Situation Context in Context-Aware Emotion Recognition:

A Structured Literature Review of Conceptualisation and Modelling Approaches

Aksu Kaypmaz[1]
Responsible Professor: Bernd Dudzik[1]
Supervisor: Sayak Mukherjee[1]

[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 22, 2026

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

Abstract

Emotion recognition systems often infer affective states from observable signals such as facial expressions, speech, body posture, language, or multimodal behaviour. Yet emotional meaning is rarely determined by these signals alone: the same expression may communicate different emotions depending on the surrounding situation context. In this review, *situation context* refers to the external, environmental, temporal, social, conversational, or event-related information that shapes how an emotional signal is interpreted. This paper presents a structured literature review (SLR) synthesising 19 papers across five search strata to examine how situation context is conceptualised and modelled in context-aware emotion recognition research. The synthesis shows that situation context is not one shared computational object. It is operationalised as visual surroundings, conversational history, social interaction, causal event structure, commonsense knowledge, or semantic situation understanding, depending on the task and modelling tradition. Modelling approaches range from implicit use of temporal or multimodal features to explicit representations through context branches, graph structures, cause labels, commonsense knowledge, or prompting. A central limitation of the field is therefore structural fragmentation: situation context is considered important, but it is defined, represented, and evaluated inconsistently across tasks and datasets.

1 Introduction

Emotion recognition research aims to infer affective states from observable cues such as facial expression, body posture, speech, language, and multimodal behaviour. In many computational systems, these cues are treated as the main source of emotional meaning. Affective science and psychology challenge this assumption: emotional signals are often ambiguous when they are separated from the situation in which they occur [1, 2, 3, 4]. A smile, for example, may signal happiness, politeness, or discomfort depending on the surrounding *situation context*: the external, environmental, social, conversational, or event-related information that shapes how an emotional signal is interpreted. This review uses the term situation context throughout to refer to this class of contextual information.

This review narrows the broader context literature to *situation context*. Building on distinctions in prior context-focused work, this review separates contextual information into sender context, perceiver context, and situation context [5, 6]. Sender context concerns information about the person expressing the emotion, such as identity, personality, or individual traits. Perceiver context concerns information about the person interpreting the emotion, such as expectations, culture, or prior knowledge. Situation context concerns external and interactional information around the emotional signal.

Sender and perceiver context are also important, but they raise different modelling and ethical questions from the ones addressed here. In this review, situation context is treated as a modelling target when a paper uses one of four situation-related cue families: visual or environmental cues, conversational or interactional cues, causal event cues, and commonsense or semantic situation cues. These cue

families can be operationalised in computational models through inputs, annotations, relations, knowledge sources, or prompts [10, 12, 15, 16, 20, 21].

The motivation for studying situation context is both theoretical and practical. Theoretically, context is part of how humans interpret affective signals, not merely an optional source of extra information [1, 2, 3]. Practically, context-aware emotion recognition matters for intelligent systems operating in real environments, such as social robots, conversational agents, tutoring systems, and healthcare interfaces, because systems that analyse signals in isolation may fail when the same expression has different meanings in different situations [5, 7, 10, 15, 21].

Computational emotion recognition already turns situation context into model input in several ways. Image-based approaches represent context through visual or environmental cues [10, 11, 12]. Emotion recognition in conversation (ERC) represents context through conversational or interactional cues [13, 14, 15, 19]. Emotion-cause analysis represents context through causal event cues [20]. Recent LVLM work represents situation context through visual, linguistic, and commonsense reasoning [21], while LLM-based ERC also uses speaker characteristics that sit closer to sender context [22]. These four cue families provide the initial scope for the search strategy; they are not treated as final findings before the literature is analysed.

The resulting literature is fragmented in a way that existing surveys only partly explain. Prior work has established that context matters for emotion perception and automatic affect detection [5, 6], but it has not yet isolated how *situation context* is defined, what situation cues are used, and how those cues become computational representations. Different papers use terms such as scene context, social context, conversational context, interaction context, event context, emotion cause, and commonsense context, and they represent situation context either implicitly through temporal or multimodal features or explicitly through context branches, cause labels, graph structures, commonsense knowledge sources, or prompts. It is therefore unclear whether this terminological variety reflects different vocabulary for similar mechanisms or substantively different modelling choices. This review investigates that distinction.

The main research question of this review is:

> **How is situation context conceptualised and modelled in context-aware emotion recognition research?**

To answer this question, the review addresses five sub-questions:

- **SQ1:** What meanings of situation context appear in the selected computational emotion-recognition literature?
- **SQ2:** Which categories of situation-context cues are studied in the selected computational emotion-recognition literature?
- **SQ3:** How do data-driven emotion-recognition models represent situation context computationally?

- **SQ4:** Is situation context modelled implicitly (captured indirectly through sequence modelling, multimodal features, or temporal structure), explicitly (represented as a separate input, branch, label, graph, knowledge source, or prompt), or through a hybrid combination?
- **SQ5:** What main trends can be identified across the selected papers?

The remainder of this paper is structured as follows. Section 2 positions the review in relation to existing theory and survey work. Section 3 describes the search, screening, extraction, and synthesis methodology. Section 4 presents the results in relation to the five sub-questions. Section 5 discusses the cross-paper synthesis and open challenges. Section 6 addresses responsible research considerations. Section 7 concludes with the review limitations and recommendations for future work.

2 Background and Related Work

Existing work motivates context-aware emotion recognition, but it also leaves open the computational question addressed in this review. Barrett et al. [1, 2] challenge face-only interpretations of emotion by showing that facial movements are interpreted differently depending on visual scenes, other faces, culture, and broader situational information. This is especially important for the present review because Barrett et al. [2] do not merely argue that context sometimes improves emotion perception; they argue that facial movements are inherently ambiguous and cannot be reliably mapped to emotion categories without considering the situation in which they occur. Wieser and Brosch [3] systematise contextual influences on affective face processing across multiple levels, while Hess and Hareli [4] provide empirical evidence that contextual information can shift human emotion labels.

The psychological literature explains why context matters, but it does not determine how situation context should become a computational object. Barrett et al. [2], for example, make a strong case against simple face-to-emotion inference, but their broad treatment of context does not specify which situation cues should become model inputs, annotations, or evaluation targets. Wieser and Brosch [3] provide a useful psychological systematisation across three levels (visual scene, bodily, and social context), but they do not address practical modelling questions such as dataset availability, feature representation, or the boundary between situation context, sender context, and perceiver context in computational pipelines.

Survey work in automatic affect detection begins to connect psychological theory to computational modelling. Dudzik et al. [5] survey context in human emotion perception for automatic affect detection and examine how contextual factors appear in audiovisual databases. Groh and Picard [6] discuss context as an important but under-specified concept in automated affect recognition. These works show that context is relevant for computational systems, but they do not focus specifically on how *situation context* is operationalised across visual, conversational, causal, commonsense, and LLM/LVLM-based modelling approaches.

Unlike Dudzik et al. [5], whose focus is on contextual factors in audiovisual databases, and Groh and Picard [6], who discuss context as a broader underspecified issue in automated affect recognition, this review focuses on how situation context is made computational. The relevant evidence is found in datasets, annotations, model architectures, and evaluation assumptions. These are the points at which a paper's treatment of situation context becomes observable and comparable, rather than remaining a general theoretical claim.

Broader survey and mapping work confirms that contextual information is studied across affective-state recognition and context-based emotion recognition [8, 9]. Dorneles et al. [8] provide a systematic mapping of context awareness in affective-state recognition, while Abbas et al. [9] survey context-based emotion recognition across techniques, challenges, datasets, evaluation metrics, and applications. These works are useful for positioning the field, but they remain broader than the present review. They do not isolate *situation context* from sender and perceiver context or compare how situation context is operationalised across visual scene, conversational, emotion-cause, commonsense, and LLM/LVLM-based modelling traditions. For this reason, they are used here as related-work positioning rather than as part of the set of 19 papers included for data extraction.

This review shifts the question from whether context matters to how one specific form of context becomes computational. Instead of reviewing all forms of context, it separates situation context from sender and perceiver context and analyses how situation-related information is conceptualised and modelled across different computational subfields. The contribution is not the claim that context matters in general; that has already been established. The contribution is a structured synthesis of how situation context is made computational: as visual environment, dialogue history, interaction structure, causal event evidence, commonsense knowledge, or semantic situation reasoning.

3 Methodology

This review followed a structured literature review (SLR) approach because the research question asks for synthesis and comparison rather than model evaluation. PRISMA was used as a reporting scheme to make the single-researcher process transparent and reproducible [24]. PRISMA specifies what should be reported at each stage; it does not prescribe a search or screening procedure. The review consisted of five stages: defining paper eligibility, searching databases, screening and selecting papers, extracting data, and synthesising results. The PRISMA-style flow summary is shown in Figure 1. Full screening details, the final paper-selection table, extraction questions, and database-specific queries are reported in Appendix A and Appendix B.

A structured literature review was chosen over two alternative review designs. A systematic mapping study, as used by Dorneles et al. [8], would have produced a broader categorical map of the field, but it would not have supported the thematic synthesis and cross-paper comparison required by the research question. A scoping review would have been appropriate for charting the extent of the literature, but the research question asks specifically how situation context is

defined and modelled, which requires structured extraction against fixed sub-questions. The SLR design, with predefined eligibility criteria and an extraction framework derived from the sub-questions, was therefore the most appropriate choice for a question that asks for comparison rather than coverage.

3.1 Eligibility criteria

Table 1 summarises the eligibility criteria used to keep the review focused on situation context rather than context in the broadest sense. The inclusion criteria define the scope: a paper had to address emotion or affect recognition, model or discuss context as part of emotion interpretation, and involve situation-related contextual information. The exclusion criteria were applied to remove papers that fell outside this scope during screening.

	Criterion
Inclusion (scope definition)	
I1	Studies emotion perception, emotion recognition, affect recognition, affective computing, or emotion recognition in conversation
I2	Discusses or models context as part of emotion interpretation
I3	Involves situation-related context: scene, physical environment, event, activity, task setting, conversational history, interaction setting, social role, emotion cause, or commonsense/situational knowledge
I4	Contains sufficient methodological or conceptual detail to extract how situation context is conceptualised or modelled
I5	Written in English and available in full text
Exclusion	
E1	Focuses only on facial-expression, speech-emotion, or physiological emotion recognition without situation-context modelling
E2	Focuses only on general sentiment analysis, sender context, or perceiver context without situation context
E3	Mentions "context" only superficially
E4	Duplicate, non-academic source, or lacking sufficient detail for extraction
E5	Published before 2017 unless retained as foundational theory or empirical context work

Table 1: Eligibility criteria. Borderline papers were retained during title/abstract screening and assessed at full-text level, because situation context is not always described using the exact phrase.

3.2 Search and screening process

The search strategy combined three concept blocks: emotion-recognition task, situation-context component, and modelling or review focus. These blocks were used to construct targeted database queries:

> (**emotion or affect recognition terms**) AND (**situation-context terms**) AND (**modelling or review terms**)

This structure was implemented as five targeted search strata per database: theory and survey work, visual scene-context modelling, conversational-context modelling, emotion-cause or commonsense/event-context modelling, and LLM/LVLM-based context reasoning. These strata were necessary because the relevant subfields use different terminology and because a single broad query produced very noisy results. They were not treated as final theoretical categories before extraction. The same inclusion and exclusion criteria in Table 1 were applied regardless of which query found a record. The full database-specific queries are reported in Appendix B. Snowballing was used as a supplementary method: two papers were identified through citation snowballing from included papers and one paper was recommended by the project supervisor, giving three additional papers. Google Scholar and Semantic Scholar were consulted informally for orientation but were not used for bulk screening because their result sets are large and their ranking procedures are not reproducible.

Database searches returned 298 records in total. After removing 44 duplicates, 254 unique records were screened at title and abstract level. Three supplementary records were assessed separately after database screening: two identified through citation snowballing and one recommended by the project supervisor. The full PRISMA-style flow summary is shown in Figure 1.

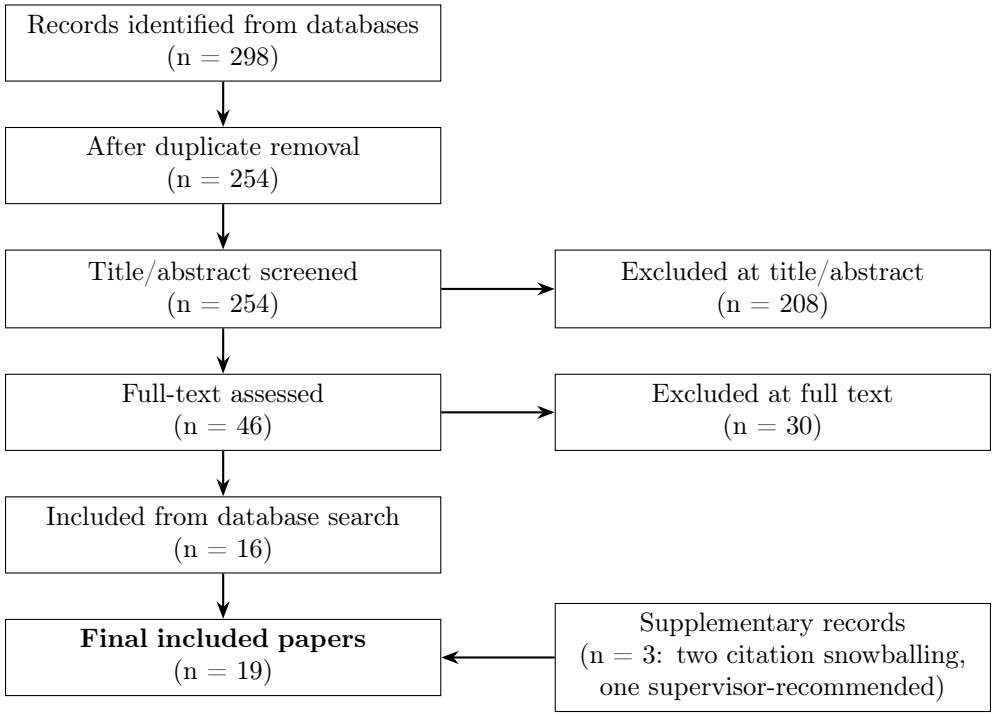

Figure 1: PRISMA-style flow diagram reporting the screening process.

3.3 Data extraction and synthesis

For each included paper, structured information was extracted into a spreadsheet following the extraction questions in Table 5 (Appendix A). The extraction

recorded bibliographic information, paper type, emotion-recognition task, definition of context, type of situation context, dataset, modality, model, context representation, context integration method, implicit, explicit, or hybrid modelling strategy, key findings, limitations, and relevance to the research question. This structure was derived from the five sub-questions of the review and informed by existing context-focused survey and mapping work [5, 8].

The analysis was thematic and comparative. The initial situation-cue families came from the search scope, but the more specific categories used in the results were assigned during extraction by comparing how each paper defined and represented the cue. Papers were then grouped by modelling strategy so that the review could compare not only what context means, but also how it is made available to a model. *Implicit context modelling* refers to approaches where situation context is available in the input, sequence, or dataset structure but is not represented as a separate model component or annotation. *Explicit context modelling* refers to approaches where context is represented as a separate input, model branch, label, metadata field, graph relation, knowledge source, or prompt component. *Hybrid context modelling* refers to approaches that combine learned representations with an explicit structural choice, such as graph relations, modality-specific branches, knowledge sources, or prompts. Because explicit structures and learned representations can coexist, these categories describe the dominant modelling mechanism rather than mutually exclusive model classes. These three categories are used in SQ4 and are defined here at first use; they are introduced in the sub-questions above and applied systematically in the results.

4 Results

This section presents the results in relation to the five sub-questions. Rather than summarising each paper individually, it compares papers along the dimensions needed to answer the research question: how they define situation context, which situation cues they study, how those cues are represented computationally, whether context is modelled implicitly, explicitly, or through a hybrid strategy, and what trends or gaps emerge across the literature.

4.1 SQ1: Meanings of situation context

The selected computational emotion-recognition literature does not treat situation context as one shared object. Instead, situation context is defined indirectly through task setup, dataset design, model architecture, or the type of contextual cue being used. In theory and survey work, context is described broadly as the information that shapes how emotional signals are interpreted [1, 2, 3, 5, 6]. This framing is conceptually useful, but it does not immediately translate into a concrete computational representation.

Visual emotion-recognition work operationalises situation context spatially: the relevant situation is the visible environment around the person. EMOTIC-style work uses the full image, surrounding objects, scene, body posture, and

activity cues to interpret a person's affective state [10, 11]. CAER-Net makes this distinction more architectural by separating face regions from context regions [12]. In these papers, situation context is mainly spatial and visual.

Conversational emotion-recognition work operationalises situation context temporally and interactionally. Previous utterances, speaker turns, party states, speaker interactions, and conversational dependencies define the situation in which the current emotion occurs [13, 14, 15, 17, 18, 19]. In emotion-cause analysis, the situation is the event or cause that explains why an emotion is expressed [20]. In LVLM-based work, situation context becomes broader semantic situation understanding, combining visual, linguistic, social, and commonsense cues [21].

The answer to SQ1 is therefore that situation context is best understood as a task-dependent modelling construct. It consistently refers to information outside the isolated emotional signal, but its concrete meaning changes across subfields: it may be spatial, temporal, interactional, causal, or semantic.

4.2 SQ2: Types of situation-context cues

The selected papers study six categories of situation-context cues. These categories refine the four initial search cue families described in the introduction and methodology: they were assigned during extraction by comparing what each paper actually used as contextual evidence. First, *visual scene context*: scene, object, body, and activity cues in the image background [10, 11, 12]. Second, *conversational context*: dialogue history, speaker turns, utterance sequence, and interaction structure [13, 14, 15, 17, 18, 19]. Third, *social-interaction context*: speaker relationships, party states, and inter-speaker dependencies, appearing especially in graph-based conversational models [15, 18].

Fourth, *emotion-cause and event context*: the cause or event that explains the emotion [20]. This is a particularly explicit form of situation context because the cue is not merely background information; it is the explanatory situation that makes the emotion understandable. Fifth, *commonsense context*: knowledge about mental states, causal relations, and likely social interpretations used to enrich the immediate dialogue context [16]. Sixth, *semantic situation context*: broader cues used by LVLM approaches, including visual descriptions, social context, natural-language prompts, and commonsense reasoning [21].

These analytical categories can overlap. Social-interaction cues, for example, frequently occur within conversational models. Nevertheless, the six categories show that situation context cannot be reduced to one modality or representation. The same broad concept appears as image background in visual emotion recognition, previous utterances in ERC, causal evidence in emotion-cause analysis, and semantic reasoning in LLM/LVLM systems.

4.3 SQ3: Computational representations of situation context

Situation context becomes computational when it is encoded as something a model or benchmark can use: an input feature, sequence, relation, label, knowledge source, or prompt. Visual models represent context as image features,

person-scene inputs, or separate face/context regions. EMOTIC represents affect through both person-level and image-level information [10, 11], while CAER-Net explicitly builds a dual-stream architecture that separates facial information from contextual visual regions [12]. This makes context more visible than face-only modelling, but it also ties situation context mainly to what is visible in the image.

Conversational models progressively make dialogue context more structured. DialogueRNN models dialogue history and party states through attentive recurrent mechanisms [14]. DialogueGCN represents utterances and speaker dependencies as graph relations [15]. MMGCN extends graph-based modelling across multimodal information [18]. DialogXL uses XLNet-style memory to handle longer conversational dependencies [19]. DialogueCRN frames ERC as multi-turn contextual reasoning rather than simple utterance classification [17]. The shift from recurrent memory to graphs and long-range transformer memory makes interaction structure more explicit, but it also defines situation context through the relations or histories that the architecture is designed to capture.

Emotion-cause analysis makes the explanatory role of situation context most direct by representing context through emotion-cause pairs, cause utterances, or multimodal cause information [20]. Commonsense-based ERC extends the immediate dialogue with external knowledge about mental states and causal relations [16]. LVLM approaches broaden the representation further through prompts, generated descriptions, natural-language explanations, and semantic reasoning over visual and linguistic inputs [21]. LLM-based ERC can instead use prompts to generate speaker-characteristic information, as in LaERC-S [22].

Across these approaches, the representation of situation context becomes richer, but also harder to compare. Each representation answers a different version of the context problem.

4.4 SQ4: Implicit, explicit, and hybrid modelling strategies

The selected papers show a continuum from implicit to explicit context modelling. *Implicit context modelling* appears when situation context is included in the input or sequence but not separately represented. For example, a recurrent or transformer-based ERC model may encode previous utterances without explicitly labelling them as situation context [14, 19]. This can improve performance, but it makes it difficult to inspect which contextual cues influenced the prediction.

Explicit context modelling appears when situation context is represented as a separate input, branch, relation, label, knowledge source, or prompt component. CAER-Net explicitly separates face and context regions [12]. DialogueGCN explicitly models conversational dependencies through graph relations [15]. COSMIC uses a pretrained commonsense-inference model to generate explicit features describing each speaker's likely mental states, such as intent, motivation, and emotional reaction, and fuses these features with the dialogue representation through dedicated GRU-based reasoning modules, making commonsense knowledge an additional explicit source of contextual reasoning [16]. Emotion-cause analysis explicitly identifies causal evidence linked to the emotion [20]. Explicit modelling makes context easier to name and inspect, but it also requires stronger assumptions about which contextual relations are relevant.

Hybrid context modelling combines learned representations with explicit structure. A model is classified as hybrid when it still learns distributed representations, but the architecture or input design explicitly constrains which contextual relation or knowledge source should be available. Graph-based models such as MMGCN are hybrid because they learn utterance embeddings while imposing speaker or utterance relations through graph topology [18]. Similarly, multimodal models distinguish modalities explicitly while learning fused features implicitly; LLM/LVLM approaches use explicit prompts while relying on opaque internal representations [21, 22].

Within the reviewed timeline, the selected papers suggest a movement toward more explicit and semantically rich forms of context modelling. This pattern is visible in the shift from recurrent approaches to graph-based models and then to LLM/LVLM-based systems [14, 15, 21, 22]. The trend does not automatically solve interpretability or evaluation problems.

4.5 SQ5: Trends across the selected papers

Three trends emerge across the selected papers. The first is a movement from isolated expression recognition toward context-aware recognition. Theory papers establish the motivation [1, 2], while datasets such as EMOTIC and MELD create practical settings where context can be modelled [10, 13]. The second is a shift from feature-level context to structured context. Later models increasingly represent context through graphs, memory mechanisms, reasoning components, or commonsense knowledge [15, 16, 18]. The third is a movement toward semantic reasoning, especially in LLM/LVLM approaches that use natural language, commonsense knowledge, and generated explanations [21, 22].

Within the selected literature, these trends suggest increasing use of context and more structurally explicit representations. At the same time, they show why comparison across subareas remains difficult: each modelling tradition develops around a different task, modality, and representation of situation context. The broader consequences of this fragmentation are discussed as open challenges in Section 5.3.

5 Discussion

5.1 Cross-paper synthesis

The main research question asked how situation context is conceptualised and modelled in context-aware emotion recognition research. The central answer is that situation context is not a single variable or one fixed type of input. Instead, it is operationalised differently depending on the task, modality, and modelling tradition. Table 2 summarises the conceptual framing and the six situation-context categories identified in the review.

The synthesis shows that situation context is best understood as a task-dependent modelling construct. In visual emotion recognition, it is usually the physical scene around the person. In conversation, it is the dialogue history and interaction structure. In emotion-cause analysis, it is the event that explains the

Paper group	Situation context type	Representation	Modelling strategy
Theory/survey	General situational, social, and environmental context	Conceptual definitions and theoretical framing	Conceptual
Visual scene context	Scene, objects, body, activity, image background	Visual features, face/context regions, person-scene inputs	Explicit or hybrid
Conversational context	Dialogue history, speaker turns, utterance sequence	Sequential states, recurrent memory, transformer context	Implicit, explicit, or hybrid
Social-interaction context	Speaker relations, party states, inter-speaker dependencies	Graph relations, party states, speaker embeddings	Explicit or hybrid
Commonsense conversation	Mental states, causal relations, social knowledge	Commonsense knowledge and dialogue context	Explicit or hybrid
Emotion-cause/event context	Cause, event, situation explaining emotion	Emotion-cause pairs, cause utterances, multimodal cause information	Explicit
LLM/LVLM reasoning	Full semantic situation, visual-social context, dialogue context	Prompts, generated descriptions, language-based reasoning	Explicit or hybrid

Table 2: Synthesis of the conceptual framing and the six situation-context categories identified in SQ2.

emotion. In commonsense and LLM/LVLM approaches, it becomes a broader semantic representation of the situation.

These are not merely different labels for the same thing. They imply different datasets, annotations, architectures, and evaluation criteria. Table 2 also makes visible a structural gap in the literature: most approaches specialise in one form of situation context, while relatively few integrate visual scene context, conversational history, emotion-cause information, and commonsense reasoning within a single modelling framework.

This gap makes cross-context integration a promising direction for future work. Shared taxonomies and benchmarks would be needed to support it.

The most important implication is that context-aware emotion recognition is not a single research line with a shared definition of context. It is a collection of related modelling traditions that operationalise situation context according to their available data and task constraints. This explains why comparison across papers is difficult: a model that uses image background, a model that uses dialogue history, and a model that identifies emotion causes may all be context-aware, but they address different versions of the context problem. A first step toward comparability would be to agree on a shared taxonomy of situation-context types and to develop cross-task evaluation protocols.

5.2 Implicit versus explicit context modelling

The implicit–explicit distinction helps compare these modelling traditions without treating them as equivalent. Implicit approaches can exploit context without requiring additional annotations, but they make it harder to verify what the model has learned. Explicit approaches make the role of context more visible, but they often require stronger assumptions, additional labels, or architectural decisions about what counts as context. Hybrid approaches are common in recent work because they combine learned representations with explicit structure, for example through graph relations, multimodal fusion, commonsense knowledge, or prompting.

This distinction also exposes a methodological weakness in the field. Improved accuracy does not necessarily prove that a model uses context in a meaningful way. A model may benefit from dataset bias, speaker identity cues, annotation artefacts, or correlations that are not genuinely situational.

Proponents of implicit approaches may view this differently. The premise of many implicit models is that with sufficient data and capacity, the model can learn the relevant contextual associations without explicit specification. Nevertheless, without targeted ablations or context-dropout evaluations, it remains unclear whether performance gains stem from situation context or from spurious correlations.

Context-aware emotion recognition therefore needs evaluation methods that test not only whether context improves performance, but also whether the model relies on the intended contextual evidence.

This observation suggests a testable direction for future work. If situation context is the genuine driver of performance gains in implicit models, then selectively removing contextual information at inference time, through context-dropout or input ablation, should produce a larger performance drop than removing non-contextual features of comparable representational size. If the performance drop is small or inconsistent, the model may be exploiting distributional shortcuts rather than situation context. This test has not been systematically applied across the papers reviewed, and its absence is an open methodological gap in the field.

5.3 Open challenges

The four open challenges below follow directly from the six situation-context categories and three modelling strategies identified in the results.

The first open challenge is conceptual fragmentation. As shown in Table 2, papers across the six categories use context without clearly specifying which dimension they mean. This weakens comparability and makes it difficult to build cumulative knowledge across subfields.

The second open challenge is integration across context types. Visual models focus on scene context, conversational models focus on dialogue history, and cause-analysis models focus on event explanations. Real emotional situations often involve all of these at once. Cross-task benchmarks that test multiple forms of situation context simultaneously are still limited.

The third open challenge is evaluation. Many models show improved classification performance when context is included, but fewer papers directly evaluate whether the model uses context in a human-interpretable or causally meaningful way. This is especially important for LLM/LVLM approaches, where generated explanations may sound plausible without faithfully reflecting the basis of the model's prediction.

The fourth open challenge is the boundary between situation context and sender context. LaERC-S illustrates this problem [22]. Speaker characteristics may improve ERC, but they are not situation context in the strict sense. Future work should define more clearly whether contextual information refers to the person expressing the emotion, the person interpreting it, or the situation in which it occurs.

6 Responsible Research

Because this review depends on search choices, screening decisions, and the interpretation of borderline papers, responsible research begins with transparent and reproducible reporting. To support this aim, the database-specific queries, result counts, PRISMA-style flow summary, inclusion and exclusion logic, final paper-selection table, extraction questions, and AI-use statement are reported in the appendix. Reproducibility alone, however, is not sufficient: the use of situation context in emotion recognition also raises ethical concerns about how systems interpret people and the settings in which they operate.

The main ethical concern is that context-aware emotion recognition systems may appear more reliable than they actually are. Adding context can improve predictions, but it can also create new risks if the system infers emotions from social setting, identity-related cues, cultural assumptions, or situational stereotypes. For example, a model may treat a workplace, classroom, hospital, or conflict setting as evidence for particular emotions even when the person's actual emotional state is different. This can produce overconfident or intrusive interpretations.

This concern is also reflected in policy. Article 5(1)(f) of the European Union Artificial Intelligence Act prohibits the use of AI systems to infer emotions in workplace and education institutions, except for medical or safety reasons [23].

Dataset bias is another responsible-research issue. Contextual cues are not neutral: scenes, social roles, activities, and dialogue patterns may reflect cultural, demographic, or annotation biases. This concern is consistent with context-focused work showing that affective datasets contain different contextual factors and that context remains under-specified in automated affect recognition [5, 6]. It is also consistent with psychological evidence that contextual information can shift emotion labels [2, 4]. If such dataset-specific or culturally specific patterns are learned by a model, context-aware systems may reproduce unfair assumptions about people and situations. This is especially problematic in high-stakes settings such as healthcare, education, hiring, surveillance, or social robotics.

Finally, explainability must be treated carefully. Explicit context representations, cause labels, graph relations, and LLM/LVLM-generated explanations

may make a system appear interpretable, but they do not automatically prove that the model's prediction is faithful to the intended contextual evidence. Responsible use of context-aware emotion recognition therefore requires cautious evaluation, clear reporting of limitations, and avoidance of claims that a system can directly read or know a person's emotional state.

7 Conclusion

This review answered the research question by showing that situation context is not conceptualised or modelled in one standard way in context-aware emotion recognition. Instead, it is operationalised differently depending on the task and modality: as visual scene and activity context, dialogue and interaction context, causal event context, commonsense knowledge, or broader semantic situation understanding.

Across these approaches, the main pattern is a movement from implicit contextual representations toward more explicit and structured forms of modelling, including context branches, graph relations, cause annotations, commonsense knowledge sources, and prompts. This diversity shows that situation context is useful for emotion recognition, but it also makes comparison across subfields difficult.

The main contribution of this review is therefore the synthesis that fragmentation is not merely terminological, but structural: different subfields operationalise context through different datasets, annotations, model architectures, and evaluation assumptions.

7.1 Limitations and threats to validity

This review has several limitations. First, it was conducted by a single researcher, which increases the risk of subjective judgement during screening and extraction. Second, only English-language papers were included, which may exclude relevant work in other languages. Third, the review is based on a final set of 19 papers, so it provides a focused synthesis rather than an exhaustive map of every context-related emotion-recognition paper. These limitations create three main threats to validity.

Construct validity concerns whether the eligibility criteria and extraction questions genuinely capture situation context as defined in the research question. The boundary between situation context, sender context, and perceiver context is not always sharp in the literature, and borderline cases such as LaERC-S required judgement calls [22]. A second independent coder would reduce this risk, but was not feasible within the scope of a single-researcher project.

Internal validity concerns whether the synthesis categories reflect patterns in the literature rather than the reviewer's prior assumptions. To mitigate this risk, the six cue categories were not fixed before extraction. They were assigned during extraction by comparing what each paper actually used as contextual evidence, and the final categories are reported in the results rather than imposed as fixed theoretical categories in the methodology.

Conclusion validity concerns whether the open challenges generalise beyond the 19 included papers. The review covers the main terminology clusters and modelling traditions through the five search strata, but it does not claim exhaustiveness. The conclusions about structural fragmentation are therefore framed as an analytical synthesis of the included papers rather than as a statistical claim about the entire field.

7.2 Future work

Four directions for future work follow from the open challenges identified in this review. First, the field would benefit from a shared taxonomy of situation context that distinguishes the six categories identified in this review: visual scene, conversational, social-interaction, emotion-cause/event, commonsense, and semantic situation context. Such a taxonomy would directly address the current conceptual fragmentation.

Second, cross-task benchmarks that combine multiple forms of situation context would make it possible to compare models beyond isolated visual, conversational, or cause-analysis settings. Third, evaluation methods should test not only whether context improves classification performance, but also whether models use contextual information in meaningful, interpretable, and causally relevant ways. This is especially important in LLM/LVLM-based systems, where explanations may sound plausible without being faithful to the model's actual reasoning.

Fourth, future work should establish clearer definitional boundaries between situation context, sender context, and perceiver context in computational pipelines, so that borderline cases such as LaERC-S can be positioned and compared without ambiguity.

As emotion-recognition systems move from controlled benchmarks into real-world interaction, modelling situation context matters not only for accuracy, but also for building systems whose affective interpretations are more appropriate, transparent, and sensitive to the situations in which people actually communicate. Such systems should be evidence-aware, situation-sensitive, and honest about uncertainty rather than claiming to read emotions directly from isolated signals.

A Screening, selection, extraction, and AI-use details

A.1 PRISMA-Style Flow

The PRISMA-style flow diagram in Figure 1 (main text) summarises the full screening process. Table 3 provides the numerical breakdown for reference.

PRISMA stage	Records	Notes
Records identified from databases	298	Scopus 148; IEEE Xplore 124; ACM Digital Library 26
Duplicates removed	−44	Duplicate records across database exports
Unique database records screened at title/abstract	254	298 database records − 44 duplicates
Excluded at title/abstract	−208	No situation-context evidence in title/abstract: 176; no relevant title terms: 22; off-topic: 7; pre-2017 non-foundational: 2; non-English: 1
Database reports assessed at full text	46	254 screened records − 208 title/abstract exclusions
Excluded at full text	−30	No situation context modelled: 21; outside scope or insufficient detail: 9
Database studies included	16	46 full-text reports − 30 full-text exclusions
Additional studies included through supplementary identification	3	Two from citation snowballing and one supervisor-recommended paper
Final studies included in review	**19**	16 database studies + 2 citation-snowballing studies + 1 supervisor-recommended study

Table 3: PRISMA-style screening table.

A.2 Final paper selection

The final extraction set contained 19 papers. The search-route labels in Table 4 indicate where the papers were found; they should not be read as final theoretical categories fixed before analysis. Six papers were included as theory or survey papers because they provide the conceptual basis for understanding context in emotion perception and automatic affect detection. Three papers were included through visual scene-context queries because they model contextual information from the visual environment. Seven papers were included through conversational-context queries because they model dialogue history, speaker interaction, multimodal conversational structure, or commonsense conversational context. One paper was included through the emotion-cause/event-context query route. Two recent papers were included through LLM/LVLM query routes because they represent the newer direction of semantic and language-based context reasoning. One of these, LaERC-S, is borderline because its main focus is speaker characteristics, which overlaps more strongly with sender context than situation context; it was therefore used to compare situation-context modelling with recent LLM-based ERC work rather than as primary evidence for visual or event-based situation context.

Paper	Reason for inclusion	Search route
Barrett et al. (2011) [1]	Reviews consistent context effects on emotion perception; calls into question face-only inference.	Scopus-Q1 / theory route
Barrett et al. (2019) [2]	Argues situation context is necessary, not optional, for emotion perception from facial movements.	Snowballing ← [1]
Wieser and Brosch (2012) [3]	Reviews and systematises contextual influences on affective face processing across multiple levels.	Scopus-Q1 / theory route
Hess and Hareli (2015) [4]	Empirical study showing context alters human emotion labelling.	IEEE-Q1 / theory route
Dudzik et al. (2019) [5]	Bridges human context perception to automatic affect detection; surveys audiovisual databases.	Supervisor-recommended / snowballing seed
Groh and Picard (2021) [6]	Identifies context as a key under-specified concept in computational affect recognition.	Snowballing ← [5]
Kosti et al. (2017) [10]	Introduces EMOTIC dataset; visual scene as explicit situation context around person.	Scopus/IEEE-Q2 / visual route
Kosti et al. (2020) [11]	Extends EMOTIC with improved person+scene modelling; explicit visual context branch.	Scopus/IEEE-Q2 / visual route
Lee et al. (2019) [12]	CAER-Net: dual-stream face + visual context region; clear explicit visual situation-context model.	IEEE-Q2 / visual route
Poria et al. (2019) [13]	MELD dataset enabling dialogue-level situation-context modelling across multi-party turns.	Scopus/ACM-Q3 / conversational route
Majumder et al. (2019) [14]	Models dialogue history and party states as sequential situational context via attentive RNN.	Scopus/ACM-Q3 / conversational route

Paper	Reason for inclusion	Search route
Ghosal et al. (2019) [15]	DialogueGCN: graph-based modelling of speaker-utterance dependencies as situation context.	ACM-Q3 / conversational route
Ghosal et al. (2020) [16]	COSMIC: commonsense knowledge combined with dialogue context; explicit commonsense modelling.	Scopus-Q4 / commonsense route
Hu et al. (2021a) [17]	DialogueCRN: multi-turn contextual reasoning over dialogue as situation context.	Scopus/ACM-Q3 / conversational route
Hu et al. (2021b) [18]	MMGCN: multimodal GCN over utterances and speakers; graph-based conversational context.	Scopus/ACM-Q3 / conversational route
Shen et al. (2021) [19]	DialogXL: XLNet-style memory for long-range conversational context dependencies.	Scopus/ACM-Q3 / conversational route
Wang et al. (2024) [20]	SemEval-2024 Task 3: explicit multimodal emotion-cause modelling as causal situation context.	Scopus-Q4 / cause route
Etesam et al. (2024) [21]	LVLMs for contextual emotion recognition; combines visual, social, and commonsense context.	IEEE-Q5 / LVLM route
Fu et al. (2025) [22]	LaERC-S: LLM-based ERC; borderline (speaker characteristics); retained as LLM comparison.	Scopus-Q5 / LLM route

Table 4: Final included papers and their search route. Search routes are reported for transparency and were not treated as final theoretical categories before extraction.

A.3 Extraction questions

To make extraction systematic, each sub-question was mapped to extraction questions in Table 5.

A.4 Use of AI-assisted tools

AI-assisted tools were used during the project for search support, note organisation, extraction assistance, and language editing. They were not used as a replacement for manual screening. Any AI-assisted summaries or suggested

SQ	Extraction questions
SQ1	How does the paper define, describe, or motivate situation context? Does it distinguish situation context from other forms of context?
SQ2	Which category of situation-context cue is used: visual scene, conversational, social-interaction, emotion-cause/event, commonsense, or semantic situation context?
SQ3	How is situation context represented computationally: visual features, textual history, temporal sequence, graph structure, labels, metadata, prompts, or knowledge representations?
SQ4	Is context modelled implicitly, explicitly, or as a hybrid? How is it integrated into the model: fusion, attention, temporal modelling, graph modelling, commonsense reasoning, prompting, or conditioning?
SQ5	What trends does the paper reveal, and how do these trends connect to the open challenges discussed in the synthesis?

Table 5: Extraction questions mapped to the review sub-questions.

classifications were checked against the original papers before being used. Extracted claims were verified against the original paper, and AI tools did not decide whether a paper was included or excluded.

B Database-specific search queries and result counts

The general search structure in Section 3 was used as the design template for the database queries. In practice, it was implemented as five targeted query variants in each main database. The variants were search strata designed to capture terminology used by different subfields: Q1 theory and survey work, Q2 visual scene-context modelling, Q3 conversational-context modelling, Q4 emotion-cause or commonsense/event-context modelling, and Q5 LLM/LVLM-based context reasoning. They were not treated as final findings before extraction. This targeted structure made the screening process manageable while still preserving the same conceptual logic across databases. The counts below are the raw database results before duplicate removal.

B.1 IEEE Xplore queries

IEEE Q1: theory and contextual emotion perception (21 results)

("All Metadata":"emotion perception" OR "All Metadata":"affect recognition" OR "All Metadata":"affective computing") AND ("All Metadata":"situation context" OR "All Metadata":"situational context" OR "All Metadata":"context in emotion" OR "All Metadata":"contextual influence")

IEEE Q2: visual scene-context modelling (11 results)

(("Document Title":"context-aware emotion recognition" OR "Document Title":"emotion recognition in context" OR "Document Title":"EMOTIC" OR "Document Title":"CAER-Net") AND ("All Metadata":scene OR "All Metadata":"visual context" OR "All Metadata":"body pose"))

IEEE Q3: conversational-context modelling (12 results)

(("Document Title":"DialogueRNN" OR "Document Title":"DialogueGCN" OR "Document Title":"DialogueCRN" OR "Document Title":"MMGCN" OR "Document Title":"DialogXL" OR "Document Title":"MELD"))

IEEE Q4: emotion-cause and commonsense/event context (19 results)

(("Document Title":"emotion cause" OR "Document Title":"emotion-cause" OR "Document Title":"COSMIC" OR "Document Title":"commonsense" OR "Document Title":"SemEval-2024") AND ("All Metadata":"emotion recognition" OR "All Metadata":"affect recognition") AND ("All Metadata":conversation OR "All Metadata":multimodal))

IEEE Q5: LLM/LVLM context reasoning (61 results)

("All Metadata":"emotion recognition" OR "All Metadata":"affect recognition") AND ("All Metadata":"large language model" OR "All Metadata":"LVLM" OR "All Metadata":"vision-language model") AND ("All Metadata":"context" OR "All Metadata":"situation")

B.2 Scopus queries

Scopus Q1: theory and contextual emotion perception (41 results)

TITLE-ABS-KEY (("emotion perception" OR "affect recognition" OR "affective computing" OR "affective face processing" OR "faces in context") AND ("situation context" OR "situational context" OR "contextual influence" OR "context in emotion"))

Scopus Q2: visual scene-context modelling (14 results)

TITLE ("context-aware emotion recognition" OR "emotion recognition in context" OR "EMOTIC" OR "CAER-Net") AND TITLE-ABS-KEY ("scene" OR "visual context" OR "body pose")

Scopus Q3: conversational-context modelling (15 results)

TITLE ("DialogueRNN" OR "DialogueGCN" OR "DialogueCRN" OR "MMGCN" OR "DialogXL" OR "MELD") AND TITLE-ABS-KEY ("emotion")

Scopus Q4: emotion-cause and commonsense/event context (40 results)

TITLE ("emotion cause" OR "emotion-cause" OR "COSMIC" OR "commonsense") AND TITLE-ABS-KEY ("emotion recognition" OR "affect recognition") AND TITLE-ABS-KEY (conversation OR multimodal)

Scopus Q5: LLM/LVLM context reasoning (38 results)

TITLE ("large vision language model" OR "large language model" OR "LVLM" OR "vision-language model" OR "LaERC-S" OR "LLM-based emotion recognition") AND TITLE-ABS-KEY ("emotion recognition" OR "affect recognition" OR "affective computing") AND TITLE-ABS-KEY ("situation context" OR "contextual reasoning" OR "scene understanding" OR "social context" OR "speaker" OR "conversation")

B.3 ACM Digital Library queries

ACM Q1: theory and contextual emotion recognition (4 results)

[Abstract: "context-aware emotion recognition"] OR [Abstract: "contextual emotion recognition"]

ACM Q2: visual scene-context modelling (5 results)

[Abstract: "context-aware emotion recognition"] OR [Abstract: "emotion recognition in context"] OR [Abstract: "emotic"] OR [Abstract: "caer-net"]

ACM Q3: conversational-context modelling (14 results)

((Title:("emotion recognition in conversation" OR "emotion recognition in conversations" OR "emotion detection in conversations" OR "emotion identification in conversations" OR "conversational emotion recognition") OR Abstract:("emotion recognition in conversation" OR "emotion recognition in conversations" OR "emotion detection in conversations" OR "emotion identification in conversations" OR "conversational emotion recognition") OR Keyword:("emotion recognition in conversation" OR "emotion recognition in conversations" OR "emotion detection in conversations" OR "emotion identification in conversations" OR "conversational emotion recognition")) AND (Title:(MELD OR DialogueRNN OR DialogueGCN OR DialogueCRN OR MMGCN OR DialogXL OR COSMIC) OR Abstract:(MELD OR DialogueRNN OR DialogueGCN OR DialogueCRN OR MMGCN OR DialogXL OR COSMIC) OR Keyword:(MELD OR DialogueRNN OR DialogueGCN OR DialogueCRN OR MMGCN OR DialogXL OR COSMIC)))

ACM Q4: emotion-cause and commonsense/event context (2 results)

[[Title: "emotion cause"] OR [Title: "emotion-cause"] OR [Title: "cosmic"] OR [Title: "commonsense"] OR [Title: "semeval-2024"]] AND [[Abstract: "emotion recognition"] OR [Abstract: "affect recognition"]] AND [[Abstract: conversation] OR [Abstract: multimodal]]

ACM Q5: LLM/LVLM context reasoning (1 result)

[[Abstract: "emotion recognition"] OR [Abstract: "affect recognition"]] AND [[Abstract: "large language model"] OR [Abstract: "lvlm"] OR [Abstract: "vision-language model"]] AND [[Abstract: "context"] OR [Abstract: "situation"] OR [Abstract: "contextual reasoning"]]

References

[1] L. F. Barrett, B. Mesquita, and M. Gendron. Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290, 2011. https://doi.org/10.1177/0963721411422522

[2] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68, 2019. https://doi.org/10.1177/1529100619832930

[3] M. J. Wieser and T. Brosch. Faces in context: A review and systematization of contextual influences on affective face processing. *Frontiers in Psychology*, 3:471, 2012. https://doi.org/10.3389/fpsyg.2012.00471

[4] U. Hess and S. Hareli. The influence of context on emotion recognition in humans. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6, 2015. https://doi.org/10.1109/FG.2015.7284842

[5] B. Dudzik, M.-P. Jansen, F. Burger, F. Kaptein, J. Broekens, D. K. J. Heylen, H. Hung, M. A. Neerincx, and K. P. Truong. Context in human emotion perception for automatic affect detection: A survey of audiovisual databases. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 206–212, 2019. https://doi.org/10.1109/ACII.2019.8925446

[6] M. Groh and R. Picard. Context in automated affect recognition. In *Meaning in Context: Pragmatic Communication in Humans and Machines Workshop at the 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021. https://mattgroh.com/pdfs/context_automated_affect_recognition.pdf

[7] A. Marpaung and A. Gonzalez. Can an affect-sensitive system afford to be context independent? In P. Brézillon, R. Turner, and C. Penco, editors, *Context in Computing*, Lecture Notes in Computer Science, volume 10257, pages 454–467. Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-57837-8_38

[8] S. O. Dorneles, R. Francisco, D. N. F. Barbosa, and J. L. V. Barbosa. Context awareness in recognition of affective states: A systematic mapping of the literature. *International Journal of Human–Computer Interaction*, 39(8):1563–1581, 2023. https://doi.org/10.1080/10447318.2022.2062549

[9] R. Abbas, B. Ni, R. Ma, T. Li, Y. Lu, and X. Li. Context-based emotion recognition: A survey. *Neurocomputing*, 618:129073, 2025. https://doi.org/10.1016/j.neucom.2024.129073

[10] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza. Emotion recognition in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1960–1968, 2017. https://doi.org/10.1109/CVPR.2017.212

[11] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza. Context based emotion recognition using EMOTIC dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2755–2766, 2020. https://doi.org/10.1109/TPAMI.2019.2916866

[12] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10142–10151, 2019. https://doi.org/10.1109/ICCV.2019.01024

[13] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 527–536, 2019. https://doi.org/10.18653/v1/P19-1050

[14] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria. DialogueRNN: An attentive RNN for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1):6818–6825, 2019. https://doi.org/10.1609/aaai.v33i01.33016818

[15] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, 2019. https://doi.org/10.18653/v1/D19-1015

[16] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, 2020. https://doi.org/10.18653/v1/2020.findings-emnlp.224

[17] D. Hu, L. Wei, and X. Huai. DialogueCRN: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP)*, pages 7042–7052, 2021. https://doi.org/10.18653/v1/2021.acl-long.547

[18] J. Hu, Y. Liu, J. Zhao, and Q. Jin. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP)*, pages 5666–5675, 2021. https://doi.org/10.18653/v1/2021.acl-long.440

[19] W. Shen, J. Chen, X. Quan, and Z. Xie. DialogXL: All-in-one XLNet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13789–13797, 2021. https://doi.org/10.1609/aaai.v35i15.17625

[20] F. Wang, H. Ma, R. Xia, J. Yu, and E. Cambria. SemEval-2024 Task 3: Multimodal emotion cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2039–2050, Mexico City, Mexico, 2024. https://doi.org/10.18653/v1/2024.semeval-1.277

[21] Y. Etesam, Ö. N. Yalçın, C. Zhang, and A. Lim. Contextual emotion recognition using large vision language models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4769–4776, 2024. https://doi.org/10.1109/IROS58592.2024.10802538

[22] Y. Fu, J. Wu, Z. Wang, M. Zhang, L. Shan, Y. Wu, and B. Liu. LaERC-S: Improving LLM-based emotion recognition in conversation with speaker characteristics. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pages 6748–6761, Abu Dhabi, UAE, 2025. https://aclanthology.org/2025.coling-main.451

[23] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. *Official Journal of the European Union*, 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[24] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372:n71, 2021. https://doi.org/10.1136/bmj.n71