Modelling Emotions is an Elusive Pursuit in Affective Computing

Anders Rolighed Larsen
Department of Applied Mathematics and Computer Science
Technical University of Denmark

Sneha Das
Department of Applied Mathematics and Computer Science
Technical University of Denmark
sned@dtu.dk

Nicole Nadine Lønfeldt
Child and Adolescent Mental Health Center, Copenhagen University Hospital, Denmark
nicole.nadine.loenfeldt@regionh.dk

Paula Petcu
Interhuman AI
paula@interhuman.ai

Line Clemmensen
Department of Applied Mathematics and Computer Science
Technical University of Denmark
lkhc@dtu.dk

Abstract

Affective computing - combining sensor technology, machine learning, and psychology - have been studied for over three decades and is employed in AI-powered technologies to enhance emotional awareness in AI systems, and detect symptoms of mental health disorders such as anxiety and depression. However, the uncertainty in such systems remains high, and the application areas are limited by categorical definitions of emotions and emotional concepts. This paper argues that categorical emotion labels obscure emotional nuance in affective computing, and therefore continuous dimensional definitions are needed to advance the field, increase application usefulness, and lower uncertainties.

1 Introduction

Affective computing systems include uni-modal research fields like facial emotion recognition (FER) [Verma, 2023], speech emotion recognition (SER) [Ma et al., 2023], as well as multi-modal sentiment analysis (MSA) [Hu et al., 2024, Soleymani et al., 2017], and human-computer interaction [Soleymani et al., 2017, Preeti, 2012]. The emotion recognition task entails mapping input signals such as speech, facial expressions, and text to affective states. This mapping is most commonly performed using categorical emotion annotations (CEA), such as *happy*, *sad*, or *angry*, derived from psychological taxonomies that include Ekman's basic emotions [Ekman, 1992] and Plutchik's psychoevolutionary model [PLUTCHIK, 1980] [Verma, 2023, Abdullah, 2021, García-Hernández et al., 2024].These

Preprint. Under review.

taxonomies serve as the primary annotation schema in many widely-used datasets, including MELD [Poria et al., 2019], eNTERFACE [Martin et al., 2005], SAVEE [Haq and Jackson, 2009], TESS [Dupuis and Pichora-Fuller, 2010], EmoDB [Burkhardt et al., 2005], and FER2013 [Goodfellow et al., 2013], where emotion labels are typically drawn from a fixed set of basic categories. While this approach offers simplicity of annotation and computational clarity, it imposes rigid boundaries on phenomena that are often ambiguous, overlapping, temporally dynamic and rarely experienced in isolation [Mower et al., 2009, Schuller, 2018, Gendron et al., 2018, Busso et al., 2008, Bradley and Lang, 1999, Park et al., 2020, Soleymani et al., 2017].

The IEMOCAP dataset [Busso et al., 2008], widely used in affective computing research, exemplifies this issue. Here, each utterance is labeled with a categorical emotion based on majority voting among three annotators, while also independently annotated with continuous valence, arousal, and dominance (VAD) scores. However, the dual annotation structure reveals a critical disjunction: VAD values often deviate significantly from the emotional categories with which they are associated, and annotators frequently disagree (see Section 3 of this paper). This inconsistency, along with a lack of a general consensus definition of emotions [Cabanac, 2002], challenges the notion of a singular ground truth based on single categorical labels. Instead, some suggest that emotions should be modeled as distributions over possible states rather than as fixed points in a discrete label space [Schipor et al., 2011].

Although datasets like IEMOCAP provide categorical labels and continuous VAD scores, these are typically used independently. The categorical label is derived via a majority vote, thereby collapsing the annotator disagreement into a singular outcome. Recent modeling strategies aim to preserve this ambiguity through soft-labeling techniques, fuzzy classifiers, or emotional profiling frameworks that treat emotional states as distributions rather than fixed categories [Palotti et al., 2023, Marioooryad et al., 2014, Davani et al., 2022]. These methods reflect a growing recognition that emotions are computationally and psychologically context dependent, temporally fluid, and rarely experienced as singular static states [Russell, 1980, Gendron et al., 2018].

These challenges have prompted a growing call for alternative representations that better reflect the fluid and context-dependent nature of human emotion. Rather than treating ambiguity and disagreement as annotation noise, emerging work suggests that they are essential signals that emotion-aware systems should accommodate.

This paper takes the position that categorical emotion labels, while computationally convenient, obscure emotional nuance, ambiguity, and subjectivity, thus limiting the fidelity, interpretability, and ethical deployment of affective computing systems.

2 Theoretical Foundations and Conceptual Incongruence in Emotion Modeling

Emotion modeling lies at the intersection of psychological theory and computational pragmatism. Although this confluence offers rich interpretative power, it also reveals fundamental tensions in how affect is conceptualized, measured, and operationalized, especially within the context of MSA. By critically examining the structure of emotional theories in psychology and the representation strategies in MSA, this section aims to expose an underlying discord: Current systems largely neglect the granularity and ambiguity inherent in real emotional experiences.

Contrasting Emotional Frameworks Psychological models of emotion have historically evolved along two primary axes: discrete and dimensional. Discrete models, exemplified by Plutchik's Wheel of Emotions [PLUTCHIK, 1980], Parrott's hierarchical taxonomy [Parrott, 2001, Schipor et al., 2011], and Ekman's universal emotions theory [Ekman, 1992], categorize emotions into bounded classes. These models offer practical advantages in clarity and universality, but critics have noted their limitations in capturing the complex, often culturally mediated, nuances of emotional expression [Cordaro et al., 2017, Gendron et al., 2018].

In contrast, dimensional models conceptualize emotions as existing on continuous spectra. The circumplex model of affect [Russell, 1980] defines emotions across the valence and arousal axes, often extended with a third dimension, such as dominance or potency [Schlosberg, 1954, Russell and Mehrabian, 1977]. This structure supports more fluid interpretations of affect, especially useful in therapeutic or introspective settings where emotions rarely conform to single-label categories [Busso

et al., 2008, Bradley and Lang, 1999, Das et al., 2022b,a]. Further, research has also shown the language dependency of frameworks and specific factors in the frameworks [Das et al., 2022c, Hjuler et al., 2025a,b].

Representation Constraints in MSA Despite the theoretical richness of the dimensional models, MSA systems have predominantly favored discrete approaches. Labels derived from Ekman's categories are easier to annotate and align with multimodal data sets [Poria et al., 2018], making them attractive for large-scale computational tasks. A gradual shift toward dimensional representations, typically valence, arousal, and dominance, indicates a growing awareness of the need for nuance, particularly in modeling emotional intensity or co-occurring states.

However, the scope of representation remains restricted. MSA tends to focus on prototypical emotions, happiness, sadness, anger, fear, and disgust, which produce high agreement between annotators and clearer detection signals [Park et al., 2020]. Emotions like guilt, nostalgia, or ambivalence, which reflect blended or ambiguous affective states, are frequently omitted due to annotation challenges and data sparsity. Practitioners [Neville, 2025] occasionally introduce custom axes such as 'Energy' or 'Tension' to refine the granularity, although such choices often stem from subjective preference rather than theoretical consistency.

This pragmatic divergence leads to a critical distinction: psychological inquiry prioritizes subjective experience and emotional complexity, while MSA is designed around observable signals and model efficiency. Consequently, the emotional spectrum in MSA is operationalized in a narrower, often oversimplified format that hampers the recognition of subtle or complex emotions [Soleymani et al., 2017].

Conceptual Vagueness and Terminological Diffusion The representational gap between psychological theory and computational practice is further compounded by the lack of terminological clarity in affective computing. Affective computing lacks a shared vocabulary for describing emotional ambiguity. In the literature, a wide range of loosely defined terms, as shown in Figure 6, are used to describe similar phenomena of interpretive uncertainty without clear agreement. Some of these terms originate from psychological theory, others from system design, but their inconsistent use has prevented convergence toward a common taxonomy.

Of particular relevance is the notion of *emotional ambivalence*, defined as the co-occurrence of opposing emotional states toward the same stimulus [Larsen et al., 2001, NeuroLaunch, 2024, Aviezer and Hassin, 2017]. It also mirrors the annotation behavior seen in evaluators who assign multiple labels to a single utterance, revealing uncertainty or a nuanced perception rather than indecision. Even the person experiencing ambivalence may have trouble describing the experience.

Adjacent to this is the concept of *prototypicality*, where consensus in annotation signals a normative emotional expression. Conversely, *non-prototypicality* and *emotional incongruity* challenge standard classification models by surfacing the ambiguity in cross-modal cues, for example, when the vocal tone contradicts facial expression [Zhang et al., 2016, Kim and Provost, 2015, Schipor et al., 2011, Deng et al., 2021]. Annotators may differ in their ability to express own or read others' emotions, in which case striving for a consensus may be seen as averaging out noise. Or it could be seen as a more continuous and natural variation in perception, in which case using distributions to model emotions seems like the better option. Finally, ambiguity across modalities may reflect more complex signals, like sarcasm, suppressed emotions, or situations where the person in question has multiple emotions or transitions between emotions.

In light of this terminological diffusion, we adopt the term *emotional ambiguity* to unify the various phenomena described. Unlike terms that focus solely on polarity or disagreement, emotional ambiguity captures both the presence of overlapping or conflicting emotional cues and the interpretive uncertainty they provoke. It emphasizes the inherently fluid, multi-layered nature of affective perception—particularly when signals across modalities or annotators diverge.

Ambiguity in Emotion Datasets Despite increasing interest in emotional nuance, the representation of ambiguity in publicly available datasets remains inconsistent. Standard corpora such as IEMOCAP [Busso et al., 2008] offer high-quality multimodal data, but largely emphasize dominant emotion labels, often masking the presence of mixed or uncertain affective cues. Although some follow-up work notes the existence of mixed emotions utterances [Mower et al., 2009], the annotation protocol

itself discourages capturing the complexity found in spontaneous expression. Tran et al. [Tran et al., 2022b,a]observe that the expressions acted in IEMOCAP tend to overrepresent prototypical emotions, thus underrepresenting the interpretive ambiguity characteristic of naturalistic interaction.

Data sets such as CMU-MOSEI [Bagher Zadeh et al., 2018] and CMU-MOSEAS [Bagher Zadeh et al., 2020] offer finer granularity by allowing per-emotion intensity ratings. However, even in these cases, the inherent ambiguity in the data points is often overlooked or relegated to future work considerations [Tran et al., 2022b,a, Manju Priya Arthanarisamy Ramaswamy, 2024, Pan et al., 2023, Aguilera et al., 2023, Siddiqui et al., 2022, Rai et al., 2025, Shou et al., 2023]. This reveals a broader trend: while datasets may technically permit multi-label or distributional annotation, prevailing usage patterns default to hard-label classification that suppresses emotional variability.

Annotation Practices and Distributional Alternatives These limitations are not only a matter of dataset design but also of annotation strategy. Majority voting schemes remain standard, despite evidence that such aggregation often conceals meaningful inter-annotator disagreement [Mostafazadeh Davani et al., 2022, Fleisig et al., 2023]. Such disagreements are not merely noise; they reflect socio-cultural diversity, perceptual idiosyncrasies, and inherent emotional ambiguity [Okur et al., 2018, Scherer and Wallbott, 1994, Plisiecki et al., 2024, Chou and Lee, 2019]. Systems that treat these disagreements as error signals risk misrepresenting emotion as an objective truth rather than a subjective construct.

Recent work offers promising alternatives. Soft-label distributions, Likert scales [Palotti et al., 2023, Psychology, 2025], and fuzzy emotion models [Schipor et al., 2011] allow expressions to exist in a multidimensional affective space. Emotional profiling, proposed by Mower et al. [Mower et al., 2009], represents emotional states as distributions between categories. Complementing this, emotional interpolation techniques weigh utterances in relation to the broader context of a dialogue, allowing for ambiguity without forcing artificial resolution.

Crowdsourcing methodologies have advanced similarly, particularly in how they accommodate emotional ambiguity. Mariooryad et al. [Mariooryad et al., 2014] proposed a three-stage Likert-based AMT pipeline designed to generate more nuanced emotional judgments by capturing individual perspectives of annotators rather than collapsing them into consensus. Building on this notion of preserving subjectivity, recent multitask learning approaches have introduced models that explicitly account for annotator-specific bias by learning separate label distributions for each rater alongside aggregate consensus labels [Mostafazadeh Davani et al., 2022, Chou and Lee, 2019, Snow et al., 2008]. These architectures not only maintain the diversity of emotional interpretation, but also demonstrate improved recall and greater interpretability, particularly in the presence of ambiguity.

Debate is emerging around the legitimacy of human annotation altogether. Plisiecki et al.[Plisiecki et al., 2024] warn that political and ideological biases embedded in annotators can corrupt model output in downstream tasks, advocating for lexicon-driven alternatives. Self-annotation, as used in IEMOCAP and later revisited by Zhang et al.[Zhang et al., 2016] and Saganowski et al. [Saganowski et al., 2022], offers one way to address this, capturing the introspective dimensions of affective experience - although this is not without its own set of limitations.

Pragmatic Merits of Categorical Annotation Despite their limitations, categorical emotion annotations remain widely used in affective computing due to their operational efficiency and empirical robustness in constrained tasks. Discrete categories such as *happy*, *angry*, and *sad* offer a shared vocabulary between annotators and systems, simplifying the annotation and model evaluation. As [Schuller, 2018] highlights in a retrospective review of speech emotion recognition, the field has made significant strides using categorical labels as benchmarks, allowing standardized comparisons and cumulative progress in studies. In emotion classification tasks, where rapid system response or interpretability is essential, such as driver monitoring [Xiao et al., 2022, Espino-Salinas et al., 2024] or call center analytics [Deschamps-Berger et al., 2021], categorical predictions serve actionable outputs that are easy to map to rules or interventions. Moreover, categorical models often outperform dimensional ones in low-data regimes or when emotion intensity is low, as shown in [Zhang et al., 2016]'s joint modeling of perceived and self-reported emotion. Categorical annotation schemes also facilitate multimodal data fusion: since modalities may align more naturally to prototypical categories than abstract affective spaces, joint modeling becomes more tractable. Even when emotional states are complex, annotators can provide multiple labels or rank primary versus secondary emotions, which is a practical compromise seen in datasets such as IEMOCAP [Busso et al., 2008].

2.1 Position and Outlook

The empirical and theoretical evidence presented throughout this paper points to a central problem: affective computing systems that rely on categorical emotion labels are structurally limited in their ability to capture the ambiguity and subjectivity inherent in emotional expression. This limitation is a fundamental mismatch between the complexity of affective experience and the representational tools we use to capture it. Rather than treating ambiguity as noise or error, we argue that it should be embraced as a modeling target. In the sections that follow, we explore the consequences of this misalignment for real-world systems, including risks to interpretability, ethical use, and user trust.

3 Empirical Breakdown of Categorical Emotion Assumptions

To ground the theoretical critiques outlined above, we present empirical analyses of the IEMOCAP dataset [Busso et al., 2008], which provides both CEA and continuous VAD scores. These analyses expose multiple forms of annotation misalignment that call into question the reliability and expressiveness of discrete labels.

Dimensional Inconsistencies within Categorical Labels. Although each IEMOCAP utterance is assigned a single categorical label (e.g., *sad*, *angry*), their associated VAD values show considerable dispersion even within the same label class. When these empirical VAD annotations are compared with theoretically expected VAD coordinates (as derived from Russell's three-factor model), we observe significant divergence patterns. For example, utterances labeled *happy* span a wide range of arousal and dominance, violating the premise of internal VAD coherence within a category. Figure 1 visualizes this divergence, plotting empirical and theoretical VAD positions, and highlighting how averaged annotator VAD scores deviate from canonical emotion coordinates. This discrepancy suggests that categorical labels often conceal heterogeneous emotional interpretations.

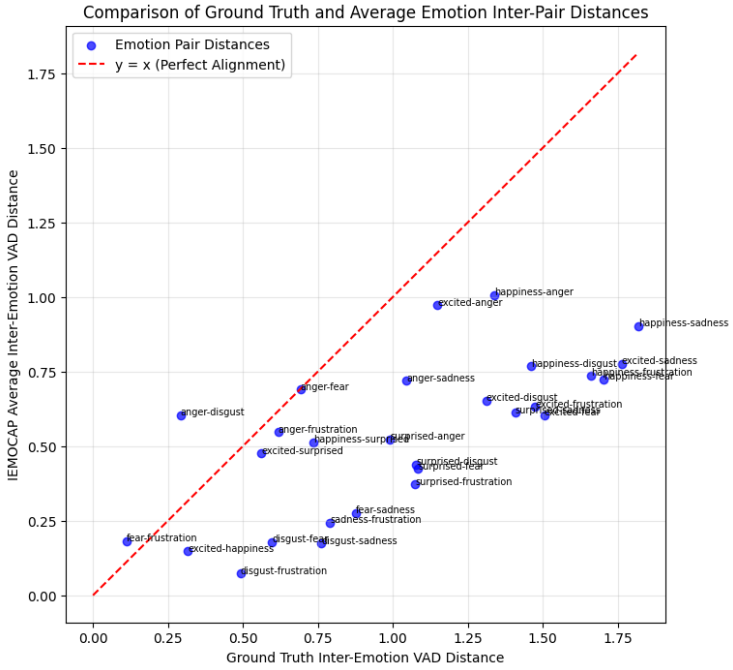

Figure 1: Pairwise emotion distances in VAD space: theoretical ground truth vs. IEMOCAP annotator averages. Most points fall below the diagonal, indicating that annotators perceived emotion categories as more similar than dimensional theory predicts.

Modality-Specific Disagreement. A second axis of ambiguity emerges when comparing emotion predictions across modalities. We report a systematic disagreement between text, audio, and visual models tested on the same utterances.Confusion matrices show, for example, that the text modality

frequently defaults to *neutral*, while the audio model tends to prefer *sadness* or *anger*, even for the same input. The general agreement between the three modalities was only 4.18%, with 54% of the utterances receiving completely distinct predictions in all models. These findings are summarized in Figure 2, emphasizing that affective perception is highly modality-dependent and that categorical alignment across channels is the exception, not the norm. Since the pre-trained audio model does not include "disgust" among its recognized categories, this class is excluded from its predictions and therefore absent in the corresponding comparison matrices. Agreement metrics are calculated based solely on samples for which both modalities yield predictions within their shared label space.

Table 1: Overview of pre-trained emotion recognition models used for each modality.

Modality	**Architecture**	**Model**
Text	Transformer (DistilRoBERTa)	`emotion-english-distilroberta-base` Hartmann [2022]
Audio	Transformer (Wav2Vec2)	`w2v-speech-emotion-recognition` Khoa [2024]
Facial	CNN + LSTM (ResNet50 + LSTM)	`EMO-AffectNetModel` Ryumina et al. [2022]

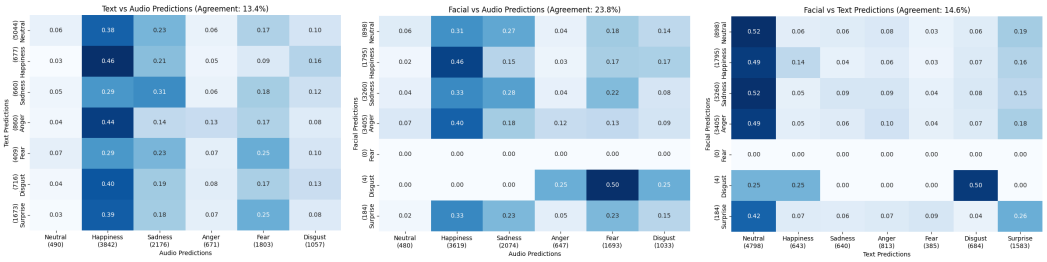

Figure 2: Cross-modal prediction alignment matrices for each modality pair, with overall agreement percentages indicated in parentheses. Left: Text vs. Audio predictions. Center: Facial vs. Audio predictions. Right: Facial vs. Text predictions. Each matrix is normalized by the number of predictions per class, by the y-axis predictor, to account for imbalanced emotion distributions across modalities. A difference in class support can be observed between modality comparisons, due to faulty data-entries to the facial modality.

Temporal Flattening of Emotional Transitions. Emotional states are not static; they often evolve substantially within the span of a single utterance. Frame-level analyses of facial expressions reveal that predicted emotional states frequently shift mid-utterance. Figure 3 illustrates this dynamic: emotion probability curves fluctuate over time, while shannon entropy on model probabilities (black line) captures moment-to-moment uncertainty. Red markers indicate identified transition points between predicted dominant emotions, which tend to align with localized spikes in entropy. This suggests that transitions correspond with heightened affective ambiguity.

This local pattern holds at scale. As shown in Appendix Figure 8, entropy across all transitional frames ($\pm$ 5 frames at transition point) is significantly higher than during baseline (stable sequences) periods, with a strong statistical effect ($p << 0$). There was discarded 5 frames between a transition window and a stable sequence in an attempt to alleviate auto-correlation. This global analysis confirms that emotion transitions are not just points of label change, but regions of elevated model uncertainty, stemming from the constraints placed by the categorical annotation domain. Moreover, entropy varies across specific emotion shifts: transitions involving surprise, neutrality, or shifts between positively and negatively valenced states exhibit the highest entropy (Appendix Figure 9). These findings underscore that not all transitions are equally ambiguous, and that ambiguity itself is a function of emotional context and trajectory.

Taken together, these results demonstrate that flattening entire utterances into single categorical labels discards rich affective dynamics and obscures the interpretive ambiguity inherent in transition phases. Emotion-aware systems would benefit from modeling such dynamics explicitly, whether via

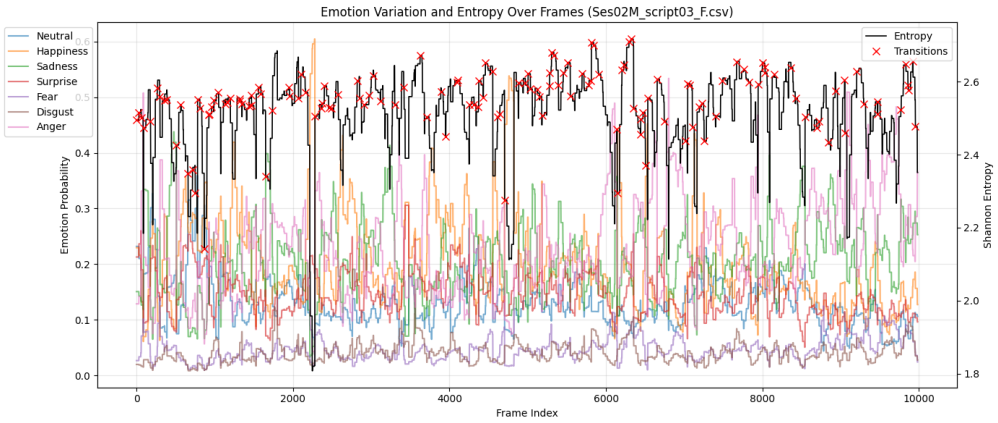

Figure 3: Framewise emotion probabilities (colored lines) and corresponding entropy (black line) for a representative utterance. Red x's mark emotion transitions. Note the alignment of entropy spikes with transitions, indicating periods of increased affective ambiguity.

frame-level representations, temporal smoothing, or probabilistic tracking of multiple concurrent emotional states.

4 Evaluation Under Ambiguity

Emotion recognition systems are typically benchmarked against discrete ground truth labels, assuming these annotations are accurate and representative of a singular emotional state. However, the findings of the IEMOCAP dataset challenge this assumption. Full annotator agreement on categorical emotion annotations (CEA) was observed in only approximately 20% of utterances, and nearly 25% of utterances lacked a clear majority label overall. These rates indicate a substantial degree of inherent ambiguity in the data itself.

4.1 Impact of Ambiguity on Model Performance

To quantify how ambiguity influences model evaluation, we implemented a filtering strategy based on categorical agreement levels. This method creates partitions of the data set based on the proportion of annotators who agreed on the majority label for each utterance. For example, a threshold of 0.75 retains only those samples in which at least 75% of the annotators selected the same emotion, thereby reducing the influence of ambiguous or contested labels.Model performance was evaluated using the weighted F1 score, which averages per-class F1 scores weighted by label frequency to account for class imbalance. As ambiguity decreased (i.e., with more stringent thresholds), model performance consistently improved for audio and facial modalities. For example, the weighted F1 of the audio model increased from 0.332 at baseline to 0.496 at higher level of prototypicality, indicating a substantial sensitivity to the clarity of the annotation. The facial model also showed marked improvements. In contrast, the performance of the text modality remained relatively stable, highlighting its greater resilience to label noise, but also possibly its lower sensitivity to affective nuance.

These effects are illustrated in Figure 4, which shows weighted F1 scores at increasing levels of cut-off values used for CEA agreement. The trend suggests that high-agreement data provide more consistent signals and a cleaner evaluation, with the implication that data, when ambiguous, is also filtered out in the process.

4.2 VAD Dispersion as an Alternative Metric

An alternative ambiguity filter was tested using the dispersion in the euclidean distance space of the valence, arousal, and dominance scores, in the [0,5] domain as is the range used in IEMOCAP. Contrary to expectations, reducing VAD disagreement did not consistently improve model performance. For the audio modality, the performance decreased slightly with stricter VAD agreement

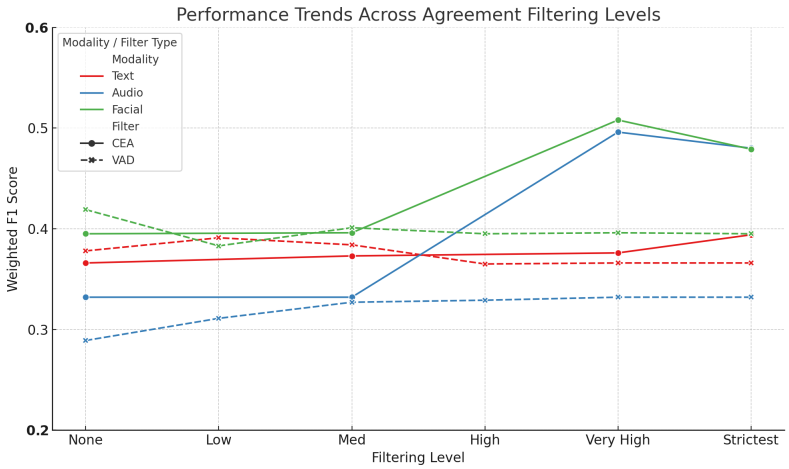

Figure 4: Weighted F1 scores for text, audio, and facial emotion recognition models across increasing levels of agreement-based filtering. Filtering is applied based on either categorical emotion annotation (CEA, solid lines) or VAD score coherence (VAD, dashed lines). Values and partioning parameters can be found in Appendix Figures 3 and 4.

filters, and the performance of the facial modality remained flat or inconsistent. This finding suggests that continuous VAD annotations from the IEMOCAP dataset, despite their finer granularity, do not necessarily correlate with categorical ground-truth reliability or clarity. In addition, it reinforces the misalignment between the two annotation paradigms.

4.3 Agreement Across Modalities

Model disagreement further complicates the evaluation landscape. Full agreement among the three unimodal models—text, audio, and facial—was achieved in just 4.18% of the utterances. This low concordance persisted even when the predictions were compared pair-wise, with a particularly weak alignment between the text and audio modalities. The divergence illustrates that each modality captures distinct affective signals and that rigid alignment to a singular label may misrepresent this variation.

Table 2 summarizes the frequency of complete agreement between the models and highlights the emotions that are the most agreed on.

Table 2: Emotion-wise agreement count across modalities (text, audio, facial).

Emotion	Agreed Count	Share of Total
Happiness	125	
Sadness	104	
Anger	67	
Neutral	37	
Total (All Emotions)	**333**	**4.18%**

This systemic misalignment, between models, annotations, and modalities, undermines the validity of single-label evaluations and calls for a rethinking of how emotion recognition systems are trained and assessed. Categorical evaluation overlooks perceptual divergence, while VAD ratings, though more granular, fail to provide a reliable proxy for consensus. Together, these findings suggest that ambiguity is not only noise to be removed, but also contains a signal to be modeled.

5 Ambiguity in Practice: Ethical and Experimental Implications

Emotion recognition technologies are increasingly being deployed in settings where outputs are not merely academic predictions, but guide real-time feedback, decision-making, and interpersonal un-

derstanding. Applications such as coaching, education, and mental health support place these systems in ethically sensitive roles, where labeling affect carries practical and psychological consequences. However, most current systems, often trained on majority-voted categorical labels, are structurally predisposed to flatten complexity.

A key concern is that such systems convey a false sense of emotional precision, masking underlying uncertainty or disagreement. Systems typically output singular emotion labels such as *happy* or *sad* without communicating underlying ambiguity, disagreement, or temporal fluctuation. Figure 12 illustrates one such case: all three unimodal models converged on the same label (*happiness*), despite annotators labeling the utterance as *anger* or *frustration*. This divergence, often tied to sarcasm, ambivalence, or multimodal mismatch, reveals how overconfident outputs can misrepresent the user's state.

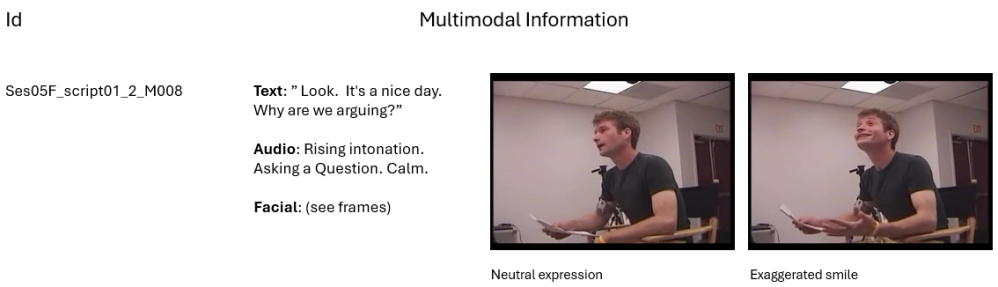

Figure 5: Case of modality agreement in system predictions diverging from ground truth. All models predicted *happy*, while annotators labeled the utterance as *angry* or *frustrated*. Examples directly taken from IEMOCAP data set [Busso et al., 2008] and used under their licensing agreement.

Such misclassifications are not neutral. In emotionally sensitive domains, users may feel unseen, invalidated, or misunderstood by systems that present misleadingly precise feedback. Over time, these mismatches can erode user trust and reduce the perceived empathy or usefulness of the technology - especially in longitudinal or therapeutic settings. When ambiguity is suppressed through majority voting or softmax outputs, interpretive nuance is lost.

Systems deployed in emotionally responsive settings must not only detect affect but track its evolution over time. Failing to register how emotional states build, shift, or resolve risks producing responses that feel uncalibrated or disconnected from lived experience. This is especially problematic in dialogue-based applications, where a flattened emotional summary can obscure meaningful turning points. A system that fails to register rising frustration, for instance, may miss the opportunity for timely intervention. As earlier analyses showed (Figure 3), these dynamics are perceptible at the frame level, yet are often lost in current labeling approaches.

Ambiguity in emotion data is not a defect to be corrected but a core feature of human expression. Systems that erase this complexity in pursuit of clarity risk technical inaccuracy and experiential harm. Ethically, developers of emotion-aware AI must treat labeling as an interpretive act — not merely a prediction task. Practically, users deserve systems that reflect the nuance and fluctuation inherent in how emotions are expressed and perceived.

6 Position and future directions

Emotion recognition systems today continue to rely on categorical annotations that reduce complex, context-dependent affective states to singular labels. As demonstrated throughout this paper, this simplification overlooks valuable information, such as variability between annotators, divergence between modalities, and transitions within utterances, that could otherwise inform more nuanced and robust models.

We do not claim that dimensional or fuzzy approaches are fully mature alternatives. However, we argue that the field must begin treating ambiguity as an informative signal rather than noise to be

discarded. Moving forward, promising directions include the adoption of soft-label distributions, emotional profiling, and annotator-specific modeling, all of which preserve subjectivity rather than collapsing it. Similarly, integrating contextual features such as dialogue structure, temporal dynamics, and multiple perceptual modalities will allow systems to infer emotional states more fluidly and adaptively—capturing the layered cues that contribute to ambiguity in human emotional perception.

The shift ahead is not merely technical, but rather it is conceptual. Affective computing must revise its assumptions about emotion itself: from discrete to continuous, from static to situated, from certain to interpretively plural. Embracing this complexity, and exploring new concepts, is the necessary next step if emotion-aware systems are to reflect the nuance and variability of human affective experience.

References

Sharmeen M.Saleem Abdullah Abdullah. *Journal of Applied Science and Technology Trends*, 2 (01):73–79, May 2021. doi: 10.38094/jastt20291. URL https://jastt.org/index.php/jasttpath/article/view/91.

Ana Aguilera, Diego Mellado, and Felipe Rojas. An assessment of in-the-wild datasets for multimodal emotion recognition. *Sensors*, 23(11), 2023. ISSN 1424-8220. doi: 10.3390/s23115184. URL https://www.mdpi.com/1424-8220/23/11/5184.

Hillel Aviezer and Ran Hassin. 333inherently ambiguous: An argument for contextualized emotion perception. In *The Science of Facial Expression*. Oxford University Press, 04 2017. ISBN 9780190613501. doi: 10.1093/acprof:oso/9780190613501.003.0018. URL https://doi.org/10.1093/acprof:oso/9780190613501.003.0018.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208/.

AmirAli Bagher Zadeh, Yansheng Cao, Simon Hessner, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. CMU-MOSEAS: A multimodal language dataset for Spanish, Portuguese, German and French. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1801–1812, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.141. URL https://aclanthology.org/2020.emnlp-main.141/.

Margaret M. Bradley and Peter J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. 1999. URL https://api.semanticscholar.org/CorpusID:145474983.

Felix Burkhardt, Angelika Paeschke, Michael Rolfes, Werner F. Sendlmeier, and Benjamin Weiss. A database of german emotional speech. *Interspeech*, 2005.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, December 2008.

Michel Cabanac. What is emotion? *Behavioural Processes*, 60(2):69–83, 2002. ISSN 0376-6357. doi: https://doi.org/10.1016/S0376-6357(02)00078-5. URL https://www.sciencedirect.com/science/article/pii/S0376635702000785.

Huang-Cheng Chou and Chi-Chun Lee. Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5886–5890, 2019. doi: 10.1109/ICASSP.2019.8682170.

Daniel Cordaro, Rui Sun, Dacher Keltner, Shanmukh Kamble, Niranjan Huddar, and Galen McNeil. Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion*, 18, 06 2017. doi: 10.1037/emo0000302.

Sneha Das, Nicole Nadine Lonfeldt, Nicklas Leander Lund, Anne Katrine Pagsberg, and Line Katrine Harder Clemmensen. Zero-shot cross-lingual speech emotion recognition: A study of loss functions and feature importance. In *2nd Symposium on Security and Privacy in Speech Communication*, 2022a.

Sneha Das, Nicklas Leander Lund, Nicole Nadine Lønfeldt, Anne Katrine Pagsberg, and Line Katrine Harder Clemmensen. Continuous metric learning for transferable speech emotion recognition and embedding across low-resource languages. In *Northern Lights Deep Learning Workshop 2022*, 2022b.

Sneha Das, Nicole Nadine Lønfeldt, Anne Katrine Pagsberg, and Line H. Clemmensen. Towards transferable speech emotion representation: On loss functions for cross-lingual latent representations. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6452–6456, 2022c. doi: 10.1109/ICASSP43922.2022.9746450.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 01 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00449. URL https://doi.org/10.1162/tacl_a_00449.

Didan Deng, Liang Wu, and Bertram E. Shi. Iterative distillation for better uncertainty estimates in multitask emotion recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3557–3566, 10 2021.

Théo Deschamps-Berger, Lori Lamel, and Laurence Devillers. End-to-end speech emotion recognition: Challenges of real-life emergency call centers data recordings. *CoRR*, abs/2110.14957, 2021. URL https://arxiv.org/abs/2110.14957.

Kate Dupuis and M. Kathleen Pichora-Fuller. Toronto emotional speech set (tess), 2010. Available: https://tspace.library.utoronto.ca/handle/1807/24487.

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

Carlos H. Espino-Salinas, Huizilopoztli Luna-García, José M. Celaya-Padilla, Cristian Barría-Huidobro, Nadia Karina Gamboa Rosales, David Rondon, and Klinge Orlando Villalba-Condori. Multimodal driver emotion recognition using motor activity and facial expressions. *Frontiers in Artificial Intelligence*, 7, November 2024. ISSN 2624-8212. doi: 10.3389/frai.2024.1467051. URL http://dx.doi.org/10.3389/frai.2024.1467051.

Eve Fleisig, Rediet Abebe, and Dan Klein. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.415. URL https://aclanthology.org/2023.emnlp-main.415/.

Rosa García-Hernández, Huizilopoztli Luna-Garcia, Jose Celaya Padilla, Alejandra García, Luis Reveles, Luis Flores-Chaires, Juan Ruben Delgado Contreras, David Rondon, and Klinge Villalba. A systematic literature review of modalities, trends, and limitations in emotion recognition, affective computing, and sentiment analysis. *Applied Sciences*, 14:7165, 08 2024. doi: 10.3390/app14167165.

Maria Gendron, Carlos Crivelli, and Lisa Barrett. Universality reconsidered: Diversity in making meaning of facial expressions. *Current Directions in Psychological Science*, 27:096372141774679, 07 2018. doi: 10.1177/0963721417746794.

Ian J. Goodfellow, Dumitru Erhan, Pierre Carrier, Aaron C. Courville, Mehdi Mirza, Ben Hamner, William Cukierski, Yuan Tang, David Thaler, Dong-Hyun Lee, Ying Zhou, Chuan Ramaiah, Yan Feng, Rui Li, Xiaoguang Wang, Alex Sherstinsky, and Tony Tang. Challenges in representation

learning: A report on three machine learning contests. *Neural Information Processing*, 2013. FER-2013 described in the ICML 2013 Challenges in Representation Learning.

S. Haq and P. J. B. Jackson. Audio-visual emotion recognition using adaboost. In *Proceedings of the International Conference on Auditory-Visual Speech Processing*, 2009.

Jochen Hartmann. Emotion english distilroberta-base. `https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/`, 2022.

Maja J Hjuler, Line H Clemmensen, and Sneha Das. Exploring local interpretable model-agnostic explanations for speech emotion recognition with distribution-shift. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025a.

Maja J Hjuler, Harald V Skat-Rørdam, Line H Clemmensen, and Sneha Das. Emotale: An enacted speech-emotion dataset in danish. *arXiv preprint arXiv:2508.14548*, 2025b.

Guimin Hu, Yi Xin, Weimin Lyu, Haojian Huang, Chang Sun, Zhihong Zhu, Lin Gui, Ruichu Cai, Erik Cambria, and Hasti Seifi. Recent trends of multimodal affective computing: A survey from nlp perspective, 2024. URL `https://arxiv.org/abs/2409.07388`.

Khoa. Wav2vec2 speech emotion recognition for english. `https://huggingface.co/Khoa/w2v-speech-emotion-recognition`, 2024.

Yelin Kim and Emily Mower Provost. Leveraging inter-rater agreement for audio-visual emotion recognition. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 553–559, 2015. doi: 10.1109/ACII.2015.7344624.

Jeff T. Larsen, A. Peter McGraw, and John T. Cacioppo. Can people feel happy and sad at the same time? *Journal of personality and social psychology*, 81 4:684–96, 2001. URL `https://api.semanticscholar.org/CorpusID:18806410`.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation, 2023. URL `https://arxiv.org/abs/2312.15185`.

Suja Palaniswamy Manju Priya Arthanarisamy Ramaswamy. Multimodal emotion recognition: A comprehensive review, trends, and challenges. 08 2024.

Soroosh Mariooryad, R. Lotfian, and Carlos Busso. Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 238–242, 01 2014.

Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas. enterface'05 audio-visual emotion database. In *Data Challenge Workshop on Emotion Recognition*, 2005.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022. doi: 10.1162/tacl_a_00449. URL `https://aclanthology.org/2022.tacl-1.6/`.

Emily Mower, Angeliki Metallinou, Chi-Chun Lee, Abe Kazemzadeh, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Interpreting ambiguous emotional expressions. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–8, 2009. doi: 10.1109/ACII.2009.5349500.

NeuroLaunch. Emotional ambivalence. `https://neurolaunch.com/emotional-ambivalence/`, 2024. Accessed: 27-01-2025.

Peter Neville. Personal communication. Interview, 2025. Interview held february 2025.

Eda Okur, Sinem Aslan, Nese Alyuz, Asli Arslan, and Ryan Baker. The importance of socio-cultural differences for annotating and detecting the affective states of students. 12 2018.

Joao Palotti, Gagan Narula, Lekan Raheem, and Herbert Bay. Analysis of emotion annotation strength improves generalization in speech emotion recognition models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5829–5837, 2023. doi: 10.1109/CVPRW59228.2023.00619.

Bei Pan, Kaoru Hirota, Zhiyang Jia, and Yaping Dai. A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. *Neurocomputing*, 561:126866, 2023. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.126866. URL https://www.sciencedirect.com/science/article/pii/S092523122300989X.

Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data*, 7 (1), September 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-00630-y. URL http://dx.doi.org/10.1038/s41597-020-00630-y.

W. Gerrod Parrott. Emotions in social psychology : essential readings. 2001. URL https://api.semanticscholar.org/CorpusID:141721437.

Hubert Plisiecki, Paweł Lenartowicz, Maria Flakus, and Artur Pokropek. High risk of political bias in black box emotion inference models. 2024. URL https://arxiv.org/abs/2407.13891.

ROBERT PLUTCHIK. Chapter 1 - a general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3–33. Academic Press, 1980. ISBN 978-0-12-558701-3. doi: https://doi.org/10.1016/B978-0-12-558701-3.50007-7. URL https://www.sciencedirect.com/science/article/pii/B9780125587013500077.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Erik Cambria, Alexander Gelbukh, and Amir Hussain. Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intelligent Systems*, 33(6):17–25, 2018. doi: 10.1109/MIS.2018.2882362.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the ACL 2019*, 2019.

Khanna Preeti. Multimodal emotion recognition for enhancing human computer interaction. 2012. URL https://api.semanticscholar.org/CorpusID:260521063.

Explore Psychology. What is the likert scale? definition, examples, and uses. https://www.explorepsychology.com/likert-scale-definition-examples-and-uses/, 2025.

Alpana Rai, Chetan Agrawal, Divya Envey, and Ijeasm Journal. A comprehensive survey of multimodal emotion recognition: Techniques, applications, and future directions. 6:2582–6948, 01 2025.

J.A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6): 1161–1178, 1980. ISSN 0022-3514.

James Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11:273–294, 09 1977. doi: 10.1016/0092-6566(77)90037-X.

Elena Ryumina, Denis Dresvyanskiy, and Alexey Karpov. In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study. *Neurocomputing*, 2022. doi: 10.1016/j.neucom.2022.10.013. URL https://www.sciencedirect.com/science/article/pii/S0925231222012656.

Stanisław Saganowski, Joanna Komoszynska, Maciej Behnke, Bartosz Perz, Dominika Kunc, Bartłomiej Klich, Łukasz D. Kaczmarek, and Przemyslaw Kazienko. Emognition dataset: emotion recognition with self-reports, facial expressions, and physiology using wearables. *Scientific Data*, 9, 2022. URL https://api.semanticscholar.org/CorpusID:248005383.

Klaus R. Scherer and Harald G. Wallbott. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66(2): 310–328, 1994. ISSN 0022-3514. doi: 10.1037/0022-3514.66.2.310. URL http://dx.doi.org/10.1037/0022-3514.66.2.310.

Ovidiu A. Schipor, Stefan G. Pentiuc, and Maria D. Schipor. Using a fuzzy emotion model in computer assisted speech therapy. In Darina Dicheva, Zdravko Markov, and Eliza Stefanova, editors, *Third International Conference on Software, Services and Semantic Technologies S3T 2011*, pages 189–193, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-23163-6.

Harold Schlosberg. Three dimensions of emotion. *Psychological review*, 61 2:81–8, 1954. URL https://api.semanticscholar.org/CorpusID:27914497.

Björn W. Schuller. Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM*, 61(5):90–99, April 2018. ISSN 0001-0782. doi: 10.1145/3129340. URL https://doi.org/10.1145/3129340.

Yuntao Shou, Tao Meng, Wei Ai, Nan Yin, and Keqin Li. A comprehensive survey on multi-modal conversational emotion recognition with deep learning, 2023. URL https://arxiv.org/abs/2312.05735.

Mohammad Faridul Haque Siddiqui, Parashar Dhakal, Xiaoli Yang, and Ahmad Y. Javaid. A survey on databases for multimodal emotion recognition and an introduction to the viri (visible and infrared image) database. *Multimodal Technologies and Interaction*, 6(6), 2022. ISSN 2414-4088. doi: 10.3390/mti6060047. URL https://www.mdpi.com/2414-4088/6/6/47.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In Mirella Lapata and Hwee Tou Ng, editors, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. URL https://aclanthology.org/D08-1027/.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 2017. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2017.08.003. URL https://www.sciencedirect.com/science/article/pii/S0262885617301191. Multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing.

Hélène Tran, Lisa Brelet, Issam Falih, Xavier Goblet, and Engelbert Mephu Nguifo. L'ambiguïté dans la représentation des émotions : état de l'art des bases de données multimodales. 01 2022a.

Hélène Tran, Issam Falih, Xavier Goblet, and Engelbert Mephu Nguifo. Do multimodal emotion recognition models tackle ambiguity? 06 2022b.

Gyanendra Verma. *Multimodal Affective Computing: Affective Information Representation, Modelling, and Analysis*. 03 2023. ISBN 9789815124453. doi: 10.2174/9789815124453123010l.

Huafei Xiao, Wenbo Li, Guanzhong Zeng, Yingzhang Wu, Jiyong Xue, Juncheng Zhang, Chengmou Li, and Gang Guo. On-road driver emotion recognition using facial expression. *Applied Sciences*, 12(2), 2022. ISSN 2076-3417. doi: 10.3390/app12020807. URL https://www.mdpi.com/2076-3417/12/2/807.

Biqiao Zhang, Georg Essl, and Emily Mower Provost. Automatic recognition of self-reported and perceived emotion: does joint modeling help? In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 217–224, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450345569. doi: 10.1145/2993148.2993173. URL https://doi.org/10.1145/2993148.2993173.

7 Appendix

- Affective Dissonance
- Affective Misalignment
- Affective Polarity Conflict
- Ambivalent Sentiment Analysis
- Annotator Idiosyncrasy
- Cognitive Dissonance
- Conflicting Sentiment Cues
- Contextual Emotion Variability
- Emotion Contradiction
- Emotion Multimodality Challenges
- Emotion Recognition Uncertainty
- Emotional Ambiguity
- Emotional Ambivalence
- Emotional Incongruity
- Expressive Modality Conflict
- Implicit vs. Explicit Emotion Mismatch
- Irony Detection
- Modality-Specific Discrepancy
- Multimodal Emotion Discrepancy
- Multimodal Sentiment Divergence
- Prototypicality / Non-prototypicality
- Sarcasm Detection
- Sentiment Incongruence
- Sentimental Misalignment
- Subjective Emotion Divergence
- Subtle Emotional Expression Analysis
- Uncertain Sentiment Prediction

Figure 6: **Terminology Related to Emotional Ambiguity and Sentiment Divergence.**
A non-exhaustive list of terms encountered or explored in the context of affective computing, emotion recognition, and multimodal sentiment analysis. All terms are used in their respective mentions to describe what could be placed under interprative uncertainty / ambiguity. These terms reflect the conceptual diversity and lack of consensus in labeling ambiguous or conflicting emotional expressions across modalities and annotation perspectives.

1. Which emotional categories are most commonly used to describe psychological states in clinical settings?
2. How are emotions such as disgust or surprise, which are less commonly emphasized in clinical psychology, typically interpreted or treated across different disciplines?
3. From a psychological standpoint, what is the typical distribution of expressed emotions in therapeutic or AI coaching contexts? Are individuals more likely to express neutrality, positivity, or negativity?
4. To what extent is emotional ambivalence considered a natural or expected phenomenon in clinical interactions?
5. How should the presence of conflicting emotional cues across different modalities be interpreted diagnostically?
6. In clinical practice, how are different modalities prioritized when assessing a patient's emotional state? Is verbal content weighed more heavily than vocal tone or facial expression?
7. Are there expected patterns in how emotional states evolve? For example, do transitions typically move through phases such as happy $\rightarrow$ neutral $\rightarrow$ sad?
8. If facial expressions pass through a neutral phase while shifting from one emotion to another, should this intermediate state be interpreted as genuine neutrality?
9. What is the general clinical view on emotional neutrality? Is it often a meaningful state or more of a fallback category in ambiguous scenarios?

Figure 7: **Interview Protocol for Clinical Insight.**
List of questions posed to a clinical psychologist, and researchers in the field, to guide interpretation of emotional expression and modality prioritization in affective computing research.

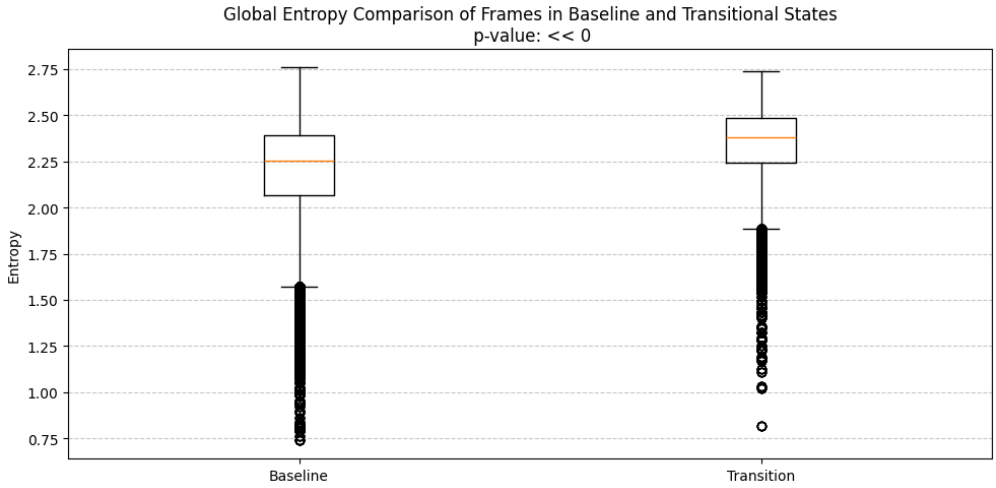

Figure 8: Comparison of entropy distributions across baseline and transitional frames. Transitional segments—defined as regions surrounding changes in dominant emotion—exhibit significantly higher entropy than stable segments, indicating greater ambiguity during emotional shifts. Statistical testing confirms a robust difference ($p << 0$), supporting the claim that transitions are periods of increased affective uncertainty.

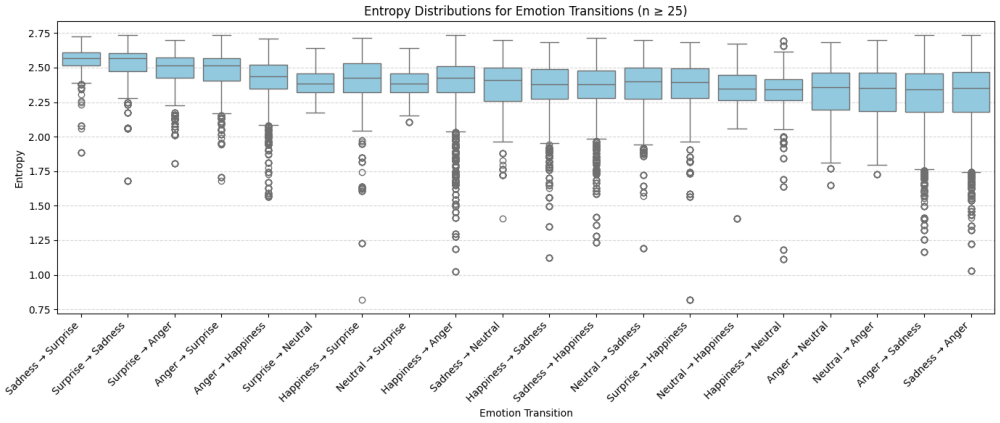

Figure 9: Entropy distributions across specific emotion-to-emotion transitions, based on frame-level predictions. Each boxplot represents a directional transition (e.g., *sadness* $\rightarrow$ *anger*) with at least 25 occurrences. Transitions involving *surprise*, *neutral*, or cross-valence shifts tend to show elevated entropy, suggesting some emotion shifts are inherently more ambiguous than others.

Table 3: Weighted F1 scores for each modality across categorical emotion annotation (CEA) agreement thresholds. Higher thresholds reflect stricter filtering (greater annotator consensus).

Modality	**No Filtering (0)**	**0.6**	**0.7**	**Strictest (1.0)**
Text	0.366	0.373	0.376	0.394
Audio	0.332	0.332	0.496	0.480
Facial	0.395	0.396	0.508	0.479

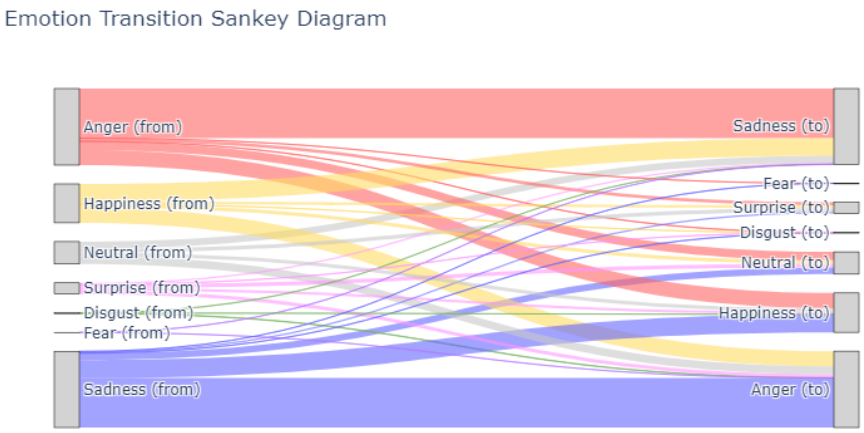

Figure 10: Sankey diagram showing directional frequency of emotion transitions across utterances. The width of each flow represents how often a given emotion transitions into another (e.g., *happiness* → *surprise*, *sadness* → *anger*). The structure and asymmetry of these flows shows how there are no emotions directly acting as transition emotions in the facial expressions picked up by the model.

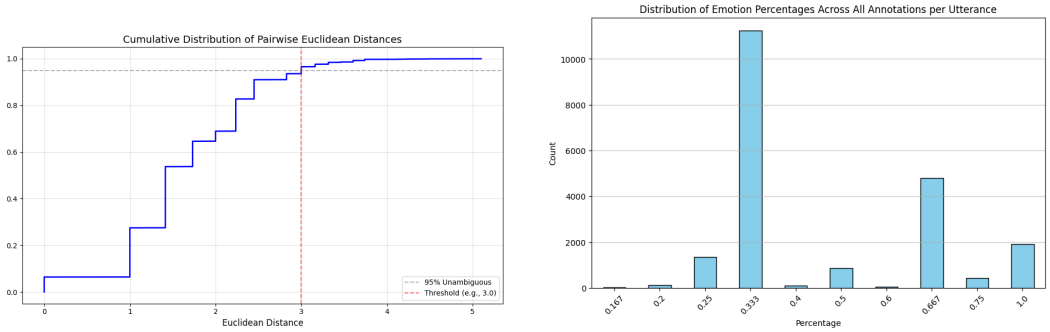

Figure 11: Illustrative distributions underlying threshold selections for ambiguity filtering. **Left**: Cumulative distribution of pairwise Euclidean distances across VAD annotations, used to define VAD-based agreement thresholds. The red dashed line indicates a potential cutoff (e.g., distance = 3.0) that captures 95%. **Right**: Distribution of categorical agreement percentages across utterances. Most utterances fall near high disagreement (0.333), moderate agreement (0.667), or full agreement (1.0), supporting the partitioning schema used in CEA-based filtering.

Table 4: Weighted F1 scores for each modality across valence-arousal-dominance (VAD) agreement filtering levels. Higher steps correspond to stricter filtering based on annotator VAD coherence.

Modality	**No Filtering (5)**	**4**	**3**	**2**	**1**	**Strictest (0)**
Text	0.366	0.366	0.365	0.384	0.391	0.378
Audio	0.332	0.332	0.329	0.327	0.311	0.289
Facial	0.395	0.396	0.395	0.401	0.383	0.419

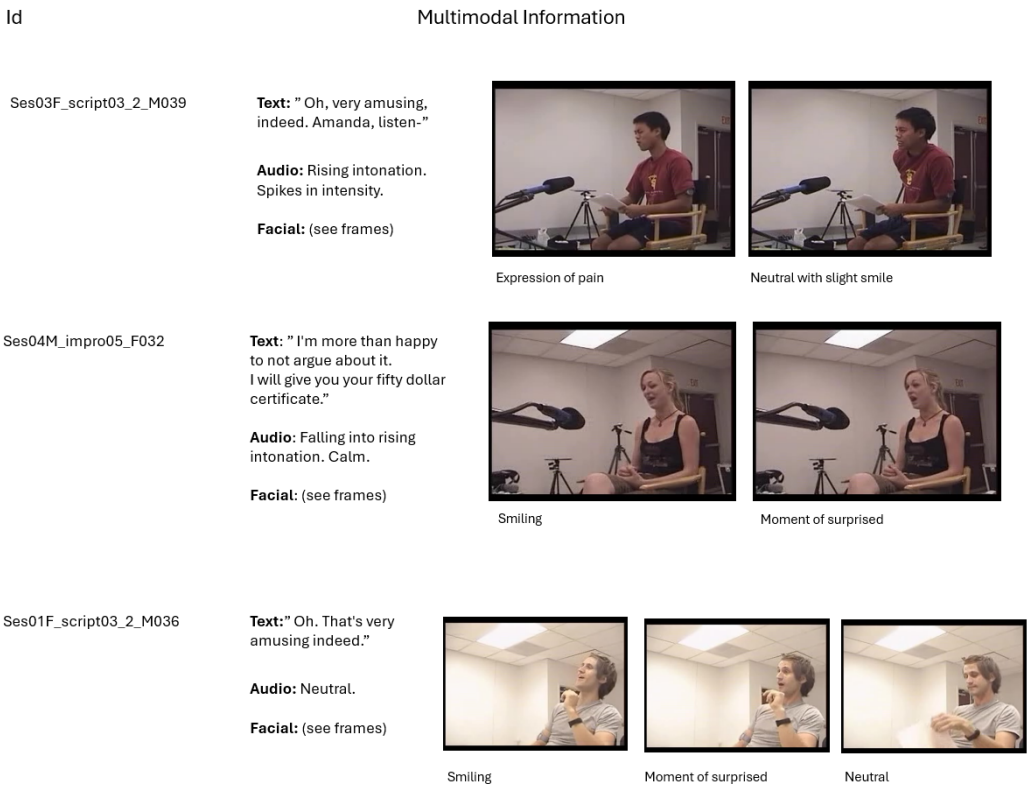

Figure 12: Extra cases of modality agreement in system predictions diverging from ground truth. All models predicted *happy*, while annotators labeled the utterance as *angry* or *frustrated*. Examples directly taken from IEMOCAP data set [Busso et al., 2008] and used under their licensing agreement.