Dynamics of Collective Group Affect: Group-Level Annotations and the Multimodal Modeling of Convergence and Divergence

Navin Raj Prabhu, *Member, IEEE*, Maria Tsfasman, Catharine Oertel, Timo Gerkmann, *Senior Member, IEEE*, and Nale Lehmann-Willenbrock

***Abstract*—Collaborating in a purposive group, whether face-to-face or virtually, involves continuously expressing emotions and interpreting those of other group members. As such, understanding group affect is essential to comprehending how groups interact and succeed in collaborative efforts. In this study, we move beyond individual-level affect and investigate group-level affect—a collective phenomenon that reflects the shared mood or emotions among group members at a particular moment. As the first in the literature, we gather annotations for group-level affective expressions in purposive group interactions using a fine-grained temporal approach (15 s windows) that also captures the inherent dynamics of this collective construct. To this end, we extensively train annotators and develop an annotation procedure specifically tuned to capture the entire scope of the group interaction from one interaction moment to the next. In addition, we model the ebb and flow of group affect by accounting for the underlying convergence (driven by emotional contagion) and divergence (resulting from emotional reactivity) of affective expressions among group members. To capture these interpersonal dynamics, we employ two approaches: (i) extracting synchrony-based handcrafted features from both audio and visual modalities, and (ii) introducing a novel, data-driven graph neural network to model interpersonal dynamics among group members. Our results highlight the advantages of the graph network over the handcrafted features in modeling group affect, while also emphasizing the importance of temporal modeling and incorporating multimodal cues. Additionally, our analysis of affective convergence and divergence reveals that groups tend to diverge in their social signals during neutral collective affect, while exhibiting convergence during more emotionally intense moments. These insights are drawn from comparative results across both modeling techniques.**

***Index Terms*—Group affect, affect dynamics, annotations, convergence, divergence, multimodal analysis, automatic affect recognition.**

I. Introduction

GROUP affect is a *collective* social construct that represents the jointly experienced shared mood or emotions that group members hold in common at a given point in time [1], [2]. There is an important conceptual and empirical difference between the affective experiences of individual group members and the collective mood of the group as a whole, not only regarding how they are measured but also in terms of their influence on group outcomes [3]. Following [2], we focus our research on affect at a collective level in *purposive groups*.[1] By purposive group, we mean "an intact social system, complete with boundaries, interdependence for some shared purpose, and differentiated member roles" [4]. This makes purposive groups distinct from spontaneous social gatherings or large groups. Purposive groups are prevalent in organizational settings, tasked with completing a wide variety of assignments across different time frames [2]. Examples of purposive groups include a group of software developers or a group of health workers assembled to brainstorm solutions to a problem. The relevance of collective group affect towards the functioning and outcomes of purposive groups is documented well in the literature [5], [6], [7], [8], [9], [10], [11], [12]. For example, a widely cited review of the literature discusses the importance of group affect for shaping (i) group member attitudes, (ii) cooperation and conflict resolution, (iii) creativity and decision making, and (iv) group performance [2]. In terms of the factors that give rise to shared group affect, recent qualitative work points to social learning and positive emotional sharing [13].

However, there is a notable dearth of *quantitative* empirical research on *collective group affect* as it emerges and fluctuates during group interactions. Whereas research has extensively examined affect as an individual-level construct [14], [15], [16], [17], [18], we know much less about group affect as a collective phenomenon (for an overview and detailed critique, see [8]). This is because gathering suitable group interaction

Received 19 December 2024; revised 16 October 2025; accepted 4 December 2025. Date of publication 12 December 2025; date of current version 10 March 2026. This work was supported by Excellence Strategy of the Federal Government and the Länder, through the project "Mechanisms of Change in Dynamic Social Interaction" (LFF-FV79, Landesforschungsförderung Hamburg). Recommended for acceptance by A. Liu. *(Corresponding author: Navin Raj Prabhu.)*

Navin Raj Prabhu and Timo Gerkmann are with the Signal Processing Lab, University of Hamburg, 20146 Hamburg, Germany (e-mail: navin.raj.prabhu@uni-hamburg.de; timo.gerkmann@uni-hamburg.de).

Maria Tsfasman and Catharine Oertel are with the Department of Intelligent Systems, Technische Universiteit Delft, 2628 Delft, The Netherlands (e-mail: m.tsfasman@tudelft.nl; c.r.m.m.oertel@tudelft.nl).

Nale Lehmann-Willenbrock is with the Department of Industrial and Organizational Psychology, University of Hamburg, 20146 Hamburg, Germany (e-mail: nale.lehmann-willenbrock@uni-hamburg.de).

This article has supplementary downloadable material available at https://doi.org/10.1109/TAFFC.2025.3643752, provided by the authors.

Digital Object Identifier 10.1109/TAFFC.2025.3643752

[1]Throughout this work, we use the term "groups" to specifically represent purposive groups.

data, annotating, and analyzing *collective group affect* and the related social signals is considerably more complex compared to that of individual members' affect [9], [11]. In this work, we address this knowledge gap through two key contributions: (1) *collecting group affect annotations* using an annotation strategy that aligns methodologically with theoretical frameworks from organisational psychology on group affect, and (2) leveraging these annotations to perform *multimodal modeling of dynamic group affect*, capturing the underlying phenomena of convergence and divergence.

Collecting annotations of group affect: We identified two key challenges in annotating group affect from a review on prior work in organizational psychology and computer science: (i) capturing fine-grained temporal dynamics [8], [10], [19], [20], and (ii) accounting for social dynamics in complex purposive groups [2], [21].

The limited literature on group affect to date has predominantly relied on annotations of static images [9], [19], [22] or temporally independent video segments [11], [23] *without accounting for any temporal context* [8], [20]. Annotating video segments in a temporally independent manner without accounting for the temporal context is resource efficient, where annotators need not continuously track the group's affective expressions and its associated changes between consecutive segments. However, such an annotation approach is agnostic to the inherently dynamic nature of group affect [2], [10], which limits the resulting empirical contributions and represents a misalignment between the theoretical construct of group affect as a dynamic process and its measurement [7], [24].

Based on our review across disciplines, we observe considerable *theory-method misalignment* between theoretical models of group affect as presented in the organizational behavior literature (e.g., [1], [2]) and the group affect recognition methodologies developed in computer science studies (e.g., [9], [19], [25]). While the theory conceptualizes group affect in complex *purposive group settings* involving interpersonal relationships and dynamics [2], group affect recognition research has focused on much simpler groups that lack social intactness and interdependence towards a shared goal. As one example, consider the groups captured in the Video Group Affect dataset, which includes concerts, silent protests, and fights [11]. Notably, annotating and modeling collective affect in purposive groups is significantly more complex, as it requires capturing the dynamics of interpersonal relationships, unlike simpler non-purposive groups where such relationships are largely absent.

The two challenges above informed the design of our annotation strategy. To address the challenge of capturing *temporal dynamics* in group affect, we conducted pilot studies to iteratively tune the annotation window size, enabling more apt representation of affective fluctuations. To handle the complexity of annotating affect in *purposive groups*, we extensively trained organizational psychology students, ensuring context-aware, high-quality labels. This approach is consistent with prior research emphasizing the need for domain-specific training and the involvement of social psychologists in annotating group-level phenomena [20], [26]. Informed by [27], our training incorporated video markers identified during the pilot study. Together, these efforts ensure that our annotation process is well-aligned with organizational psychology theory, which frames group affect as a dynamic social construct within purposive groups.

Multimodal modeling of dynamic, group-level affect: Annotating dynamic group affect allows for (i) the *development of multimodal predictive models* capable of automatic recognition of group affect with more real-world application possibilities, coping with changing affect over time, and higher robustness [8], and (ii) the *quantitative analysis of convergence and divergence* in affective expressions that shape the dynamics of group affect [28].

Social signals, such as facial expressions [29] and vocal pitch [14], encode micro-level behavioral cues like synchrony and convergence [30], which are fundamental to the development of dynamic interpersonal relationships [30] and higher-order group constructs [26], [31]. In this work, we introduce two complementary methods for capturing these micro-level patterns using multimodal, individual-level social signals. The first approach leverages handcrafted features designed to quantify synchrony and convergence, thereby modeling interpersonal dynamics (e.g., [26], [31]). The second approach employs a novel graph neural network (GNN), inspired by social network theory [32], which represents individuals as nodes and their relationships as edges, allowing it to learn the structure and temporal evolution of social interactions. Unlike handcrafted features, the GNN offers a data-driven framework for modeling dynamic interpersonal relationships.

Barsade [2], in a review of the organizational psychology literature, highlighted the need for fine-grained analyses of the temporal dynamics underlying collective group affect. In particular, the study of *convergence* and *divergence* in affective expressions are ways to study this ebb and flow of group affect. Hareli and Rafaeli [28] proposed a cyclical model where individual affect converges via contagion or diverges through reactivity, creating a feedback loop that shapes group functioning. Recent conceptual work further emphasizes the need to account for social dynamics in that lead to shared affect in groups [3]. Despite its theoretical significance, empirical research on affective convergence and divergence in group interactions remains limited. To address this gap, we conduct a quantitative analysis of affective convergence and divergence using both the handcrafted features and the data-driven representations from the proposed graph network.

To the best of our knowledge, this is the first study to develop predictive models of *dynamic* group affect in complex *purposive* group interactions, and the first to present an empirical analysis of affective *convergence and divergence* in such dynamic settings. For a detailed comparison of our study contribution beyond existing approaches in the literature, see Appendix Table S2.

II. Background and Related Work

A. Affect in Purposive Groups

While individual affect has been extensively studied in the past few decades [14], collective group affect has received considerably less research attention [2], [8], [20]. Empirical research [33] has demonstrated that group-level affect is distinct from individual-level affect and that they are shared among interlocutors in purposive interactions. This has consequences

for research design, requiring scholars to pay attention to the actual social interaction dynamics that give rise to group affect (for an overview, see [7]). Understanding group affect is a key element of understanding how groups interact and achieve collaborative performance [2], [7]. This study specifically focuses on the under-researched construct of collective group affect.

To analyze affect in groups, quantification of affect is the first step. Two types of quantification techniques have been widely used in the literature: (1) Ekman's six basic emotions [34] (e.g., happy, angry) or (2) Russell's circumplex model [35], where affect is quantified using two continuous, bipolar, and orthogonal dimensions of arousal and valence. In recent years, research in social science as well as state-of-the-art datasets for affect recognition in computer science have moved from categorical representations to the circumplex model, owing to the fact that the circumplex model is more suitable for capturing the ambiguous and fuzzy nature of affective expressions (e.g., [10], [15], [36]). In line with this development across disciplines, in this work we rely on the circumplex model to quantify collective group affect.

B. Dynamics Inherent in Group Affect

Theorizing on group affect as a dynamic, temporally evolving social construct characterized by bottom-up and top-down processes, while intuitively appealing, has received only limited empirical validation efforts [2]. This is likely due to the challenge of annotating collective group behavior while also accounting for its inherent temporal dynamics [19], [20]. Two key challenges in this space are: (i) the time and cost inefficiencies stemming from the complexity of group constructs [10], [11], [19], and (ii) the need for domain expertise, particularly the involvement of social psychologists, and intensive annotator training [20], [26].

Moreover, the appropriate window size required to capture the dynamics of group-level social constructs remains an open research question across disciplines. Choices regarding the temporal granularity of annotations differ substantially in the prior literature. For example, Mo et al. [37] used 20-second windows, whereas Barsade [38] or Lei et al. [10] used 2-minute windows to capture dynamic changes in group affect. Of note, recent research on *individual*-level affect has moved from segment-level annotations ($\approx$ 5 secs) [17] to much smaller time windows, as small as 10 ms [16], [18]. Annotating *larger* time windows is simpler and more resource-efficient, as annotators need not track micro-level behaviors and events, and require fewer labels for an entire interaction. However, such windows are less suited for capturing the dynamics of group affect, which often shifts and changes from one conversational moment to the next. In contrast, *smaller* time windows better reflect these dynamics but make the annotation process significantly more complex and time-consuming [18].

In this work, we address these challenges by training annotators with a background in social psychology to attentively track the full scope of group interactions and annotate for dynamic group affect. This training enables annotators to incorporate both top-down and bottom-up processes that underlie the emergence of group-level affect. To aptly capture the dynamic fluctuations of group affect, we eploy a window size that is iteratively tuned with respect to the construct and the social setting at hand.

C. Automatic Recognition of Dynamic Group Affect

Dhall et al. [9] were among the first to explore automatic group affect recognition using static web images of non-purposive groups (e.g., concerts, silent protests [39]), annotated with six levels of happiness intensity. This foundational work was extended by [25] and [40], and later evolved into a multimodal framework through the efforts of Sharma et al. [11], [41]. To promote further progress in the field, a series of benchmark challenges were introduced [22], [23], [39], and [42] open-sourced an interactive, multimodal group affect analysis toolkit for use in diverse social settings.

More recently, GNNs have gained attention for this purpose due to their ability to model relational structures [20]. Early GNN-based approaches [19], [43] primarily focused on static group images, capturing only spatial relationships among individuals and their environments, without the temporal context. Addressing this gap, Wang et al. [44] introduced a unimodal, video-based GNN that represents individuals as nodes and includes pseudo-nodes to aggregate spatial and temporal features between neighboring time frames, within a fixed-length, temporally independent video segment. Of note, the model is limited to fixed-duration inputs and does not scale well to longer temporal contexts, as the GNN's adjacency matrix expands proportionally with input duration and the growing number of pseudo-nodes.

Despite recent progress, a key limitation remains in group affect recognition research: insufficient temporal modeling—most existing methods treat video segments as temporally independent, overlooking how group affect unfolds over time. To address this, propose a GNN-based model that represents individuals as nodes initialized with multimodal, individual-level social cues, and with edges trained to capture interpersonal relationships. Subsequently, a long short-term memory (LSTM) layer operates on the edge representations to model temporal dynamics between consecutive video segments. Leveraging the collected, temporally evolving annotations, the model captures fluctuations in group affect across variable-length interactions—integrating both bottom-up and top-down processes, thus aligning computational modeling more closely with established psychological theory [2], [8], [10].

D. Fine-Grained Analyses of Dynamic Group Affect

The investigation of affective convergence and divergence allows insights into collective group-level affect and its underlying dynamics [28]. Organizational psychology literature [21], [28] identifies emotional contagion and reactivity as key mechanisms through which convergence and divergence emerge within social groups. These processes involve one person or group influencing the emotions or behaviors of others through the conscious or unconscious induction of emotional states and behavioral responses [28], [45]. For emotional contagion to occur, individuals must first express their emotional states through observable social behaviors, which others may then mirror (i.e., convergence) or contrast with (i.e., divergence). Capturing this dynamic, reciprocal exchange of micro-level behaviors requires the extraction

of *synchrony* and *convergence* features, as outlined in [30]. For example, group members may express more activated speech adapting to other group members' level of activation (emotional contagion), or they might gasp in response to a group member's ill-fated story (emotion reactivity).

To capture these affective contagion and reactivity within groups, we employ graph *attention* networks (GATs), a variant of GNNs that learn attention weights between interlocutors represented as nodes. Alongside an analysis based on handcrafted synchrony and convergence features, we examine the learned GAT attention weights to gain insight into convergence and divergence processes.

III. Dynamic Group Affect: Conceptualization and Annotation

The experimental workflow of this research comprises two pipelines: (1) the Annotation Strategy, used to collect group affect annotations, and (2) the Modeling process, which leverages these annotations to analyze dynamic group affect. An illustration of this workflow is in Appendix Fig. S2. This section focuses on the first pipeline, detailing the annotation strategy employed.

Barsade and Knight [2] defined group affect as the amalgamation of group members' affective states and the mutual influence of a group's affective context. Recently, theorists have expanded on the temporal dynamics of group affect, showing how momentary individual- and group- affective experiences become inputs for future group affective experiences [28], [46]. In line with these organization psychology theories on group affect [2], [7], [28], [46], we conceptualize group affect as a dynamic and continually evolving social phenomenon in groups. Specifically, we define *group affect* as the collective affective state of the group, which is the amalgamation of group members' affective states expressed during group interactions (i.e., a *bottom-up* process) and which in turn affects future affective experiences of the group (i.e., *top-down* influence). Of note, this amalgamation requires individual behavior to be expressed in terms of social signals which can be annotated by external annotators.

A. Dataset: MEMO

To investigate dynamic group affect in the context of a purposive group interactions, we drew suitable data from the multimodal longitudinal meeting corpus (MeMo) [47], consisting of 45 unscripted video-call discussions in groups of three to six participants. The MeMo corpus, with its setup of purpose-driven group discussions and longitudinal recordings, is an ideal dataset for studying group affect in purposive groups. To achieve maximum affect elicitation, we selected those MeMo group discussions that focused on group members' experiences during the COVID-19 pandemic, recorded in the year 2021. A group discussion on COVID-19 is likely to evoke strong emotional responses due to several factors, including polarized opinions, economic impact, hope and resilience, loss and grief, and shared experiences. The recorded group interactions lasted for approximately 45 mins (average duration of 41 mins and 35 secs; with a standard-deviation of 7 mins and 30 secs). As a longitudinal meeting corpus, participants were divided into 15 groups, and each group met 3 times over the course of 2 weeks using a video-conferencing software. At the start, the participants were reported to be complete strangers, having never met each other before the first session.

The interactions were guided by professional facilitators in order to promote active discussions among the group members. The selected facilitators were experienced in moderating meetings and facilitating creative sessions. To maximize the diversity of in-group opinions, participants were recruited from various COVID-19 affected demographics in every group, e.g., parents with young children, adults of age 50+, students, (ex)-business owners. Overall, 15 groups, totaling 53 participants (25 Male, 28 Female; 18-76 years old) and 4 moderators (3 Male, 1 Female; 24-45 years old) took part in the interactions collected as part of MeMo.

To facilitate future research on these novel group affect annotations, we commit to making the annotations public along with the MeMo corpus [47]. Initially, the annotations will be publicly accessible following the review process and can be obtained by agreeing to the end-user agreement.[2] It is important to note that the annotations can be released immediately, as they have already been anonymized using pseudonyms for both annotators and interlocutors and do not include any sensitive personal information. Conversely, the raw behavioral group interaction data recorded in the MeMo corpus contains sensitive and private information shared by the interlocutors [47]. To protect the data of the interlocutors, a thorough review process is currently being conducted to eliminate any sensitive or potentially private information. Following this process, the MeMo corpus will also be made publicly available in the same data repository as the annotations (please see [47] for more details).

For this work, we used only the spontaneous interaction segments within the MeMo corpus. All interactions had eye-gaze calibration and administrative tasks at the beginning and at the end, respectively. Furthermore, in some interactions, participants were late to join the interaction. We cropped these segments out of all the interaction videos for our annotation procedure. The timestamps of these segments were manually annotated by the lead author.

B. Annotators

In [26], [48], efforts to annotate conversation quality at both individual and group levels reveal that group-level constructs are more complex to annotate and typically have lower inter-annotator agreement than individual-level constructs. The importance of involving social psychologists in the annotation of group affect is further underscored in [20].

Following this, instead of relying on naive annotators, we trained annotators for the task of annotating dynamic group affect. This approach helped us overcome several limitations regarding the annotation of group affect in the prior literature. These include: (i) accurately capturing the dynamic bottom-up

[2]For access to the annotations, contact co-author Catharine Oertel.

and top-down processes involved in group affect [10], [19], and (ii) addressing the difficulty of conceptualizing group affect within complex purposive group interaction settings [9], [37], [41]. To achieve this, we recruited students with an organizational psychology background, either at the bachelor's or master's level. They were familiar with the topic of affect in groups through their study curriculum. A total of eight annotators (3 male and 5 female; ages 18-25) were recruited.

C. Initial Pilot Studies

We began with an initial pilot study to (1) refine the annotation procedure for the social construct and dataset, and (2) train our annotators. We selected three videos from the MeMo corpus that were perceived to have the maximum affective variances, which were then randomly assigned to four out of the eight annotators to annotate group-level arousal and valence using INTERACT [49]. As a primer for further training of annotators, we held individual meetings with each of the four annotators to discuss the definitions of dynamic group affect and the circumplex model.

1) Tuning the Annotation Procedure: We tuned the annotation procedure in terms of two key parameters: (i) the *scale* to be used to annotate group affect and (ii) the *time-window size* to be used to aptly capture the affective dynamics. For the pilot study, we began by utilizing the Self Assessment Manikin (SAM) [50] to annotate group-level arousal and valence [16], employing 20-second time windows in accordance with [37].

After the four annotators completed the annotations of the videos part of the pilot study, a meeting was setup with the annotators to discuss on the two parameters to be tuned. The discussion during the meeting was specifically on two questions: (i) "Should we increase or decrease or keep the same time-window size?", and, (ii) "Do you accept the scale used? Does it well capture the affective expressions and fluctuations in the interactions?".

Based on the consensus reached during this discussion, we made two changes to the initial setup for the annotation procedure. First, we reduced the time-window size to 15 secs. This decision was based on the consensus that the dynamics of group affect in the videos fluctuate rather faster and a smaller time-window size would capture the fluctuations more adequately. Second, we replaced the SAM scale with an ordinal scale ranging from 1 to 9. All annotators felt that the SAM scale had extreme arousal and valence categories on the scale that were not usually present in the observed spontaneous interactions in the MEMO corpus. Because of this, the resulting affect annotations had only limited affective variance. Our decision to replace the SAM scale with an ordinal scale aligns with arguments in favor of the ordinal nature of emotions in previous work [51].

We used an evidence-based approach to determine the most appropriate window-size to capture meaningful fluctuations in observed group-level affect. To this end, we iterated our pilot study until a consensus was reached between annotators to freeze the annotation procedure. While the annotators in our pilot study agreed that the initial window size of 20 secs was not adequate to capture all fluctuations in collective group affect, they reached a consensus that a 15 secs window was more effective in capturing these fluctuations. They also concurred that a smaller window size was unnecessary and would not add value in terms of capturing additional nuance in group affect dynamics. This choice aligns with the literature which indicates that collective group-level affect tends to change somewhat more slowly compared to individual-level affect [7]. More specifically, the development of group affect necessitates the expression of individual affect and the occurrence of a bottom-up process, leading to relatively slower fluctuations of collective group affect compared to fluctuations in individual affect.

Moreover, during the iterative tuning rounds, a critical problem raised by most of the annotators was that a particular moderator in the MeMo corpus dominated most of the interaction preventing the interaction from being an active discussion amongst all group members. With respect to this, we got rid of all 9 interactions from that particular moderator. After this, the final dataset had in total 35 group interactions. We attribute the richness of the discussions with the annotators to their educational backgrounds and the prior experience of some in annotating social behaviors.

2) Video Markers: A pitfall in using an ordinal scale over the SAM scale is that the annotators do not have a reference affective expression associated with each ordinal value of the scale, such as the illustrations in SAM. This leaves room for interpretation and confusion, as annotators are unsure when to assign a relatively higher or lower value on a scale. To tackle this problem, when training our annotators, we used *video markers* developed for each point on the ordinal scale. The objective was to link each point on the scale to a specific expression of collective group affect by identifying a fitting 15 secs video segment from within the MeMo corpus, which would serve as a behavioral anchor for the respective point on the scale. This video marker-based training of the annotators was inspired by the nonverbal behavioral anchors for affect expressions developed by Bartel and Saavedra [27], which were subsequently adopted by other works for the annotation of group affect and group mood (e.g., [10], [36], [38]). Notably, whereas Bartel and Saavedra [27] devised a list of nonverbal behaviors describing each point of their rating instrument, we aimed to contextualize this information and provide a temporal setting for each type of group affect expression, so annotators would have specific video segments that clearly illustrated each point on our ordinal scale for annotating group affect, for their reference during the annotation procedure.

To create these video markers, we began by utilizing the annotations derived from the fine-tuning processes of the annotation procedure. We selected several potential candidates for each element of the ordinal scale. These candidates were discussed with the four annotators who participated in the pilot study, using the two definitions provided to them (i.e., the definition of dynamic group affect and the circumplex model). Of note, the development of video markers was carried out alongside the training of the annotators. Throughout the training and the accompanying discussions among the annotators, the video markers were consistently adjusted, with the goal of establishing a single, definitive video marker for each element of the ordinal scale. The specific topics covered in these discussions are presented in the next section.

3) Training Annotators: The main objective of the training was to help annotators focus on the most important aspects of dynamic group affect, with discussions conducted on the following topics.

Aggregation of affective expressions: Two key ideas were discussed under this topic: (i) "How do you *aggregate the individual affective* expressions to group affect?" (i.e., on the bottom-up phenomena), and, (ii) "How do you track and *aggregate the dynamic fluctuations* in group affect?" (i.e., on the top-down phenomena). Moreover, owing to the ordinal nature of scale, the scale differences between video markers belonging to adjacent scale elements were also discussed.

Contribution of the Group Facilitators: The role and the contribution of the group facilitator in the interaction towards the group affective state was also discussed. To capture the unscripted nature of the interaction, we instructed the annotators to treat the moderator as one of the group members and not to perceive them differently from the participants, especially in light of the bottom-up nature of collective group affect.

Focusing on Affective Expressions: Research on emotional contagion and group affect indicates that behaviorally expressing an emotion, such as smiling, can lead to experiencing it, such as feeling happiness [38]. Accordingly, in this work, we instructed annotators to focus solely on the affect expressed by group members, following the circumplex model of affective expressions [35]. For example, when one of the observed group members used sarcasm (i.e., a positive valence expression to convey negatively valenced conversational content), the annotators were instructed to only focus on the affect *expressed* (in this case, positive valence). This approach generally aligns with the affect recognition literature that prioritizes *expressed* over experienced emotion [14], [16], [17].

Based on these discussion topics, one candidate was selected as the emotion marker for a particular scale item. These emotion markers are then used as reference videos to explain the respective scale item. At the end, all other annotators who were not part of the initial pilot studies were also trained using the outcome of the studies and the video markers derived for the ordinal scale.

D. Annotation Procedure

1) Annotation Software Setup and Location: During training and for the entire annotation procedure that followed, we used INTERACT software [49] which provides a graphical user interface where annotators can scroll through videos to annotate observed behavior directly from an audio or video file. The software allowed us to systematize and synchronize the entire annotation procedure across annotators. INTERACT was set up to request an annotation for every 15 seconds segment of each video. Clicking on each segment would play the respective slice of the video. The annotators were allowed to watch each 15 seconds time-window any number of times. They used their number pad in the keyboard to input a number between 1 and 9 as their ordinal scale annotations, and any wrong input would not be accepted by INTERACT. To ensure an appropriate setting without distractions, annotators were asked to come to the laboratory where they were provided with an individual workspace, including a desktop computer with INTERACT installed and a two-screen setup.

TABLE I
INTER-ANNOTATOR AGREEMENTS

	Ours		Comparison	
	Arousal	Valence	Arousal	Valence
Quadratic κ	0.41	**0.58**	**0.50**	0.54 [18]
Cronbach's α	**0.82**	**0.89**	0.80	**0.89** [37]
Correlation (PCC)	**0.51**	**0.64**	0.44	0.41 [16]

2) Distribution of Videos: The 35 interaction videos were provided in a randomized order to each of the annotators, to prevent any potential annotator bias which might occur if all the annotators received the videos in the same order. The videos were distributed in batches of 5 videos to ensure that the annotators followed the order provided. The annotators worked 10-20 hours per week, annotating approximately 2-5 videos per week. The annotation procedure ensured that each group discussion received at least five annotations throughout (i.e., five annotations for each 15-second window of each of the 35 group discussions). In comparison, most state-of-the-art individual-level affect recognition datasets have three to five annotators [17], [52], with literature on uncertainty modeling suggesting that at least four annotations should be collected for a reliable annotation distribution and its ground-truth consensus [53]. The annotations for arousal and valence were done independently, similar to [17], [18], with the annotators watching the entire video separately while annotating for each dimension of affect.

3) Inter-Annotator Agreement: To measure the inter-annotator agreement, we used three different metrics: (i) quadratic weighted kappa (κ) [54], (ii) Cronbach's alpha (α) [55], and (iii) Pearson's correlation coefficients (PCC) [56]. The three metrics are chosen with their respective advantages in mind. Firstly, the quadratic weighted kappa measure κ is a variant of Cohen's kappa that is specifically designed to measure agreements in the ordinal scale data [54]. However, it is constraint to calculating agreement between only pairs of annotators. Hence, following common practices [18], [26], [48], we report the average in all possible combinations of annotators. Secondly, Cronbach's alpha (α) measure overcomes this limitation by allowing one to measure the agreement between an arbitrary number of annotators. Finally, as we annotate a time-dependent dynamic construct, we also use the PCC measure as the inter-annotator agreement metric.

As the group affect annotations collected here are novel, both in terms of the social setting at hand and its dynamic nature that captures the temporal fluctuations, there is no direct comparison that can be made in terms of the agreement scores. However, to better understand the agreement scores, we compare their value with other agreement scores reported in the existing literature on collecting affect annotations (see Table I). The criteria for selecting the literature to compare with is as follows: (i) the annotations are performed on affect either at the individual-level or group-level, (ii) the agreement scores are presented on the metrics chosen above. Note that no hard

TABLE II
QUADRATIC WEIGHTED KAPPA κ SCORES WHEN A PARTICULAR ANNOTATOR IS EXCLUDED

	Arousal		Valence	
All	0.407		0.578	
Excluded Annotator	κ	Δ	κ	Δ
1	0.402	−0.005	0.573	−0.005
2	0.398	−0.009	0.572	−0.006
3	0.416	+0.009	0.581	+0.003
4	0.402	−0.005	0.575	−0.002
5	0.412	+0.005	0.567	−0.011
6	0.401	−0.006	0.598	+0.020
7	0.399	−0.008	0.569	−0.009
8	0.385	−0.022	0.582	−0.004

filter was set on the social setting at hand, due to the less availability of literature on group affect in an purposive interaction setting.

Quadratic weighted kappa κ measure: From Table I, we observe a κ =0.41 and κ =0.58 for arousal and valence, respectively. This indicates a moderate agreement for both arousal and valence dimensions of group affect [54]. As a comparison, which is also presented in Table I, the MSP-Conversation dataset [18], on a much simpler annotation task of individual-level affect, has an agreement of κ =0.50 and κ =0.54 for arousal and valence, respectively (i.e., higher than ours on arousal and lower on valence).

Cronbach's α measure: From Table I, we observe an α of 0.82 and 0.89 for arousal and valence, respectively, indicating a good and satisfactory level of agreement [55]. For comparison, the annotations of individual affect in groups collected by [37] have an α agreement of 0.80 for arousal and 0.89 for valence, which is lower than our arousal annotations and virtually the same as our valence annotations. Note here that [37] worked on a much simpler social settings without accounting for interactions between interlocutors.

Pearson Correlation (PCC) measure: From Table I, we observe a PCC of 0.51 and 0.64 for arousal and valence, respectively. The individual-level affect annotations in the RECOLA dataset [16], in comparison, have a PCC agreement of 0.435 for arousal and 0.407 for valence, which is much less than our agreement scores, despite being on a much more simpler social construct.

From Table I we note that in general the agreement for valence is higher than that of arousal, which aligns with the literature [18], [53]. Furthermore, in Table II, we present the κ scores when a particular annotator was excluded, where Δ is the increase (+) or decrease (−) in κ when the annotator was excluded. Table II reveals that in most cases (i.e., for six out of eight annotators), excluding the annotator only decreases the κ score (−) in terms of both arousal and valence. Only for annotator 5 a large increase in Δ is noted when excluded, i.e., Δ of $+0.020$ is noted for valence. In all other cases, the Δ is rather minimal, i.e., $\Delta < +0.01$. This points to the reliability of the collected annotations, even when annotations from an annotator are omitted.

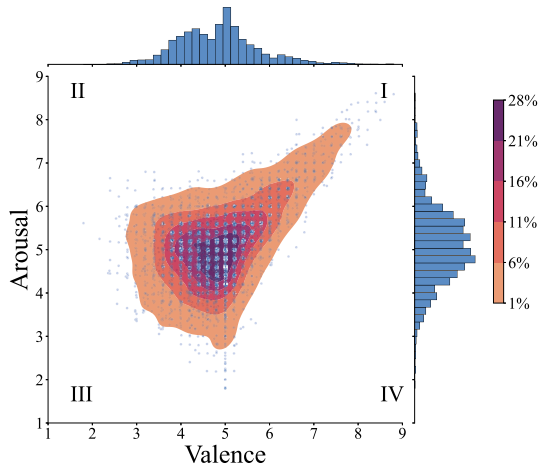


Fig. 1. Distribution of final ground-truth for dynamic group affect $y_{s,t}^{\mathrm{EWE},(i)}$, in terms of arousal and valence.

E. Ground-Truth

To derive the final ground-truth of dynamic group affect, we use the Evaluator Weighted Estimator (EWE) [57], a common technique to derive ground-truth for individual-level affect recognition [16], [18]. The EWE, to derive the ground-truth, weights annotations with respect to the annotator-wise correlation coefficients, and is formulated for a time segment t in an interaction sample s as follows:

$$y_{s,t}^{\mathrm{EWE},(i)} = \frac{1}{\sum_{k=1}^{K} r_{k,s}^{(i)}} \sum_{k=1}^{K} r_{k,s}^{(i)} \quad y_{s,t}^{(i)} \tag{1}$$

where $r_{k,s}^{(i)}$ denotes the average correlation coefficient of the annotator k with other annotators for a particular interaction sample s and emotion dimension i. For unreliable annotators with $r_{k,s}^{(i)} < 0$ a lower bound of zero is defined. In cases where all k annotators have the same correlation coefficients $r_{k,s}^{(i)}$, they result in the same correlation weighting and thereby $y_{s,t}^{\mathrm{EWE},(i)} = y_{s,t}^{(i)}$. For the rest of the experiments and analyses in this work we will use $y_{s,t}^{\mathrm{EWE},(i)}$ as the ground-truth of dynamic group affect.

The distribution of the ground-truth in terms of the arousal and valence dimensions can be seen in Fig. 1. From the plot, we note that the independently annotated arousal and valence, depicted by the histograms of the marginal distribution, have considerable variances. Arousal annotations are distributed with a mean (m) of 5.10, a standard-deviation (s) of 0.88, a minimum value ($\min$) of 1.80, and a maximum value ($\max$) of 8.61. Similarly, the valence annotations are distributed with $m = 4.84$, $s = 0.90$, $\min = 1.80$, and $\max = 8.80$.

Samples in Quadrant I denote high-arousal and positive-valence (e.g., emotions such as happy and excitement). Quadrant II denotes high-arousal and negative-valence (e.g., emotions such as anger and frustration). Quadrant III denotes low-arousal and negative-valence (e.g., emotions such as depressed and gloomy). Quadrant IV denotes low-arousal and positive-valence

(e.g., emotions such as relaxed). With respect to the joint distribution of arousal and valence, depicted by the heatmap and the scatter plots (in Fig. 1), the variances are noted only in some of these quadrants of the circumplex model. For example, good number of samples are noted in Quadrant I. Similarly, a good number of samples can also be noted for low-arousal and neutral-valence (i.e., between Quadrant III and IV; emotions like tired and melancholic), and for neutral-arousal and negative-valence (i.e., between Quadrant II and III; emotions like bored and sad). However, in Quadrant IV and Quadrant II extreme samples are not noted. This could be because the participants in MeMo were reported to be non-preacquainted and complete strangers at the beginning of the longitudinal study that spanned over 3 interactions in two weeks. In such short longitudinal cases, amongst non-preacquainted participants, extreme expressions are very rare [58], likely due to professional behavioral norms.

See Appendix Section 2 for examples of the collected annotations and corresponding video frames, providing a qualitative illustration of how the annotations capture the dynamic ebb and flow of group affect. These examples, together with the inter-annotator agreement results, thereby substantiate the *first contribution*—the development of an annotation strategy for group affect that establishes theory-method alignment with concepts from organizational psychology—outlined in Section I.

IV. Modeling Group Affect Dynamics

This section outlines our approach to modeling dynamic group affect. We begin with preprocessing of raw audio and video data (Section IV-A), followed by the extraction of individual-level features (Section IV-B). Finally, in Section IV-C, we present two complementary methods for modeling group affect as a dynamic and collective construct: (i) handcrafted features capturing synchrony and convergence, and (ii) a graph neural network-based approach that models interpersonal relationships in a data-driven manner.

A. Preprocessing

The MeMo corpus provides with manually diarized and synchronized audio for each of the interlocutors, collected at a sample rate of 16 kHz [47]. Similarly, video recordings of online group discussions are also provided at a frame rate of 60 fps. The group discussion video frames are cropped to obtain individual-level frames of each of the interlocutors.

B. Individual-Level Feature Extraction

Existing research works have revealed the multimodal nature of affect. For instance, the audio modality has shown to be more informative of the arousal dimension of affect, whereas the video and text modalities better explain the valence dimension [59], [60], [61], [62]. To this, we employ a multimodal strategy to model group affect, by employing both audio-based and video-based features.

Audio features: As the audio features we extract the first 5 *MFCC* coefficients, *Voice Intensity*, *Pitch*, *VGGish*, and the *Speech Rate*. This set of individual-level paralinguistic audio cues are selected owing to their demonstrated effectiveness in the automatic recognition of individual-level affect [63] and also other group-level constructs such as cohesion [31]. The MFCC and pitch features were calculated using the `librosa` package, for every 10 ms with a sliding window of 30 ms. The voice intensity and speech rate features were extracted using `Praat` [64]. The voice intensity was calculated at the same rate as the MFCCs while the speech rate was calculated in vowels per second using [65] at a rate of 1.5 secs following [31]. The VGGish features are pretrained deep learning based features and was extracted using the pretrained weights from [66].

Video features: As the video features we extract *Facial Action Units (AUs)*, *Face Pose*, and *ResNet50*. Action units, Face pose and ResNet50 features have been successfully used in existing literature for several tasks, such as sentiment analysis [67] and emotion recognition [29]. Individual-level AUs (subset available in the `OpenFace` toolkit) and the face pose (pitch, roll and yaw) were extracted using the `OpenFace` toolkit [29], for every 0.5 sec. Similar to the VGGish audio features, the ResNet50 network was used to extract framewise pretrained deep learning based features and was extracted using the pretrained weights from [68].

C. Group-Level Modeling Techniques

1) Handcrafted Features Based Group Modeling: Social interactions are multilevel systems where interpersonal relationships and affective states emerge at multiple levels of the interaction, i.e., at the individual, dyadic and group level [69]. With respect to this theoretical framework of group-level constructs, in this work, to study dynamic group affect, from the individual-level features we extract dyadic-level and group-level features that are descriptive of the interpersonal relationships shared between a dyad in the interaction and the group as a whole, respectively. The complete list of handcrafted features extracted can be seen in the Appendix Table S1.

Dyad-level features: We extract two sets of dyad-level interpersonal relationship-based features: (1) Synchrony and (2) Convergence. As the synchrony feature set, following [26], we use linear correlation coefficient-based measures: (i) the correlation coefficient ρ, the linear correlation without a time-lag, (ii) lagged correlation ρ_δ, the linear correlation with the best time-lag, and (iii) the best lag δ defined as the time-lag used to obtain the maximum linear correlation between the two individual-level signals. Existing literature notes that synchronous behavior is displayed by interlocutors often in a time-lagged manner, with a leader and a follower [26], [30]. The three synchrony measures are extracted using the formulation below:

$$\begin{aligned} &\text{Correlation coeff. } \rho : X \circledast Y \\ &\text{Lagged correlation } \rho_\delta : \max z(X,Y) \\ &\text{Best lag } \delta : \arg\max_l z(X,Y,l) - ||X|| + 1 \end{aligned} \tag{2}$$

$$z(X,Y,l) = \sum_{k=0}^{||X||-1} X_l \circledast Y_{k-l+N-1} \tag{3}$$

where z(.,.) is the cross-correlation function, $\circledast$ denotes linear correlation between two signals, $||X||$ denotes the length of

signal X, $l = 0, 1, \ldots, ||X|| + ||Y|| - 2$ denoting the time-lags possible, and $N = \max(||X||, ||Y||)$.

Following the technique proposed in [70] and [26], we extract three convergence features: (i) global, (ii) symmetric and (iii) asymmetric convergence. Global convergence captures the change in similarity between two individual's social signals, specifically between the initial time-segments and the later time-segments. Similarly, symmetric and asymmetric convergence features capture the decrease or increase in similarity between the two individual's social signals, without and with a time-lag, respectively. The three measures are formulated as:

$$\begin{aligned} \text{Global } \Theta_{\text{gbl}}: & \sum_{i=0}^{||X||/2} (X_i - Y_i)^2 - \sum_{j=||X||/2}^{||X||} (X_j - Y_j)^2 \\ \text{Symmetric } \Theta_{\text{s}}: & \ (X_l - Y_l)^2 \circledast L \\ \text{Asymmetric } \Theta_{\text{as}}: & \ p(Y_b/\theta_{X_a}) \circledast L \end{aligned} \tag{4}$$

where, $L = [0, 1, \ldots, ||X||]$, $l \in L$, and θ is the parameter of a Gaussian mixture model (GMM) trained using the expectation-maximization procedure on the data points from X in the initial period of the interaction, i.e., $a \in [0, m]$, $m = 2 \cdot ||X||/3$. Similarly, Y_b are data points from Y in the later period of the interaction, i.e., $a \in [m, ||Y||]$.

Group-level features: To extract *group-level* features from the individual and dyadic features, following [26], [31], we use six aggregation techniques that are agnostic to group size: `average`, `standard-deviation`, `median`, `minimum`, `maximum`, and `gradient`. The core idea here is that these different types of aggregations, each with an unique approach, describe the distribution of dyadic-level features within a group thereby capturing the interpersonal nuances within all possible dyads in the group. The `average` and `standard-deviation` explains the average and the deviation from the average of synchrony measures across all possible dyads in a group. Similarly, the `gradient` aggregator explains the deviation between the least and most synchronous dyad, i.e., the absolute difference between the `minimum` and `maximum`.

It is important to note that the group-level aggregations were computed from fine-grained temporal features at the dyad level, capturing detailed nuances in the temporal dimension. However, one could argue that these aggregations may miss finer details in the interpersonal relationship dimension. To address this, we propose using GNNs in the following section as a data-driven alternative to handcrafted features for modeling group affect.

2) Graph-Based Group Modeling: Grounded in social network theory [32], we frame group affect modeling as a graph classification task using GNNs, where the group interaction is represented as an undirected graph $\mathcal{G} = (V, E)$, with V denoting the set of M nodes and E the set of edges. Each node $V_i \in V$ represents an interlocutor with individual-level features $h_i \in \mathbb{R}^F$, and each edge $E_{i,j} \in E$ denotes a connection between two nodes. The adjacency matrix A captures the edge structure, where $A_{i,j} = 1$ if nodes i and j are connected and $A_{i,j} = 0$ otherwise.

A standard GNN training pipeline, as introduced in Graph Convolutional Networks (GCN) by Kipf et al. [71], consists of two main steps: (1) *convolution*, where each node transforms its feature representation (h_i) to share with adjacent nodes, and (2) *message passing*, wherein these features are propagated to the adjacent nodes. Nodes subsequently update their representations by aggregating information from their neighbors, typically using simple operations such as summation or averaging. However, this approach can produce identical output features for nodes with identical neighborhoods, thereby limiting the model's expressiveness for certain graph structures. For example, this poses a challenge in group affect modeling, where interlocutors are connected through a fully connected graph that captures overall group membership but overlooks the distinct relationships between individuals. To address this limitation, we employ an attention-based message passing mechanism, as proposed in Graph Attention Networks (GAT) [72].

Unlike basic sum or average aggregations in GCN, the attention based GAT compute a weighted average of multiple node features. These weights are dynamically determined based on a combination of the node's own features, the interlocutor's individual-level features (h_i), and the features of adjacent nodes (h_j), which represent the interacting counterparts of the interlocutor. The GAT layer obtains an attention based aggregation formulated as:

$$h_i^{(l+1)} = \phi\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}^{(l)} h_j^{(l)}\right), \tag{5}$$

where $\mathbf{W}^{(l)}$ is the learnable weight parameters that transforms the node features, ϕ represents an arbitrary activation function, and α_{ij}, the learnable attention weight between nodes i and j. The attention weight α_{ij} is formulated as:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}\left[\mathbf{W}h_i || \mathbf{W}h_j\right]\right)\right)}{\sum_{k \in \mathcal{N}_i} \exp\left(\text{LeakyReLU}\left(\vec{\mathbf{a}}\left[\mathbf{W}h_i || \mathbf{W}h_k\right]\right)\right)}, \tag{6}$$

where $||$ denotes concatenation, $\vec{\mathbf{a}}$ represents the attention mechanism, implemented as a single-layer feedforward neural network parameterized by a weight vector $\vec{\mathbf{a}} \in \mathbb{R}^{2F}$, LeakyReLU is the chosen activation function, and $\mathcal{N}_i$ denotes the indices of the adjacent nodes of node i. To accommodate varying group sizes (g), we fix the number of nodes M to the maximum group size in the dataset. For groups with fewer than M members, dummy nodes without edges are added, ensuring that $A_{x,j} = 0$ for such nodes ($g < x \leq M$).

The overall architecture consists of three GAT layers followed by an LSTM, enabling temporal modeling across consecutive 15-second time segments. Mini-batching is employed to ensure that each batch contains segments from the same interaction and maintains temporal continuity. This batching strategy is commonly used in end-to-end continuous emotion recognition techniques [53], [59], [73].

V. Results

This section analyzes and discusses the results of group affect modeling. First, Section V-A presents predictive modeling based

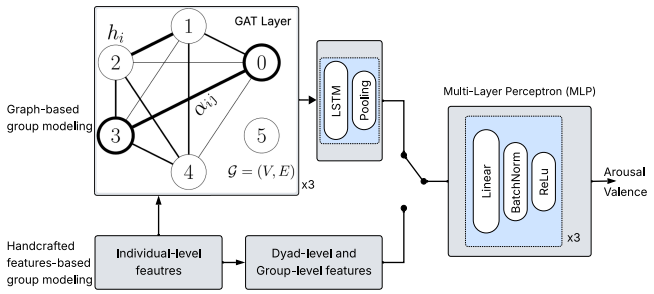


Fig. 2. Overview of the predictive modeling architecture.

on the collected group affect annotations. Then, Section V-B explores quantitative analyses of the convergence and divergence phenomena.

A. Predictive Modeling

Using the handcrafted features described in Section IV-C1 and the graph neural network in Section IV-C2, we frame the predictive modeling of group affect as a regression task. Fig. 2 presents the block diagram of the modeling technique's architecture. As shown, a simple Multi-Layer Perceptron (MLP) performs the regression task, using as input either the handcrafted features or the average pooled output of the GAT layers. The MLP is made up of three linear layers with ReLU and Batch Norm after each of the layers. The MLP's architecture was tuned with respect to the loss obtained on the validation dataset. The code for the modeling technique, including the extraction of handcrafted features and the implementation of the graph neural network, is publicly available and can be found here[3].

1) Experimental Setup: Feature Sets: To evaluate the predictive capabilities of the extracted *handcrafted features* (**HCF**), we employ three sets of group-level features: (1) Basic, where individual-level features are directly aggregated to the group level using group-level aggregators. This set does not account for interpersonal or dyadic relationships and only partially captures social signal dynamics due to the use of average temporal and group-level aggregations. (2) Synchrony (or Sync.), where dyadic synchrony and convergence-based features are extracted before applying group-level aggregations. This set effectively captures both social signal dynamics and interpersonal relationships among interlocutors. (3) Combined (or Comb.), which fuses the Basic and Synchrony feature sets, integrating both aspects into a unified model. These feature sets are further classified into audio, video, and audio-visual categories.

Contrarily, to study the predictive capabilities of the proposed *graph based modeling* (**GAT**), we use only individual-level features (or Indiv.) as input, categorized into audio, video, and audio-visual feature sets. For this, we initialise the node features h_i of the graph $\mathcal{G}$ with the individual-level features of the respective interlocutor.

Loss Function: The concordance correlation coefficient (CCC) [74] is used as the loss function and as the metric to validate the performances. The CCC has been widely used in literature for the task of individual-level affect recognition [14]. The CCC measures the agreement between two variables and ranges from -1 to $+1$, with perfect agreement at $+1$. In contrast to Pearson's correlation, the CCC takes both the linear correlation and the bias in to account, which makes it preferable over Pearson's correlation as the loss function and as the evaluation metric.

[3]https://github.com/sp-uhh/group_affect

TABLE III
RESULTS OF THE PREDICTIVE MODELING

	Feature Set	TM	Model	Arousal CCC ↑	Valence CCC ↑	Avg. CCC ↑
Audio	Basic	×	*HCF*	0.242	0.245	0.244
	Sync.	×	*HCF*	0.215	0.203	0.209
	Comb.	×	*HCF*	0.261	0.222	0.241
	Comb.	✓	*HCF*	0.274	0.242	0.258
	Indiv.	×	*GAT*	0.253	**0.263**	0.253
	Indiv.	✓	*GAT*	**0.286**	0.255	**0.271**
Video	Basic	×	*HCF*	0.315	0.405	0.360
	Sync.	×	*HCF*	0.342	0.428	0.385
	Comb.	×	*HCF*	0.355	0.448	0.401
	Comb.	✓	*HCF*	0.339	0.445	0.392
	Indiv.	×	*GAT*	0.344	0.456	0.400
	Indiv.	✓	*GAT*	**0.361**	**0.483**	**0.422**
Audio-Visual	Basic	×	*HCF*	0.293	0.332	0.313
	Sync.	×	*HCF*	0.403	0.428	0.401
	Comb.	×	*HCF*	0.416	0.431	0.424
	Comb.	✓	*HCF*	0.396	0.447	0.422
	Indiv.	×	*GAT*	0.384	0.484	0.434
	Indiv.	✓	*GAT*	**0.420**	**0.508**	**0.464**

Training Strategy: The models are trained using the ADAM optimizer with a learning rate of 10^{-4}. The strategy includes an early-stopping on the validation loss improvements with a patience of 10 epochs. The best model during training is selected as the one with the best validation loss.

Data Partition: The dataset is partitioned following the strategy proposed in [17]. The partitions were made such that there is no speaker overlap between the training and testing datasets. This also includes non-overlapping moderators in the MeMo corpus. The validation dataset however may have an overlap of moderators in some samples, but not overlapping participants. Overall, the training-testing split is made with a 80-20% split, and 10% of the training datasets is split for the validation dataset.

2) Discussion: Table III presents the results of the predictive modeling. The performance is evaluated across four feature sets (Basic, Synchrony, Combined, and Individual-level), three modalities (Audio, Video, and Audio-Visual), two modeling techniques (HCF and GAT), and two temporal settings (with and without modeling temporal relationships between consecutive 15 s segments).

Multimodal nature of group affect: The results reveal that dynamic group affect is best captured in a multimodal manner, with the audio-visual feature set obtaining the best performance of 0.420 and 0.508 in terms of arousal and valence, respectively. Furthermore, the visual feature set outperforms the audio feature sets in predicting group affect, demonstrating superior performance across feature sets and modeling techniques. We also note that the audio modality better explains the arousal dimension than the valence dimension, while the video modality better predicts the valence dimension. However, their performance

differences are not as large as noted in individual-level affect recognition literature [62].

Relevance of temporal modeling: In Table III, a ✓ in the "TM" column indicates that the method models temporal relationships between consecutive 15-second segments, while a × denotes it does not. Capturing these temporal dynamics allows the model to represent affective processes such as the top-down influence of group-level affect on individuals through emotional contagion (as discussed in [2], [75], and partially explored in [19]). Results show that temporal modeling notably enhances group affect prediction, particularly in GAT-based models, where CCC improvements are more consistent across modalities. To our knowledge, this is the first work to effectively capture this top-down mechanism of group affect, addressing a key limitation in prior studies [9], [19] (as discussed in Section II-C). This was made possible by collecting group affect annotations that reflect temporal dynamics.

Relevance of capturing interpersonal dynamics: The results highlight the significance of capturing interpersonal dynamics in group affect modeling, as both HCF and GAT modeling techniques demonstrate improved performance. For example, incorporating synchrony and convergence features into the video and audio-visual sets increases average CCC from 0.360 to 0.401 and from 0.313 to 0.445, respectively. However, in the audio modality, synchrony features perform worse than the basic set, likely due to segments where most interlocutors are silent, limiting synchrony extraction. This issue does not occur in the video modality, where participants are typically active, such as in attentive listening scenarios [76].

HCF vs GAT for capturing interpersonal relationships: Regarding the modeling technique used to capture interpersonal relationships, the data-driven GAT consistently outperforms the handcrafted synchrony and convergence features (HCF) across all modalities. The most notable improvement is observed in the audio-visual setting, where GAT achieves a CCC of 0.410 for arousal and 0.508 for valence, compared to HCF's 0.396 and 0.447, respectively. The performance gain is greater for valence than for arousal, highlighting GAT's stronger ability to model interpersonal dynamics associated with emotional valence. The GAT model's improved performance underscores its effectiveness in modeling group affect by capturing micro-level, multimodal social dynamics.

B. Analysis on Affective Convergence and Divergence

With the majority of empirical research focused on static group affect [9], [19], [22], Kelly & Barsade [75] emphasized on the dynamic nature of affect, i.e., how, over time, the nature of collective affect can change. The ebb-and-flow of collective group affect over time is primarily characterized by the affective convergence and divergence underlying the bottom-up and top-down processes of group affect [2], [28]. Foundational theories on affective dynamics (e.g., [45] and [28]) further describe how several individual interaction- and behavior-level mechanisms, including facial mimicry, emotional similarity and dissimilarity, and empathy, contribute to affective *convergence* and *divergence* in groups.

Building on these theorizations, in this section, we present a quantitative analysis on the relationship between interaction- and behavior-level cues, that quantify the level of affective convergence and divergence within interlocutors, and the collected annotations of dynamic group affect. To this end, we analyse affective convergence and divergence using the two techniques, HCF and GAT, that were employed in the predictive modeling discussed in Section V-A.

1) Experimental Setup: Analyses using HCF: As the *independent variable*, we use the group-level handcrafted features that explain the within-group convergence and divergence, the similarity and dissimilarity in social signals amongst interlocutors. Specifically, the `mean` (μ) aggregation of the *dyad*-level features, and the `standard-deviation` (σ) of the *individual*-level features are used. Intuitively, *larger* σ of individual level features indicate a group that is *diverging*, while *smaller* values indicate that it is *converging*. Contrarily, *larger* μ of dyad level features imply convergence and *smaller* values that of divergence. As the *dependent variable*, the ground-truth annotations of group affect is used. To model the relationship, a least-squares based polynomial regression with *two* degrees of freedom is used, where the relationship is modeled as a 2^{nd} degree polynomial in the independent variables. The regression model is formulated as $y_t = \alpha x_t^2 + \beta x_t + c$.

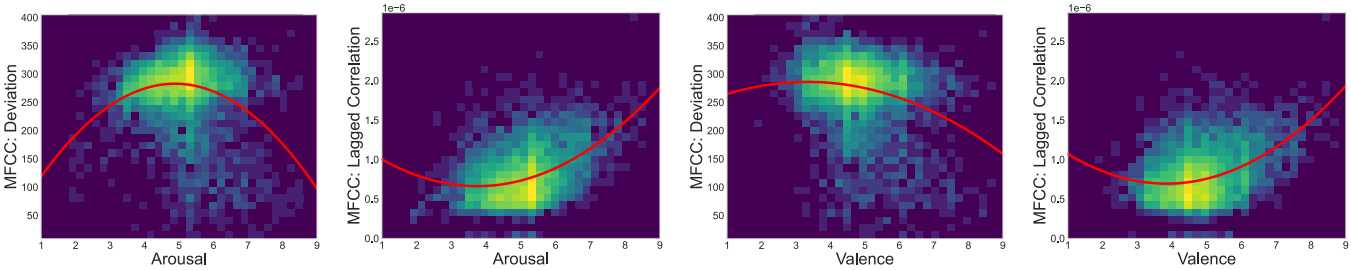
For the quantitative analysis the regression coefficients of the polynomial model (α, β, and c) are analyzed for all the independent measures used. Along with the regression coefficients, the polynomial model's R-Squared (R^2) is also analyzed. Additionally, the Kendal's rank correlation coefficient τ, along with its statistical significance ascertained with a two-tailed p-value $\leq 10\%$ (denoted by $*$), is used to reveal the direction (positive or negative) of linear relationship in the ordinal scale of group affect.

Analyses using GAT: The GAT layer models interpersonal relationships within a group by learning an attention weight α_{ij} for each dyad (as discussed in Section IV-C2), where higher values indicate greater importance in modeling group affect—suggesting more synchronous (convergent or divergent) behavior between interlocutors i and j. Lower values imply weaker synchronization and less affective alignment. For visualization, in Fig. 4 we plot the graph $\mathcal{G}$ with edge width and opacity proportional to the corresponding α_{ij} values. Additionally, we also analyze the standard deviation $\sigma(\alpha_{ij})$ and mean $\mu(\alpha_{ij})$ of attention weights of the graph at a particular time segment. To preserve the full dynamic range and expressivity of the attention weights, we compute the mean and standard deviation on the raw, non-normalized values. A lower value of $\sigma(\alpha_{ij})$ indicates that all interpersonal relationships are of equal equal (un-)importance, with none being especially prominent, whereas a higher value implies that certain interpersonal relationships are assigned a notably greater level of importance than others. Similarly, a higher value of $\mu(\alpha_{ij})$ suggests that, on average, the model assigns greater importance to interpersonal connections overall, while a lower value indicates a more uniformly low weighting across relationships.

2) Quantitative Analyses and Discussion: The convergence-divergence analyses using HCF is presented in Table IV and

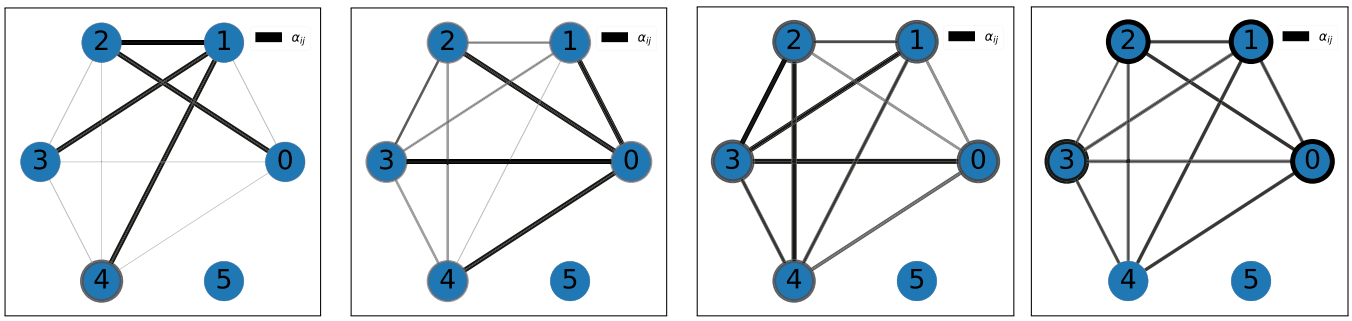
TABLE IV
QUANTITATIVE ANALYSIS OF THE CONVERGENCE-DIVERGENCE PROCESSES

		Arousal					Valence				
		Regression Analysis				Kendal's	Regression Analysis				Kendal's
		α	β	c	R^2	τ	α	β	c	R^2	τ
Pitch	σ	-2.038	+22.322	-12.183	1.2%	-0.081	-0.650	-6.654	+63.554	1%	-0.002
	ρ	+0.002	-0.019	+0.023	6%	$+0.210^*$	+0.002	-0.013	+0.001	4%	$+0.212^*$
VGGish	σ	-0.090	+0.851	+0.646	2.1%	-0.105	-0.027	-0.462	+4.168	3.7%	-0.194^*
	ρ	+0.013	-0.146	+ 1.123	7.2%	$+0.027$	-0.005	+ 0.018	+ 0.660	2%	$+0.171^*$
	Θ_{as}	+0.030	-0.026	+0.324	13%	$+0.112$	-0.001	+0.018	+0.211	18%	$+0.168^*$
MFCC (1st Coeff.)	σ	-10.925	+106.629	+22.751	7.7%	-0.170^*	-3.957	+26.370	+242.151	6.4%	-0.210^*
	ρ_δ	+0.010	+0.029	+0.001	16.4%	$+0.293^*$	+0.021	+0.000	+0.017	15%	$+0.270^*$
	Θ_s	+0.001	+0.008	-0.022	2%	$+0.023$	+0.003	-0.034	+0.094	2%	$+0.037$
AU06	σ	-0.002	-0.010	+0.058	3.3%	-0.210^*	-0.001	-0.004	+0.077	6.6%	-0.259^*
	ρ	+0.018	-0.125	+0.209	15.5%	$+0.264^*$	+0.017	-0.102	+0.150	18.5%	$+0.275^*$
AU07	σ	+0.002	+ 0.001	+0.067	7.0%	-0.213^*	-0.002	0.025	-0.004	14.8%	-0.343^*
	ρ	+0.077	-0.059	+0.107	2.0%	$+0.138$	+0.067	-0.047	+0.076	2.0%	$+0.141$
AU12	σ	-0.004	-0.007	+0.277	1.5%	-0.104	-0.003	+0.02	+0.211	2.2%	-0.130
	ρ	+0.019	-0.135	+0.240	16.1%	$+0.230^*$	+0.015	-0.085	0.115	18.4%	$+0.272^*$
AU25	σ	-0.012	+0.009	+0.135	3.1%	$+0.027$	-0.004	+0.069	-0.118	3.5%	$+0.233^*$
	ρ	+0.013	-0.109	+0.224	8%	$+0.125$	+0.012	-0.090	+0.172	8%	$+0.132$
Head Roll	ρ_δ	+0.178	-1.799	+14.633	2%	$+0.015$	+0.113	-1.243	13.511	2%	$+0.035$



(a) Standard deviation (b) Time-lagged correlation (c) Standard deviation (d) Time-lagged correlation

Fig. 3. Relationship between convergence-divergence measures and group affect: *Arousal* (a,b) and *Valence* (c,d).

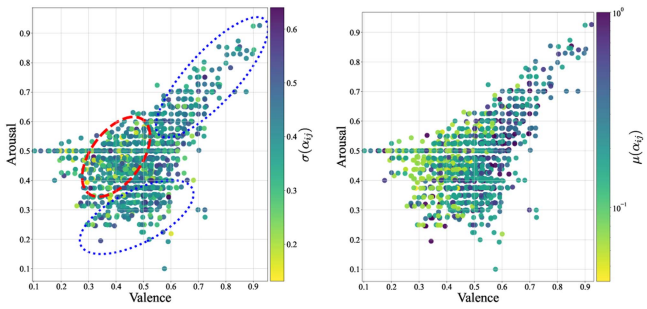


(a) Arousal:0.75, Valence:0.73 (b) Arousal:0.29, Valence:0.35 (c) and (d): Arousal:0.47, Valence:0.50

Fig. 4. Visualization of attention weights α_{ij}, for Positive (a), Negative (b), and Neutral (c and d) samples.

Fig. 3. Similarly, the analyses using GAT can be seen in Figs. 4 and 5. Based on both these analyses we make the following observations.

Trends across the affect scale: We note that the interacting groups tend to *diverge* in terms of their social signals along neutral levels of arousal and valence (i.e., mid-scale values of 4 to 6) and *converge* along extreme levels of arousal and valence (i.e., strong positive affect values of 8-9, or, strong negative values of 1-2). This trend is inferred using the *negative* α values for deviation based group-level features (i.e., σ features), and *positive* α values for synchrony and convergence based features (i.e., ρ, ρ_δ, Θ_s features). Note that negative α values denote concave curves (e.g., seen in Fig. 3(a) and (c)), and positive α values denote convex curves (e.g., seen in Fig. 3(b) and (d)).



(a) Distribution of $\sigma(\alpha_{ij})$ (b) Distribution of $\mu(\alpha_{ij})$

Fig. 5. Distribution of the aggregates of attention weights (α_{ij}) relative to arousal and valence.

However, there are instances where this trend does not align, such as with the ρ and Θ_{as} features of VGGish. Notably, in these cases, the degree of convexity or concavity is rather minimal; that is, the $|\alpha| \leq 0.005$. As a result, the significance of the negative or positive sign diminishes, particularly when α tends towards 0, making the relationship more linear.

Similar patterns emerge in the analysis of the attention weights α_{ij} within the GAT layers. As illustrated in Fig. 4, the distribution of α_{ij} varies distinctly across different levels of arousal and valence. A notable contrast is observed between extreme affective states and neutral levels. First, in the case of extreme arousal and valence levels (see Fig. 4(a) and (b)), several α_{ij} values stand out prominently, exhibiting significantly higher weights compared to other dyadic relationships (captured by α_{ij}) as well as individual-level contributions (captured by α_{ii}). Importantly, every interlocutor in such groups is involved in at least one high-weight dyadic connection, suggesting the presence of strong interpersonal dynamics indicative of extreme group affect levels. These groups are characterized by a *high* $\sigma(\alpha_{ij})$, reflecting greater variability in dyadic attention weights. In contrast, for groups exhibiting neutral affective states, no dyadic connection is notably strong, with all α_{ij} values remain uniformly low. This indicates that interpersonal synchrony is weak or undifferentiated, and such groups exhibit a *low* $\sigma(\alpha_{ij})$. Second, for extreme affective states, dyadic interactions dominate the graph structure, with $\alpha_{ij} > \alpha_{ii}$, highlighting the significance of interpersonal relationships over individual-level cues. Conversely, in neutral conditions, individual features are equally or more influential, with $\alpha_{ij} \leq \alpha_{ii}$. Additional examples supporting these two key observations are provided in Appendix Section 5.1.

To further investigate this pattern, Fig. 5(a) presents a plot of $\sigma(\alpha_{ij})$ as a function of the corresponding arousal and valence levels. The visualization reveals two distinct clustering patterns. First, two clusters outlined with blue dotted boundaries represent group samples exhibiting *high* $\sigma(\alpha_{ij})$ values, corresponding to *extreme* levels of arousal and valence. Second, a cluster enclosed by a red dashed boundary captures group samples with *neutral* affective states, characterized by *low* $\sigma(\alpha_{ij})$ values. Appendix Section 5.2 presents additional analyses illustrating the evolution of GAT-based attention weights during transitions between different levels of affect.

Positive vs negative affect: We note that convergence is higher for positive affect than for negative affect, in both arousal and valence, suggesting that interlocutors align their social signals more during activities that lead to the emergence of positive affect than when negative affect arises. In the context of the HCF-based analysis (see Table IV), this is reflected by the *negative* Kendall's τ for σ features, and the *positive* τ values for features related to synchrony and convergence (e.g., ρ and ρ_δ).

In the context of GAT layers, although the attention weight patterns α_{ij} appear similar for both positive and negative affect (see Fig. 4(a) and (b), respectively), we further examine the differences by plotting the mean attention weights $\mu(\alpha_{ij})$ as a function of arousal and valence in Fig. 5(b). This plot reveals that the average α_{ij} associated with group interaction segments tends to be higher for *positive* affect compared to *negative* affect. There is a clear decreasing trend in the mean attention weights, with $\mu(\alpha_{ij})$ gradually reducing as group affect shifts from high arousal-valence to low arousal-valence states. This suggests that the GAT learns attention weights α_{ij} in a way that the aggregated node representations h_i (5) positively correlate with the corresponding ground-truth affect. This behavior is consistent with the earlier observations of positive Kendall's τ coefficients (see Table IV) between group affect and handcrafted features capturing synchrony and convergence.

The results of the predictive modeling, along with the convergence-divergence analyses, thereby substantiate the *second contribution*—the multimodal modeling of dynamic group affect through capturing the underlying phenomena of convergence and divergence—outlined in Section I.

VI. Conclusion

In this work, we move beyond the traditional emphasis on individual-level affect [14], [17], [77] to address the relatively underexplored collective, group-level affect [8], [20], [78]. Our contributions to this area are twofold: (1) we proposed a novel group affect annotation methodology grounded in organizational psychology theory, and (2) we introduced a multimodal modeling approach capable of capturing the complex dynamics of affective convergence and divergence underlying dynamic group affect.

For the *first contribution*, we developed an annotation strategy that not only addresses key challenges in group affect annotation, but also ensures methodological alignment with theoretical frameworks from organizational psychology [2], [58], [78]. Specifically, we tackled the critical challenge of capturing the temporal dynamics inherent in group affect. While previous literature often neglected these dynamic aspects [19], [22], [25], existing modeling techniques were also constrained by the lack of annotations that reflect the temporal context within which affective expressions occur [9], [19]. Our approach leverages an iteratively tuned 15-second window to more accurately capture the evolution and fluctuation of group affect over time. Furthermore, our study focuses on complex, purposive groups characterized by dynamic interpersonal interactions—unlike prior works [9], [19], [43], [44], which often centered on non-purposive groups lacking social intactness and goal-oriented interdependence. The inter-annotator agreement analysis on the

collected annotations indicated a moderate level of agreement, as interpreted using Cohen's κ. The quality of the obtained group affect annotations is also evidenced by their ability to capture the ebb and flow of affect, including subtle differences in social signals between consecutive segments of the observed group interactions (see Appendix Section 2 for more detail).

Regarding the *second contribution*, leveraging the collected annotations for group affect, we explored two modeling paradigms for predicting group affect: (i) a handcrafted approach based on synchrony and convergence features, and (ii) a data-driven model employing graph attention mechanisms informed by social network theory. Our findings underscore the critical role of interpersonal dynamics in modeling group affect: the graph-based model consistently outperformed the handcrafted method. Moreover, integrating both audio and visual modalities improved prediction performance for arousal and valence. To further investigate group affect dynamics, we analyzed patterns of convergence and divergence using both feature-based measures and attention weights from the graph model. Quantitative results indicate that when social signals among group members *diverged*, the group affect tended toward *neutral levels*, whereas *convergence* corresponded with more *extreme affect* (either strongly positive or negative). Our results also reveal that group members are more synchronized during the emergence of positive affect compared to negative affect.

In summary, our work advances the study of collective affect by providing a theoretically grounded annotation methodology and demonstrating that modeling fine-grained group dynamics is essential to understanding how group affect emerges and fluctuates in social interactions.

A. Limitations and Future Research Avenues

The group interactions present in MeMo [47] occur among groups of unacquainted participants, with a short longitudinal study spanning 3 interactions over the course of 2 weeks. While this zero-history group setup is very suitable to study emergence processes such as group affect [79], we cannot draw conclusions regarding collective affect and its convergence or divergence mechanisms in *groups that share a history*. Despite our observations of rather vivid discussions among participants in MeMo [47], the interactions are still not "real" groups that collaborate on a day-to-day basis, which may explain the relatively small nuances of affective variance in our annotations (see Fig. 1). Hence, it would be of interest to collect group affect annotations on longitudinal data of "real" groups that collaborate on a day-to-day basis as a future research endeavor.

Moreover, in line with the affect recognition literature that prioritizes *expressed* over experienced emotion [14], [17], we instructed annotators to focus solely on the affect collectively expressed by group members. However, ambiguity can arise when the experienced emotion is not fully expressed [15], such as when individuals surface-act to maintain a happy face [80], down-regulate their emotional expressions in line with social norms [81], or manage their emotional expressions for strategic purposes [82]. To capture such possibilities is beyond the scope of this research work, however this might be addressed in future research by adding self-report measures of emotional labor (e.g., [83]) following a group interaction and examine to what extent observable expressions of collective group affect as studied in the current work may be influenced by individual emotion regulation.

Finally, future work could extend large language models (LLMs) beyond individual-level affect recognition toward modeling group-level emotions and their temporal evolution—two critical yet largely underexplored frontiers in affective computing. Current LLM-based approaches predominantly focus on static, categorical emotion classification and lack mechanisms to capture fine-grained, time-varying affective dynamics across interacting individuals [84], [85], [86], [87]. A promising research direction lies in integrating the relational and temporal reasoning strengths of GNNs with the generative and multimodal reasoning capabilities of LLMs [88]. Such a hybrid framework could enable more holistic modeling of group affect by combining structured interpersonal representations with the contextual understanding of language and vision offered by LLMs. Developing this integration would require advancing LLMs to reason over temporal dependencies, interpersonal dynamics, and continuous affective dimensions—thereby opening new possibilities for understanding and modulating collective emotions in social interactions.

REFERENCES

[1] A. Knight and N. Eisenkraft, "Positive is usually good, negative is not always bad: The effects of group affect on social integration and task performance," *J. Appl. Psychol.*, vol. 100, pp. 1214–1227, Dec. 2014.
[2] S. G. Barsade and A. P. Knight, "Group affect," *Annu. Rev. Organ. Psychol. Organ. Behav.*, vol. 2, no. 1, pp. 21–46, 2015.
[3] V. B. Hinsz and L. Bui, "Socially shared affect: Shared affect, affect sharing, and affective processing in groups," *Group Dyn., Theory, Res., Pract.*, vol. 27, no. 4, 2023, Art. no. 229.
[4] J. R. Hackman and N. Katz, "Group behavior and performance," in *Handbook of Social Psychology*, vol. 2. Hoboken, NJ, USA: Wiley, 2010, pp. 1208–1251.
[5] S. W. Kozlowski and D. R. Ilgen, "Enhancing the effectiveness of work groups and teams," *Psychol. Sci. Public Int.*, vol. 7, no. 3, pp. 77–124, 2006.
[6] A. L. Collins, S. A. Lawrence, A. C. Troth, and P. J. Jordan, "Group affective tone: A review and future research directions," *J. Organizational Behav.*, vol. 34, no. S1, pp. S43–S62, 2013.
[7] C. Jones, S. Volet, and D. Pino-Pasternak, "Observational research in face-to-face small groupwork: Capturing affect as socio-dynamic interpersonal phenomena," *Small Group Res.*, vol. 52, no. 3, pp. 341–376, 2021.
[8] E. A. Veltmeijer, C. Gerritsen, and K. V. Hindriks, "Automatic emotion recognition for groups: A review," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 89–107, Jan.–Mar. 2023.
[9] A. Dhall, J. Joshi, K. Sikka, R. Goecke, and N. Sebe, "The more the merrier: Analysing the affect of a group of people in images," in *Proc. Int. Conf. Automat. Face Gesture Recognit.*, 2015, pp. 1–8.
[10] Z. Lei and N. Lehmann-Willenbrock, "Affect in meetings: An interpersonal construct in dynamic interaction processes," in *The Cambridge Handbook of Meeting Science*. Cambridge, U.K.: Cambridge Univ. Press, 2015, pp. 456–482.
[11] G. Sharma, A. Dhall, and J. Cai, "Audio-visual automatic group affect analysis," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1056–1069, Apr.–Jun. 2023.
[12] J. Rösemeier, X. Hu, and B. A. Nijstad, "Toward a dynamic social process view: An integrative, multidisciplinary review of the relationship between affect and creativity," *Acad. Manag. Ann.*, 2025.
[13] B. Ya-Hui Lien, Y.-C. Hsu, Y.-h. Chen, and L.-W. Chen, "The formation of positive group affective tone: A narrative practice," *Small Group Res.*, vol. 54, no. 2, pp. 277–301, 2023.

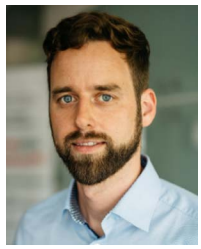
[14] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018.
[15] V. Sethu, E. M. Provost, J. Epps, C. Busso, N. Cummins, and S. Narayanan, "The ambiguous world of emotion representation," 2019, *arXiv:1909.00360*.
[16] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, Shanghai, China, Apr. 2013, pp. 1–8.
[17] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 471–483, Oct.–Dec. 2019.
[18] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The MSP-conversation corpus," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 1823–1827.
[19] X. Wang, D. Zhang, and D.-J. Lee, "Implementing the affective mechanism for group emotion recognition with a new graph convolutional network architecture," *IEEE Trans. Affect. Comput.*, vol. 15, no. 3, pp. 1104–1115, Jul.–Sep. 2023.
[20] X. Huang, J. Xu, W. Zheng, Q. Mao, and A. Dhall, "A survey of deep learning for group-level emotion recognition," 2024, *arXiv:2408.15276*.
[21] S. G. Barsade, C. G. Coutifaris, and J. Pillemer, "Emotional contagion in organizational life," *Res. Organizational Behav.*, vol. 38, pp. 137–151, 2018.
[22] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, "From individual to group-level emotion recognition: Emotiw 5.0," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2017, pp. 524–528.
[23] A. Dhall et al., "Emotiw 2023: Emotion recognition in the wild challenge," in *Proc. 25th Int. Conf. Multimodal Interaction*, 2023, pp. 746–749.
[24] N. Lehmann-Willenbrock, "Dynamic interpersonal processes at work: Taking social interactions seriously," *Annu. Rev. Organizational Psychol. Organizational Behav.*, vol. 12, pp. 133–158, 2025.
[25] X. Huang, A. Dhall, R. Goecke, M. Pietikäinen, and G. Zhao, "Multimodal framework for analyzing the affect of a group of people," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2706–2721, Oct. 2018.
[26] C. Raman, N. Raj Prabhu, and H. Hung, "Perceived conversation quality in spontaneous interactions," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 2901–2912, Oct.–Dec. 2023.
[27] C. A. Bartel and R. Saavedra, "The collective construction of work group moods," *Administ. Sci. Quart.*, vol. 45, no. 2, 2000, pp. 197–231.
[28] S. Hareli and A. Rafaeli, "Emotion cycles: On the social influence of emotion in organizations," *Res. Organizational Behav.*, vol. 28, pp. 35–59, 2008.
[29] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–10.
[30] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen, "Interpersonal synchrony: A survey of evaluation methods across disciplines," *IEEE Trans. Affect. Comput.*, vol. 3, no. 3, pp. 349–365, Jul.–Sep. 2012.
[31] M. C. Nanninga, Y. Zhang, N. Lehmann-Willenbrock, Z. Szlávik, and H. Hung, "Estimating verbal expressions of task and social cohesion in meetings by quantifying paralinguistic mimicry," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2017, pp. 206–215.
[32] S. S. Singh, S. Muhuri, S. Mishra, D. Srivastava, H. K. Shakya, and N. Kumar, "Social network analysis: A survey on process, tools, and application," *ACM Comput. Surv.*, vol. 56, no. 8, 2024.
[33] E. R. Smith, C. R. Seger, and D. M. Mackie, "Can emotions be truly group level? Evidence regarding four conceptual criteria," *J. Pers. Social Psychol.*, vol. 93, no. 3, 2007, Art. no. 431.
[34] P. Ekman, *Are There Basic Emotions?*. Washington, DC, USA: Amer. Psychological Assoc., 1992.
[35] J. A. Russell, "A circumplex model of affect," *J. Pers. Social Psychol.*, vol. 39, no. 6, 1980, Art. no. 1161.
[36] N. Lehmann-Willenbrock, R. A. Meyers, S. Kauffeld, A. Neininger, and A. Henschel, "Verbal interaction sequences and group mood: Exploring the role of team planning communication," *Small Group Res.*, vol. 42, no. 6, pp. 639–668, 2011.
[37] W. Mou, H. Gunes, and I. Patras, "Alone versus in-a-group: A multi-modal framework for automatic affect recognition," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 2, 2019.
[38] S. G. Barsade, "The ripple effect: Emotional contagion and its influence on group behavior," *Administ. Sci. Quart.*, vol. 47, no. 4, pp. 644–675, 2002.
[39] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon, "Emotiw 2018: Audio-video, student engagement and group-level affect prediction," in *Proc. 25th Int. Conf. Multimodal Interaction*, 2018, pp. 746–749.
[40] S. Ghosh, A. Dhall, and N. Sebe, "Automatic group affect analysis in images via visual attribute and feature networks," in *Proc. 25th IEEE Int. Conf. Image*, 2018, pp. 1967–1971.
[41] G. Sharma, S. Ghosh, and A. Dhall, "Automatic group level affect and cohesion prediction in videos," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction Workshops Demos*, 2019, pp. 161–167.
[42] S. Ghosh et al., "Emolysis: A multimodal open-source group emotion analysis and visualization toolkit," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction Workshops Demos*, 2024, pp. 116–118.
[43] Y. Wang, S. Zhou, Y. Liu, K. Wang, F. Fang, and H. Qian, "ConGNN: Context-consistent cross-graph neural network for group emotion recognition in the wild," *Inf. Sci.*, vol. 610, pp. 707–724, 2022.
[44] X. Wang, T. Chen, and D. Zhang, "A spatial-temporal graph convolutional network for video-based group emotion recognition," in *Proc. Int. Conf. Pattern Recognit.*, 2024, pp. 339–354.
[45] E. Hatfield, J. Cacioppo, and R. Rapson, *Emotional Contagion Cambridge*. Cambridge, U.K.: Cambridge Univ. Press, 1994.
[46] F. Walter and H. Bruch, "The positive group affect spiral: A dynamic model of the emergence of positive affective similarity in work groups," *J. Organizational Behav., Int. J. Ind., Occup. Organizational Psychol. Behav.*, vol. 29, no. 2, pp. 239–261, 2008.
[47] M. Tsfasman, B. Dudzik, K. Fenech, A. Lorincz, C. M. Jonker, and C. Oertel, "Introducing MeMo: A multimodal dataset for memory modelling in multiparty conversations," 2024, *arXiv:2409.13715*.
[48] N. Raj Prabhu, C. Raman, and H. Hung, "Defining and quantifying conversation quality in spontaneous interactions," in *Proc. Companion Publ. Int. Conf. Multimodal Interaction*, Sep. 2020, pp. 196–205.
[49] P. Mangold, "Discover the invisible through tool-supported scientific observation," in H. Böttger, K. Jensen, and T. Jensen, in *Proc. editors-Mindful Evol. Conf.*, 2018.
[50] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Ther. Exp. Psychiatry*, vol. 25, pp. 49–59, 1994.
[51] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," in *Proc. IEEE Int. Conf. Affect. Comput. Intell. Interact.*, 2017, pp. 248–255.
[52] C. Busso et al., "Iemocap: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.
[53] N. Raj Prabhu, N. Lehmann-Willenbrock, and T. Gerkmann, "End-to-end label uncertainty modeling in speech emotion recognition using Bayesian neural networks and label distribution learning," *IEEE Trans. Affect. Comput.*, vol. 15, no. 2, pp. 579–592, Apr.–Jun. 2024.
[54] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, 2012.
[55] J. A. Gliem et al., "Calculating, interpreting, and reporting cronbach's alpha reliability coefficient for likert-type scales," in *Proc. Midwest Res. Pract. Conf. Adult, Continuing, Community Educ.*, Columbus, OH, USA, 2003, vol. 1, pp. 82–87.
[56] D. Freedman, R. Pisani, and R. Purves, *Statistics*, 4th ed. New York, NY, USA: Norton, 2007.
[57] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, Jan. 2005, pp. 381–385.
[58] Z. Lei and N. Lehmann-Willenbrock, "Contagious peers in teams: Peer affective influence on individual emotions and performance," *Proc. Acad. Manage.*, vol. 2014, Oct. 2014, Art. no. 13936.
[59] P. Tzirakis, J. Chen, S. Zafeiriou, and B. Schuller, "End-to-end multimodal affect recognition in real-world environments," *Inf. Fusion*, vol. 68, pp. 46–53, 2021.
[60] J. Wagner et al., "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10745–10759, Sep. 2023.
[61] D. de Oliveira, N. Raj Prabhu, and T. Gerkmann, "Leveraging semantic information for efficient self-supervised emotion recognition with audio-textual distilled models," in *Proc. Interspeech*, 2023, pp. 3632–3636.
[62] K. Sridhar and C. Busso, "Unsupervised personalization of an emotion recognition system: The unique properties of the externalization of valence in speech," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 1959–1972, Oct.–Dec. 2022.
[63] F. Eyben et al., "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr.–Jun. 2016.

[64] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing parselmouth: A python interface to praat," *J. Phonetics*, vol. 71, 2018, pp. 1–15.
[65] N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behav. Res. Methods*, vol. 41, no. 2, pp. 385–390, 2009.
[66] S. Hershey et al., "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 131–135.
[67] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, Feb. 2001.
[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[69] K. J. Klein and S. W. Kozlowski, "A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes," in *Proc. Multilevel Theory, Res., Methods Org., Found., Extensions, New Directions*, 2000, pp. 3–90.
[70] J. Vargas-Quiros, Ö. Kapcak, H. Hung, and L. Cabrera-Quiros, "Individual and joint body movement assessed by wearable sensing as a predictor of attraction in speed dates," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2168–2181, Jul.–Sep. 2023.
[71] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–14.
[72] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Representations*, 2018.
[73] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 5089–5093.
[74] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.
[75] J. R. Kelly and S. G. Barsade, "Mood and emotions in small groups and work teams," *Organizational Behav. Hum. Decis. Process.*, vol. 86, no. 1, pp. 99–130, 2001.
[76] C. Oertel, K. A. Funes Mora, J. Gustafson, and J.-M. Odobez, "Deciphering the silent participant: On the use of audio-visual cues for the classification of listener categories in group discussions," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2015, pp. 107–114.
[77] S. Alisamir and F. Ringeval, "On the evolution of speech representations for affective computing: A brief history and critical overview," *IEEE Signal Process., Mag.*, vol. 38, no. 6, pp. 12–21, Nov. 2021.
[78] S. Barsade and D. Gibson, "Group affect," *Curr. Directions Psychol. Sci.*, vol. 21, pp. 119–123, Mar. 2012.
[79] S. W. Kozlowski, "Advancing research on team process dynamics: Theoretical, methodological, and measurement considerations," *Organizational Psychol. Rev.*, vol. 5, no. 4, pp. 270–299, 2015.
[80] A. S. Gabriel, M. A. Daniels, J. M. Diefendorff, and G. J. Greguras, "Emotional labor actors: A latent profile analysis of emotional labor strategies," *J. Appl. Psychol.*, vol. 100, no. 3, 2015, pp. 863–879.
[81] D. Geddes and D. Lindebaum, "Unpacking the 'why' behind strategic emotion expression at work: A narrative review and proposed taxonomy," *Eur. Manage. J.*, vol. 38, no. 5, pp. 708–722, 2020.
[82] F. Liu and S. Maitlis, "Emotional dynamics and strategizing processes: A study of strategic conversations in top team meetings," *J. Manage. Stud.*, vol. 51, no. 2, pp. 202–234, 2014.
[83] T. M. Glomb and M. J. Tews, "Emotional labor: A conceptualization and scale development," *J. Vocational Behav.*, vol. 64, no. 1, pp. 1–23, 2004.
[84] Z. Cheng et al., "Emotion-LLaMA: Multimodal emotion recognition and reasoning with instruction tuning," in *Proc. Adv. Neural Inf. Proc. Syst.*, 2024, vol. 37, pp. 110805–110853.
[85] S. Dutta and S. Ganapathy, "LLM supervised pre-training for multimodal emotion recognition in conversations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2025, pp. 1–5.
[86] S. Madan et al., "MIP-GAF: A MLLM-annotated benchmark for important person localization and group context understanding," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2025, pp. 1467–1476.
[87] J. Zhang, Z. Mai, Z. Xu, and Z. Xiao, "Is LLAMA 3 good at identifying emotion? A comprehensive study," in *Proc. Int. Conf. Mach. Learn. Mach. Intell.*, 2024, pp. 128–132.
[88] S. Wang et al., "Graph machine learning in the era of large language models (LLMs)," *Trans. Intell. Syst. Technol.*, vol. 16, no. 5, pp. 1–40, 2025.

Navin Raj Prabhu (Member, IEEE) received the BTech degree in computer science from SRM University, India, in 2015, and the MS degree in computer science from the Delft University of Technology, Delft, The Netherlands, in 2020. He is currently working toward the PhD degree with the Signal Processing Lab and also with the Organization Psychology Lab, University of Hamburg, Germany. His research interests include affective computing, social signal processing, deep learning, uncertainty modeling, generative modeling, and emotional speech synthesis, and group affect.

Maria Tsfasman received the BSc degree in fundamental and computational linguistics from HSE University, Moscow, and the MSc (with distinction) degree in artificial intelligence from Radboud University, Nijmegen. Her research focuses on cognitive and computer science: training machines to understand humans better and using the insights from these machines to expand global understanding of human social processing and cognition.

Catharine Oertel is currently an assistant professor with TU Delft, The Netherlands. She is also the co-principal investigator of the Designing Intelligence Lab (DI Lab), an effort aiming to bridge research done in computer science with industrial design engineering. Her research focuses on understanding and modeling human interaction to build socially aware conversational agents able to engage with people in a human-like manner.



Timo Gerkmann (Senior Member, IEEE) is currently a professor of signal processing with the University of Hamburg, Germany. He has held positions with Technicolor Research & Innovation in Germany, University of Oldenburg, Germany, KTH Royal Institute of Technology, Sweden, Ruhr-Universität Bochum, Germany, and Siemens Corporate Research, Princeton, NJ, USA. His research interests include statistical signal processing and machine learning for speech and audio applied to communication devices, hearing instruments, audio-visual media, and human-machine interfaces. Timo Gerkmann was a member of IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing (2018-2023), Associate Editor (2019-2022) and since 2022, he has been the senior area editor of *IEEE/ACM Transactions on Audio, Speech and Language Processing*. He was the recipient of VDE ITG award 2022.

Nale Lehmann-Willenbrock received the PhD degree in psychology from Technische Universität Braunschweig in 2012. She held positions as an associate professor with the University of Amsterdam and assistant professor with Vrije Universiteit Amsterdam. She is currently a professor of industrial and organizational psychology, director of the Center for Better Work, and vice dean of research and transfer with the Faculty of Psychology and Human Movement Science, University of Hamburg. She studies dynamic social interaction patterns in groups and teams, interpersonal processes among leaders and followers, and meetings as a core interaction site in organizations. Her research program blends organizational psychology, management, communication, and social signal processing. She was an associate editor for *Small Group Research* (2019-2024). She is an associate editor for *Journal of Business and Psychology* and *Group & Organization Management*.