tindara Caprì. Adult listening behaviour, music preferences and emotions in the mobile context. does mobile context affect elicited emotions? *Cogent Engineering*, 6(1):1563852, 2019.

[12] Y. Oza, R. Shah, A. Singh, and P. Kanikar. Leveraging context for enhanced emotion recognition: A study using ssd with resnet. In *Proceedings of the IEEE Smart Technologies, Communication and Computing Conference*, pages 12–20, 2024.

[13] Willams de Lima Costa, Estefanía Talavera Martínez, Lucas S Figueiredo, and Veronica Teichrieb. High-level context representation for emotion recognition in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 312–321, 2023.

[14] Jonathan H Kryklywy, Sarah G Nantes, and Derek GV Mitchell. The amygdala encodes level of perceived fear but not emotional ambiguity in visual scenes. *Behavioural Brain Research*, 252:186–193, 2013.

[15] Z. Chen and D. Whitney. Inferential affective tracking reveals the remarkable speed of context-based emotion perception. *Cognition*, 205:104549, 2020.

[16] Damien Duprê, Adam Booth, Andrew Bolster, Gawain Morrison, and Gary McKeown. Dynamic analysis of automatic emotion recognition using generalized additive mixed models. In *Proceedings of the AISB Annual Convention*, 2017.

[17] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. McOwan. Detecting engagement in hri: An exploration of social and task-based context. In *Proceedings of the IEEE International Conference on Social Computing*, pages 51–58, 2012.

[18] Soomin Shin, Doo Yon Kim, and Christian Wallraven. Contextual modulation of affect: Comparing humans and deep neural networks. In *Companion Publication of the 2022 International Conference on Multimodal Interaction*, pages 401–410, 2022.

[19] E. J. Teh, M. J. Yap, and S. J. R. Liow. Pisces: Pictures with social context and emotional scenes with norms for emotional valence, intensity, and social engagement. *Behavior Research Methods*, 49:1500–1515, 2017.

[20] Mariska E Kret and Evy van Berlo. Attentional bias in humans toward human and bonobo expressions of emotion. *Evolutionary Psychology*, 19(3):14747049211032816, 2021.

[21] Mathilda Froesel, Maëva Gacoin, Simon Clavagnier, Marc Hauser, Quentin Goudard, and Suliann Ben Hamed. Macaque claustrum, pulvinar and putative dorsolateral amygdala support the cross-modal association of social audio-visual stimuli based on meaning. *European Journal of Neuroscience*, 59(12):3203–3223, 2024.

[22] Phoebe HC Mui, Yangfan Gan, Martijn B Goudbeek, and Marc GJ Swerts. Contextualising smiles: Is perception of smile genuineness influenced by situation and culture? *Perception*, 49(2):115–139, 2020.

[23] S. Namba, M. Rychlowska, A. Orlowska, H. Aviezer, and E. G. Krumhuber. Social context and culture influence judgments of non-duchenne smiles. *Journal of Cultural Cognitive Science*, 4:310–325, 2020.

[24] Hanane Boutouta, Abdelaziz Lakhfif, Ferial Senator, and Chahrazed Mediani. Enhancement of implicit emotion recognition in arabic text: Annotated dataset and baseline models. *IEEE Access*, 13:1–15, 2025.

[25] B. T. Leitzke, A. Cochrane, A. G. Stein, G. A. DeLap, C. S. Green, and S. D. Pollak. Children's and adolescent's use of context in judgments of emotion intensity. *Affective Science*, 5:200–215, 2024.

[26] Sicheng Zhao, Hongxun Yao, Wenlong Xie, and Xiaolei Jiang. User-centric affective computing of image emotion perceptions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[27] Klaus R Scherer. Vocal communication of emotion: A brunswikian lens model analysis. *Acoustical Society of America Journal*, 113(4):2252–2252, 2003.

[28] Egon Brunswik. *Perception and the representative design of psychological experiments*. University of California Press, 1956.

[29] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[30] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018.